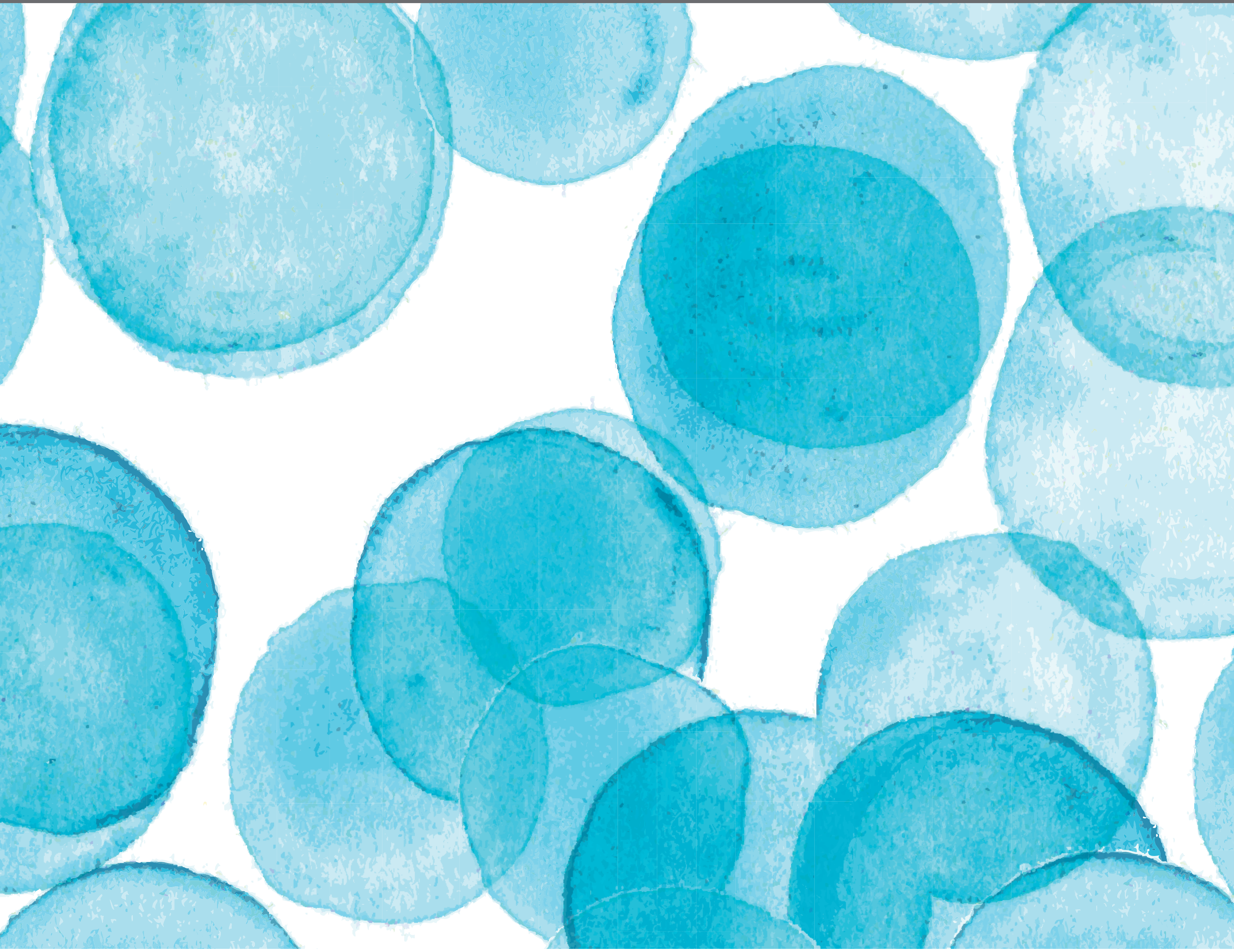
# HIV-1 GENETIC DIVERSITY, VOLUME II

EDITED BY: Kok Keng Tee, Michael M. Thomson and Joris Hemelaar

PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-940-7
DOI 10.3389/978-2-88976-940-7

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

# HIV-1 GENETIC DIVERSITY, VOLUME II

Topic Editors:
**Kok Keng Tee,** University of Malaya, Malaysia
**Michael M. Thomson,** Carlos III Health Institute (ISCIII), Spain
**Joris Hemelaar,** University of Oxford, United Kingdom

**Citation:** Tee, K. K., Thomson, M. M., Hemelaar, J., eds. (2022). HIV-1 Genetic Diversity, Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-940-7

# Table of Contents

frontiers | Frontiers in Microbiology

TYPE Editorial
PUBLISHED 22 August 2022
DOI 10.3389/fmicb.2022.1007037

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Linqi Zhang,
Tsinghua University, China

*CORRESPONDENCE
Kok Keng Tee
k2tee@um.edu.my
Michael M. Thomson
mthomson@isciii.es
Joris Hemelaar
joris.hemelaar@ndph.ox.ac.uk

SPECIALTY SECTION
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

RECEIVED 29 July 2022
ACCEPTED 05 August 2022
PUBLISHED 22 August 2022

CITATION
Tee KK, Thomson MM and Hemelaar J (2022) Editorial: HIV-1 genetic diversity, volume II. *Front. Microbiol.* 13:1007037. doi: 10.3389/fmicb.2022.1007037

# Editorial: HIV-1 genetic diversity, volume II

Kok Keng Tee[1]*, Michael M. Thomson[2]* and Joris Hemelaar[3]*

[1]Department of Medical Microbiology, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia, [2]HIV Biology and Variability Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain, [3]Infectious Disease Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

KEYWORDS

**HIV, subtype, circulating recombinant form (CRF), unique recombinant form (URF), phylogenetic analysis, recombinant**

Editorial on the Research Topic
HIV-1 genetic diversity, volume II

The HIV pandemic continues to be a major global health problem. In 2021, 38.4 million people were living with HIV worldwide. Despite the increasing availability of antiretroviral therapy (ART) worldwide, around 650,000 deaths and 1.5 million new HIV infections occurred in 2021 (UNAIDS, 2022).

A key characteristic of the HIV pandemic is its extraordinary global genetic diversity. After zoonotic transmission of simian immunodeficiency virus from chimpanzees to humans in the beginning of the twentieth century, HIV-1 group M diversified in central Africa in the first half of the century, leading to distinct subtypes, designated by the letters A, B, C, D, F, G, H, J, K, and L (Robertson et al., 2000; Worobey et al., 2008). The second half of the twentieth century was characterized by the global spread of HIV-1 and ongoing diversification (Tebit and Arts, 2011; Hemelaar, 2012; Faria et al., 2014). Genetic divergence between HIV-1 subtypes is around 17–35% at the amino acid level, depending on the subtypes and genome regions considered (Korber et al., 2001).

HIV-1 genetic variability arises due to the error-prone reverse transcriptase enzyme, which leads to high rates of mutation and recombination. A prerequisite for recombination is that an individual is co-infected with two or more different strains of HIV (Vuilleumier and Bonhoeffer, 2015). Recombinants between subtypes are designated as either circulating recombinant forms (CRFs) or unique recombinant forms (URFs) (Robertson et al., 2000). CRFs are defined as recombinant HIV-1 genomes that are identified in three or more epidemiologically unrelated individuals. URFs refer to unique recombinant sequences without evidence of onward transmission. CRFs are consecutively named, in accordance with an internationally defined nomenclature (Robertson et al., 2000). To date, more than 120 distinct CRFs have been described, a number that continues to increase at pace (Hemelaar et al., 2019, 2020b). CRFs can undergo further recombination with other pure subtypes or recombinants, resulting in secondary recombinants, leading to an increasingly complex array of recombinants (Hemelaar, 2012). The proportion of recombinants has been increasing over time, both globally and in most regions, and recombinants now constitute close to a quarter of all HIV-1 infections (Hemelaar et al., 2020a).

Global molecular epidemiological studies have demonstrated that HIV-1 genetic diversity is extremely complex and evolving (Hemelaar et al., 2019, 2020b). The global spread and evolution of HIV-1 has caused differential global distributions of HIV-1 subtypes, CRFs, and URFs, leading to large regional variation in numbers, types, and proportions of HIV-1 genetic variants. The global HIV-1 epidemic is therefore diversifying and recombinants play particularly important roles in Africa, Asia, and South America (Hemelaar et al., 2020a,b). The diverse distribution patterns of HIV variants are determined by complex factors, including social transmission networks, urbanization, transportation networks, migration, founder effects, and population growth. It may also be that different HIV variants could have an evolutionary advantage in terms of transmission and pathogenesis (Arien et al., 2007; Tebit and Arts, 2011; Faria et al., 2014). Increasing global HIV-1 genetic diversity clearly forms a major obstacle to development of a globally effective HIV-1 vaccine (Gaschen et al., 2002). It also impacts the design of diagnostic, resistance, and viral load assays. Finally, the variability and rapid evolution of HIV-1 provide the means to examine the evolutionary relationships and origins of strains (phylogenetics), the growth dynamics of transmission networks (phylodynamics), and to track the geographical spread of HIV-1 (phylogeography) (Hemelaar, 2012, 2013).

This Research Topic brings together studies that further expand our knowledge of the origins and spread of HIV-1 genetic variants, and examines the impact of HIV-1 diversity on prevention and treatment efforts, including HIV-1 vaccine development and drug resistance.

## Recombination

In this issue, Bacqué et al. conducted a HIV-1 molecular epidemiological study among patients recruited in Spain. A novel CRF derived from subtypes B and F1, designated CRF66_BF, was characterized using whole genome sequencing and detailed phylogenetic and recombination analyses. Bayesian coalescent analyses, which estimated the divergence time of the most recent common ancestors of the sampled genomes, showed that the probable origin of CRF66_BF was in Paraguay around 1984.

Inter-continental transmission of HIV-1 among countries with close socioeconomic relationship is efficient in driving the dissemination of novel recombinants. For example, CRF47_BF of South American origin has expanded considerably in Spain since it was first reported in 2010, with a predominant transmission via heterosexual contact. Hill et al. in this issue revealed that CRF47_BF originated in Brazil, before it spread into Spain and expanded rapidly until the mid-2010s, with evidence of spillover into the men who have sex with men (MSM) population. This and other studies established the repeated introduction and expansion of CRFs in Spain, which highlights the need to establish molecular epidemiological surveillance systems that could provide timely information on cross-border introduction and dynamics of HIV-1 strains.

However, precision in CRF characterization can be compromised by the extreme genome plasticity of HIV-1, in addition to the lack of complete genome sequences and a standardized parameter in recombination analyses. Here, Cañada-García et al. identified a novel HIV-1 BF1 recombinant (CRF122_BF1) in Europe and South America that was previously unidentified (or, rather, misclassified as CRF72_BF1). Apart from different recombination signals detected in the polymerase and envelope genes, CRF122_BF1 was otherwise highly similar to CRF72_BF1 that was described in Brazil. The study highlighted the continuous emergence of new CRFs as a result of co-circulation of multiple viral lineages.

The expanding complexity of HIV-1 recombinants in Africa and Asia are also illustrated in this issue. Among acutely and recently infected patients in Kigali, Rwanda, Umviligihozo et al. reported an increasing frequency of URFs from 23% in a preceding cohort in 2005–2011 to 57% in 2016–2019 that comprised of inter-subtype A1/C and A1/C/D recombinants. Similarly, He et al. reported significant prevalence of URFs among the newly diagnosed MSM population in Shenyang city of Liaoning province, northeast China, between 2016 and 2020, which involved the CRF01_AE/CRF07_BC and CRF01_AE/B recombinants. Taken together, these studies highlight the power of molecular epidemiological surveillance in tracking the evolutionary dynamics of HIV-1 worldwide.

## Molecular epidemiology and phylodynamics

While the study of defined HIV-1 subtypes and recombinants is clearly important, the emergence and expansion of diverse lineages *within* subtypes and CRFs have also become the focus of attention. Such lineages represent viruses sharing a common ancestry propagating within a transmission network, some of which are associated with peculiar biological features (Cid-Silva et al., 2018; Song et al., 2019; Ge et al., 2021; Wymant et al., 2022). The study of the emergence, spatiotemporal propagation, and growth dynamics of HIV-1 variants, based on the analysis of viral sequence evolution, is the subject of phylodynamics. Tracking the expansion of HIV-1 lineages through phylodynamic and phylogenetic methods can inform the design of public health interventions aimed at epidemic control (Brenner et al., 2013; Paraskevis et al., 2016; German et al., 2017; Oster et al., 2018; Vasylyeva et al., 2020).

In this Research Topic several papers fall within the field of HIV-1 molecular epidemiology. Three of them are focused on the use of phylodynamics to track HIV-1 epidemic spread of intra-subtype lineages. Arantes et al. and Arimide et al.

focus on estimation of growth rates of lineages circulating in the Amazonas state of Brazil and Ethiopia, respectively. Arantes et al. found continuous expansion until most recent times and comparable epidemic growth rates of Amazonian subtype B non-pandemic [derived from the original subtype B radiation from Haiti (Gilbert et al., 2007)] and pandemic (derived from the subtype B expansion from the USA that disseminated worldwide) lineages. Arimide et al. found sharp declines in transmission parameters in all Ethiopian subtype C lineages coinciding with public health awareness campaigns and behavioral interventions in the mid-1990s, a decade before ART roll-out. On the other hand, Nduva et al. focused on phylogeographic analyses to estimate geographic dissemination of HIV-1 lineages among MSM in Kenya, finding significant dissemination from the Coast to Nairobi and Nyanza provinces and from Nairobi to Nyanza. The public health implications of these results are emphasized by the authors.

The Eastern European and Central Asian region has the fastest growing HIV-1 epidemic in the world, but is insufficiently studied. Sivay et al. examined HIV-1 genetic diversity in Kyrgyzstan, a country in this region for which there were few prior data, analyzing 555 samples. In contrast to most countries in the region, where A6 sub-subtype predominates, in Kyrgyzstan, a Central Asian CRF02_AG variant is predominant, although A6 is also common. No phylogenetic structure was seen in A6, but four geographically-associated lineages were found in CRF02_AG.

The importance of dense sampling for cluster detection and the usefulness of phylogeny for estimating the place of HIV-1 acquisition in migrants is highlighted by Gil et al. who analyzed two densely-sampled Spanish regions, finding an association of clusters with MSM and native Spaniards, but 35% of Latin American immigrants belonged to Spanish clusters (and, therefore, probably acquired HIV-1 in Spain), compared to 1.2% of Sub-Saharan Africans.

While most HIV-1 molecular epidemiology studies focus on the coding regions of HIV, the study by Bhange et al. highlights intra-subtype genetic variation present in the long terminal repeat (LTR) promoter region. In their study of 764 ART-naïve individuals in India they find nine different promoter variant strains of subtype C, which contain additional copies of existing transcription factor binding sites, created by duplication, which may impact viral gene expression and latency.

Methodological improvements for HIV-1 cluster detection are needed in molecular epidemiological studies, and, in line with this, Guang et al. describe a new method based on next generation sequencing incorporating within-host diversity that can detect clusters not detected by consensus sequence approaches.

## Antiretroviral drug resistance

The relationship of antiretroviral (ARV) drug resistance to phylogeny and phylodynamics derives from the fact that HIV-1 drug resistant strains may persist for many years, propagating in phylogenetically-identifiable transmission networks, and from the frequent use of polymerase sequences obtained for drug resistance testing for molecular epidemiology studies. Surveillance of ARV drug resistance transmission is important to monitor the expansion of drug resistant strains, which may affect the choice of first-line ART regimens.

Two papers in this issue focus on such surveillance. Pingarilho et al. analyzed transmitted drug resistance (TDR) and transmission clusters in newly-diagnosed patients in Portugal, finding higher proportions of TDR and clustered TDR among heterosexuals than among MSM, attributing this difference to higher pre-exposure prophylaxis usage and HIV testing among MSM. Miranda et al., using the EuResist database, examined trends of TDR and acquired drug resistance (ADR) in Europe in 1981–2019, comparing late presenters (LP) and non-late presenters (NLP), finding a decreasing trend in both TDR and ADR, and similar TDR frequencies and mutation profiles in LP and NLP.

Monitoring sequence changes in proteins targeted by ARV drugs can inform drug usage in a geographic area and serve as guide to optimize therapeutic choices. Along this line, Bimela et al., using sequences from the Los Alamos HIV Sequence Database, analyzed changes in frequencies of drug resistance mutations (DRM) and naturally occurring polymorphisms in Cameroon, where HIV-1 is highly diverse, before and after implementation of combination ART, finding much more frequent changes in reverse transcriptase than in protease and integrase, mirroring the usage of drugs targeting these enzymes in Cameroon.

The Los Alamos database was also used by Troyano-Hernáez et al. to perform extensive analyses of capsid and polymerase sequences of all circulating HIV-1 genetic forms (groups, subtypes, and CRFs) for variant-specific markers and DRM, finding that mutations in the capsid associated with resistance to lenacapavir (the most promising drug targeting this protein) and major DRM in polymerase in drug-naïve individuals were infrequent in all genetic forms. Valadés-Alcaraz et al. derived HIV transmembrane glycoproteins consensus sequences for subtypes and CRFs and assessed their level of conservation in the different gp41 structural domains, with no natural major resistance mutation to fusion inhibitor T-20 observed.

## Global evolution and vaccines

As highlighted by many studies in this Research Topic, HIV-1 continues to evolve around the world. To reflect this change, Linchangco et al. updated the global whole genome consensus sequences for HIV-1 subtypes and CRFs from 2002 to 2021, based on sequences deposited in the Los Alamos database. Finally, global HIV-1 diversity forms a major obstacle to the development of a globally effective HIV-1 vaccine. Given the genetic divergence between HIV-1 subtypes, it may be necessary to employ subtype-specific vaccines in individual countries according to their HIV-1 subtype distribution. A study by Elangovan et al. estimated the global and regional need for subtype-specific therapeutic and prophylactic HIV-1 vaccines, indicating that to achieve global coverage, HIV-1 vaccines should be mainly directed against subtypes A, B, and C.

## Conclusions

The papers in this Research Topic highlight the incessantly increasing genetic diversification of HIV-1, through the generation of recombinant forms and the emergence and expansion of new lineages. These studies also exemplify the important roles of analyzing the growth dynamics and tracking the geographic spread of HIV-1 variants, through phylogenetic, phylodynamic, and phylogeographic methods, for epidemiological and public health purposes. Another area of interest in this issue relates to the importance of surveying the constantly evolving picture of HIV-1 genetic diversity to inform vaccine immunogen design and to ensure the effectiveness of ARV drugs.

To date, many molecular epidemiology studies have been unsystematic and uncoordinated. There is a need for a coordinated global molecular epidemiological surveillance system that could provide up-to-date, accurate, and geographically representative information on the evolution and spread of HIV variants to aid prevention and treatment efforts. At present, most subtyping is done as an adjunct to resistance testing, performed by *pol* sequencing. This means that samples are often unrepresentative of populations and recombinants may be missed, as no information is available outside *pol*. Representative sampling from key populations and/or the general population will be essential and will depend on the state of the HIV epidemic in each country (Aldrich and Hemelaar, 2012). Moreover, whole genome sequencing is crucial to adequately characterize HIV strains and detect recombinants. With the increasing availability of deep sequencing, whole genome sequencing is becoming increasingly feasible, and also enhances detecting dual/multiple infections and minority drug-resistant strains (Aldrich and Hemelaar, 2012).

The use of phylogenetic and phylodynamic methods in public health is increasingly being advocated, and, in fact, the plan for ending the HIV epidemic in the USA includes the use of such tools for rapid HIV-1 outbreak detection and response (Fauci et al., 2019). However, further empirical data on the effectiveness of such strategies are needed. More research is needed on the impact of HIV-1 diversification and the increasing proportion of recombinants on transmission and pathogenesis. Finally, in order to end the global HIV epidemic, it is imperative that the recent successes in developing COVID vaccines are harnessed to accelerate HIV-1 vaccine development, with immunogen sequence design that accounts for global HIV-1 diversity.

## Author contributions

All authors contributed equally to the writing of this editorial and read and approved the final version of the manuscript.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Aldrich, C., and Hemelaar, J. (2012). Global Hiv-1 diversity surveillance. *Trends Mol. Med.* 18, 691–694. doi: 10.1016/j.molmed.2012.06.004

Arien, K. K., Vanham, G., and Arts, E. J. (2007). Is HIV-1 evolving to a less virulent form in humans? *Nat. Rev. Microbiol.* 5, 141–151. doi: 10.1038/nrmicro1594

Brenner, B., Wainberg, M. A., and Roger, M. (2013). Phylogenetic inferences On Hiv-1 transmission: implications for the design of prevention and treatment interventions. *AIDS* 27, 1045–1057. doi: 10.1097/QAD.0b013e32835cffd9

Cid-Silva, P., Margusino-Framiñán, L., Balboa-Barreiro, V., Martín-Herranz, I., Castro-Iglesias, Á., Pernas-Souto, B., et al. (2018). Initial treatment response among HIV subtype f infected patients who started antiretroviral therapy based on integrase inhibitors. *AIDS* 32, 121–125. doi: 10.1097/QAD.0000000000001679

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). Hiv epidemiology. The early spread and epidemic ignition of Hiv-1 in human populations. *Science* 346, 56–61. doi: 10.1126/science.1256739

Fauci, A. S., Redfield, R. R., Sigounas, G., Weahkee, M. D., and Giroir, B. P. (2019). Ending The Hiv epidemic: a plan for the United States. *JAMA* 321, 844–845. doi: 10.1001/jama.2019.1343

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., et al. (2002). Diversity considerations in Hiv-1 vaccine selection. *Science* 296, 2354–2360. doi: 10.1126/science.1070441

Ge, Z., Feng, Y., Li, K., Lv, B., Zaongo, S. D., Sun, J., et al. (2021). Crf01_Ae and Crf01_Ae Cluster 4 are associated with poor immune recovery in Chinese patients under combination antiretroviral therapy. *Clin. Infect. Dis.* 72, 1799–1809. doi: 10.1093/cid/ciaa380

German, D., Grabowski, M. K., and Beyrer, C. (2017). Enhanced use of phylogenetic data to inform public health approaches to hiv among men who have sex with men. *Sex Health* 14, 89–96. doi: 10.1071/SH16056

Gilbert, M. T., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., and Worobey, M. (2007). The emergence of Hiv/Aids in the Americas and beyond. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18566–18570. doi: 10.1073/pnas.0705329104

Hemelaar, J. (2012). The origin and diversity of the Hiv-1 pandemic. *Trends Mol. Med.* 18, 182–192. doi: 10.1016/j.molmed.2011.12.001

Hemelaar, J. (2013). Implications of Hiv diversity for the Hiv-1 pandemic. *J. Infect.* 66, 391–400. doi: 10.1016/j.jinf.2012.10.026

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of Hiv-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis*. 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020a). Global and regional epidemiology of Hiv-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, E772–E781. doi: 10.1016/S2352-3018(20)30252-6

Hemelaar, J., Loganathan, S., Elangovan, R., Yun, J., Dickson-Tetteh, L., and Kirtley, S. (2020b). Country level diversity of the Hiv-1 pandemic between 1990 and 2015. *J. Virol.* 95, E01580–E01520. doi: 10.1128/JVI.01580-20

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary Hiv-1 variation. *Br. Med. Bull.* 58, 19–42. doi: 10.1093/bmb/58.1.19

Oster, A. M., France, A. M., and Mermin, J. (2018). Molecular epidemiology and the transformation of Hiv prevention. *JAMA* 319, 1657–1658. doi: 10.1001/jama.2018.1513

Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I., and Hatzakis, A. (2016). The application of Hiv molecular epidemiology to public health. *Infect. Genet. Evol*. 46, 159–168. doi: 10.1016/j.meegid.2016.06.021

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). Hiv-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Song, H., Ou, W., Feng, Y., Zhang, J., Li, F., Hu, J., et al. (2019). Disparate impact on Cd4 T cell count by two distinct Hiv-1 phylogenetic clusters from the same clade. *Proc. Natl. Acad. Sci. U.S.A.* 116, 239–244. doi: 10.1073/pnas.1814714116

Tebit, D. M., and Arts, E. J. (2011). Tracking a century of global expansion and evolution of Hiv to drive understanding and to combat disease. *Lancet Infect. Dis.* 11, 45–56. doi: 10.1016/S1473-3099(10)70186-9

UNAIDS. (2022). *Global AIDS Update*. Geneva: UNAIDS.

Vasylyeva, T. I., Zarebski, A., Smyrnov, P., Williams, L. D., Korobchuk, A., Liulchuk, M., et al. (2020). Phylodynamics helps to evaluate the impact of an Hiv prevention intervention. *Viruses* 12, 469. doi: 10.3390/v12040469

Vuilleumier, S., and Bonhoeffer, S. (2015). Contribution of recombination to the evolutionary history of Hiv. *Curr. Opin.* 10, 84–89. doi: 10.1097/COH.0000000000000137

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., et al. (2008). Direct evidence of extensive diversity of Hiv-1 in kinshasa by 1960. *Nature* 455, 661–664. doi: 10.1038/nature07390

Wymant, C., Bezemer, D., Blanquart, F., Ferretti, L., Gall, A., Hall, M., et al. (2022). A highly virulent variant of Hiv-1 circulating in the Netherlands. *Science* 375, 540–545. doi: 10.1126/science.abk1688

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/fmicb.2021.690647

# Global and Regional Estimates for Subtype-Specific Therapeutic and Prophylactic HIV-1 Vaccines: A Modeling Study

***Ramyiadarsini Elangovan[1†], Michael Jenks[1†], Jason Yun[1], Leslie Dickson-Tetteh[1], Shona Kirtley[2], Joris Hemelaar[1]* and WHO-UNAIDS Network for HIV Isolation and Characterisation[‡]***

*[1] Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom, [2] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, United Kingdom*

OPEN ACCESS

***Edited by:***
*Hirofumi Akari,*
*Kyoto University, Japan*

***Reviewed by:***
*Suvankar Ghorai,*
*Raiganj University, India*
*Hiroshi Ishii,*
*National Institute of Infectious, Japan*

****Correspondence:***
*Joris Hemelaar*
*joris.hemelaar@ndph.ox.ac.uk*

*[†]These authors have contributed equally to this work*

*[‡]Members are listed in Supplementary Material*

***Specialty section:***
*This article was submitted to HIV-1 Genetic Diversity, a section of the journal Frontiers in Microbiology*

***Received:*** *03 April 2021*
***Accepted:*** *27 May 2021*
***Published:*** *15 July 2021*

***Citation:***
*Elangovan R, Jenks M, Yun J, Dickson-Tetteh L, Kirtley S, Hemelaar J and WHO-UNAIDS Network for HIV Isolation and Characterisation (2021) Global and Regional Estimates for Subtype-Specific Therapeutic and Prophylactic HIV-1 Vaccines: A Modeling Study. Front. Microbiol. 12:690647. doi: 10.3389/fmicb.2021.690647*

Global HIV-1 genetic diversity forms a major obstacle to the development of an HIV vaccine. It may be necessary to employ subtype-specific HIV-1 vaccines in individual countries according to their HIV-1 subtype distribution. We estimated the global and regional need for subtype-specific HIV-1 vaccines. We took into account the proportions of different HIV-1 variants circulating in each country, the genetic composition of HIV-1 recombinants, and the different genome segments (*gag*, *pol*, *env*) that may be incorporated into vaccines. We modeled different scenarios according to whether countries would employ subtype-specific HIV-1 vaccines against (1) the most common subtype; (2) subtypes contributing more than 5% of HIV infections; or (3) all circulating subtypes. For therapeutic vaccines targeting the most common HIV-1 subtype in each country, 16.5 million doses of subtype C vaccine were estimated globally, followed by subtypes A (14.3 million) and B (4.2 million). A vaccine based on *env* required 2.6 million subtype E doses, and a vaccine based on *pol* required 4.8 million subtype G doses. For prophylactic vaccines targeting the most common HIV-1 subtype in each country, 1.9 billion doses of subtype A vaccine were estimated globally, followed by subtype C (1.1 billion) and subtype B (1.0 billion). A vaccine based on *env* required 1.2 billion subtype E doses, and a vaccine based on *pol* required 0.3 billion subtype G doses. If subtype-specific HIV-1 vaccines are also directed against less common subtypes in each country, vaccines targeting subtypes D, F, H, and K are also needed and would require up to five times more vaccine doses in total. We conclude that to provide global coverage, subtype-specific HIV-1 vaccines need to be directed against subtypes A, B, and C. Vaccines targeting *env* also need to include subtype E and those targeting *pol* need to include subtype G.

**Keywords: HIV, subtype, recombinant, CRF, URF, vaccine**

## INTRODUCTION

Thirty-eight million people globally were living with HIV in 2019 (UNAIDS, 2020). Despite the increased availability of antiretroviral therapy, there were 690,000 deaths and 1.7 million new infections in 2019 (UNAIDS, 2020). A globally effective preventative HIV vaccine is likely necessary to end the HIV pandemic (Fauci, 2017). Furthermore, a therapeutic vaccine that augments the immune system of HIV-infected individuals may reduce the need for antiretroviral therapy (Dorrell, 2005). However, a key stumbling block to the development of an HIV vaccine is the extensive global genetic diversity of HIV (Barouch, 2008; Hemelaar, 2012, 2013).

HIV has its origins in the zoonotic transmission of Simian Immunodeficiency Virus (SIV) from chimpanzees to humans a century ago (Gao et al., 1999). Subsequent to this, HIV-1 Group M diversified in Central Africa into multiple distinct subtypes: A, B, C, D, F, G, H, J, K, and L (Robertson et al., 2000; Worobey et al., 2008; Yamaguchi et al., 2020). The global spread of HIV throughout the second half of the twentieth century led to the differential global distribution of HIV-1 subtypes (Tebit and Arts, 2011; Hemelaar, 2012; Faria et al., 2014).

Genetic divergence between HIV-1 subtypes is substantial, with Env displaying a median difference of 25% (range 20–36%) at the amino acid level between strains from different subtypes, whereas the difference is 17% (15–22%) for Gag (Korber et al., 2001). Recombination between different HIV strains has led to further diversification of the HIV pandemic (Hemelaar, 2013). Recombinant forms are classified as circulating recombinant forms (CRFs) if they are found in three or more epidemiologically distinct individuals or unique recombinant forms (URFs) if there is no evidence of onward transmission (Robertson et al., 2000). To date, 106 distinct CRFs have been identified (Zhou et al., 2020), and collectively, these CRFs have been estimated to account for 16.7% of HIV-1 infections worldwide (Hemelaar et al., 2019).

The immune response to HIV is multifaceted, with antibodies mainly directed against the envelope component of the virus, whereas cytotoxic T lymphocyte responses are preferentially directed against Gag and/or Pol (Goulder et al., 2000). The large genetic divergence between HIV-1 subtypes makes it difficult to elicit immune responses that are sufficiently cross-reactive between HIV-1 subtypes (Korber et al., 2001). Given the variation between HIV-1 subtypes, it has been a common approach in HIV vaccine design to match the HIV-1 subtype(s) of the immunogen in the candidate vaccine to the HIV-1 subtype(s) circulating in the target population. To date, a number of different vaccine concepts, each using immunogen HIV-1 subtype(s) matched to circulating HIV-1 subtype(s), have been tested in large-scale efficacy trials.

Firstly, recombinant HIV-1 envelope proteins were used as immunogens aimed at generating broadly neutralizing antibodies. A bivalent subtype B/B recombinant glycoprotein gp120 vaccine was trialed in North America and The Netherlands, where subtype B dominates, and a bivalent subtype B/E recombinant gp120 vaccine was tested in Thailand, where subtype B and CRF01_AE cocirculate. Neither of these vaccines proved efficacious (HIV Vaccine Study Group et al., 2005; Pitisuttithum et al., 2006).

Next, viral vectors were used with the aim of eliciting cytotoxic T lymphocyte responses. The first such vaccine consisted of adenovirus type-5 (Ad5) vectors expressing subtype B Gag, Pol, and Nef proteins. This vaccine was tested in North America, the Caribbean, South America, and Australia, where subtype B is the predominant HIV-1 subtype, and in South Africa, where subtype C dominates. In both trials, the vaccine did not prevent HIV-1 infection or lower the viral-load setpoint (Buchbinder et al., 2008; Gray et al., 2011).

A subsequent approach aimed to elicit both antibody and T-cell responses. This strategy consisted of priming with DNA plasmids expressing subtype B Gag, Pol, and Nef and subtypes A, B, and C Env proteins, followed by a boost consisting of Ad5 vectors expressing a subtype B Gag-Pol fusion protein and Env glycoproteins of subtypes A, B, and C. When tested in MSM populations in the United States (mainly subtype B), the vaccine did not reduce the rate of HIV-1 acquisition or the viral-load set point (Hammer et al., 2013).

A further attempt aimed at eliciting both humoral and cell-mediated immunity used a combination of a canarypox vector expressing subtype B Env gp41TM, Gag and Pol and CRF01_AE Env gp120 followed by a boost with bivalent subtype B/E recombinant gp120 proteins, chosen to match the B and CRF01_AE strains circulating in Thailand. The RV144 trial showed modest efficacy of this vaccine (Rerks-Ngarm et al., 2009). This vaccine concept was then adapted for use in South Africa by replacing the B/CRF01_AE immunogens with subtype C immunogens to match HIV-1 subtype C endemic in South Africa (Bekker et al., 2018). However, the trial of this vaccine was recently halted due to lack of efficacy (Adepoju, 2020).

Given the genetic divergence between HIV-1 subtypes and their differential global spread, it may be necessary to employ subtype-specific HIV-1 vaccines in individual countries according to their HIV-1 subtype distribution. To aid prioritization of HIV-1 subtypes for vaccine development, we aimed to estimate the global and regional need for therapeutic and prophylactic vaccines specific for different HIV-1 subtypes, taking account of the proportions of different HIV-1 variants circulating in each country, the genetic composition of HIV-1 recombinants, and the different genome segments of HIV that may be incorporated into a vaccine.

## MATERIALS AND METHODS

### HIV-1 Molecular Epidemiology Data

Country-level HIV-1 molecular epidemiology data was collected by conducting a global survey and a comprehensive systematic review, as described previously (Hemelaar et al., 2019; Hemelaar et al., 2020). In total, 2,203 datasets with 383,519 samples were obtained from 116 countries across 1990–2015, with the data analyzed for four time periods: 2010–2015, 2005–2009, 2000–2004, and 1990–1999. In the current study, we used data from the latest time period (2010–2015) for analysis for most countries.

For the small number of countries for which no data for 2010–2015 was available, data from the next most recent time period (2005–2009, 2000–2004, or 1990–1999) was used. These latter countries, as well as the time period for which data was used, are listed in **Supplementary Material**, p. 7.

## Reassignment of CRFs to "Pure" HIV-1 Subtypes

Given the large number of different CRFs (106 distinct CRFs identified to date (Los Alamos National Laboratory; Zhou et al., 2020), it would be impractical to make a vaccine specific for each CRF. Hence, we determined which "pure" HIV-1 subtypes contributed most to each CRF, both to each genome segment (*gag*, *pol*, *env*) as well as the full-length genome (which also includes accessory genes (*vif*, *vpu*, *vpr*, *nef*), the regulatory genes (*rev*, *tat*), and the 5′ and 3′ long terminal repeat regions). Information on the genetic composition of individual CRFs was obtained from the Los Alamos National Laboratory (LANL) website (Los Alamos National Laboratory). CRFs were reassigned to "pure" HIV-1 subtypes according to the HIV-1 subtype making the largest contribution to the full-length genome as well as each genome segment. In situations where unclassified sequences made the largest contribution, the next largest contributing HIV-1 subtype was taken. If the subtype composition of certain genome regions was unclear from the LANL website, the original paper describing the CRF was examined. The full list of CRFs and their reassignment to the "pure" HIV-1 subtypes, for full length as well as each genome segment, can be found in **Supplementary Material**, pp. 4–6. Unfortunately, reassignment could not be performed for CRF30_0206, CRF75_01B, CRF77_cpx, CRF79_0107, CRF80_0107, CRF81_cpx, and CRF84_A1D due to lack of data on their subtype composition.

## HIV-1 Subtype Proportions in Countries After Reassignment of CRFs to "Pure" HIV-1 Subtypes

Upon completion of the reassignment of CRFs to "pure" HIV-1 subtypes, the proportions of infections accounted for by each CRF in each country, as previously estimated (Hemelaar et al., 2020), were added to those of the relevant "pure" HIV-1 subtypes, thereby generating new proportions of infections that could be ascribed to each "pure" HIV-1 subtype for the full-length genome and each genome segment. Country-level "pure" HIV-1 subtype distributions were combined with UNAIDS data on the number of people living with HIV in each country in 2016 (UNAIDS, 2017) to generate estimates of regional and global "pure" HIV-1 subtype proportions (**Supplementary Material**, p. 2).

## Estimation of Numbers of Doses for Subtype-Specific Therapeutic and Prophylactic HIV-1 Vaccines

Upon estimation of the proportions of "pure" HIV-1 subtypes in each country, estimates for the numbers of doses needed for either therapeutic or prophylactic subtype-specific HIV-1 vaccines were calculated. Calculations were performed for three different scenarios, each using a different cut-off for HIV-1 subtypes eligible for inclusion in vaccines for each country: (1) Vaccinating against only the most common subtype circulating in each country ("most common subtype" scenario), (2) Vaccinating against subtypes with a prevalence of >5% in people living with HIV in each country (">5% prevalence" scenario), and (3) Vaccinating against all circulating subtypes in each country ("all circulating subtypes" scenario). A fourth scenario was assessed for therapeutic vaccines, in which each HIV-infected individual would be vaccinated with a vaccine based on the HIV-1 subtype they were infected with. All calculations for each scenario were conducted for the full-length genome and each genome region (*gag*, *pol*, *env*).

For therapeutic vaccines, the target population was all people living with HIV in 2016, as estimated by UNAIDS (UNAIDS, 2017). For prophylactic vaccines, the target population was 10–49-year-old men and women, chosen to include most of the sexually active population as well as other high-risk groups, using estimates of population numbers in 2015 reported by the United Nations (United Nations, 2017). For both the therapeutic and prophylactic HIV-1 vaccine analyses, the entire target population in each country was to be vaccinated against every subtype that made the cut-off in each scenario, i.e., the most common subtype, all subtypes which contributed >5% of HIV infections, or all circulating subtypes. The estimated number of subtype-specific HIV-1 vaccine doses per country were subsequently aggregated at both the regional and global levels.

The term "dose" in this study was used to describe a "course of vaccination" with a subtype-specific HIV-1 vaccine. It may, however, be that a course of vaccination may consist of multiple doses of the same vaccine or a combination of different types of vaccines in a "prime-boost" configuration. All calculations were performed in Microsoft Excel.

# RESULTS

## Global and Regional Distribution of HIV-1 Subtypes

Given the difficulty in generating a vaccine for each individual HIV-1 recombinant, we first reassigned each CRF to a "pure" HIV-1 subtype (A-K) according to the HIV-1 subtype that contributed most to each CRF, both for the full length genome and the *gag*, *pol*, and *env* regions (**Supplementary Material**, pp. 4–6). Following reassignment of CRFs, we determined the global and regional proportions of infections caused by each of the "pure" HIV-1 subtypes, based on the most recent available HIV-1 subtype distribution data for each country (**Figures 1**, **2** and **Supplementary Material**, pp. 8–10).

After reassignment of CRFs based on the full-length genome, half of the global HIV infections were attributable to subtype C (49.1%), followed by subtype A (24.8%), B (12.0%), G (5.1%), D (2.5%), F (0.7%), and H, J, and K (0.1% each) (**Figure 1B** and **Supplementary Material**, pp. 8–10). Major changes in global HIV-1 subtype distribution resulting from reassignment of CRFs to "pure" HIV-1 subtypes were driven by the major recombinants CRF01_AE and CRF02_AG, as CRF01_AE is

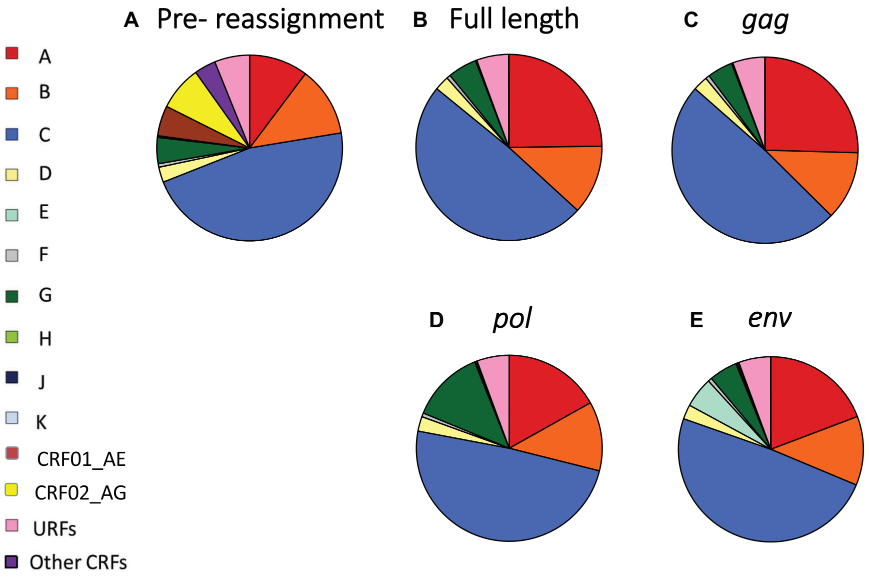


**FIGURE 1 |** Global distribution of HIV-1 variants before and after reassignment of CRFs to "pure" HIV-1 subtypes. **(A)** Global proportions of HIV-1 subtypes A–K, CRF01_AE, CRF02_AG, other CRFs and URFs, based on most recent data available for each country. **(B–E)** Global proportions of HIV-1 variants after reassignment of CRFs to "pure" HIV-1 subtypes, based on full-length sequence **(B)**, *gag* **(C)**, *pol* **(D)**, and *env* **(E)**. CRF, circulating recombinant form; URF, unique recombinant form. Data underlying this figure is displayed in **Supplementary Material**, pp. 8–10.

composed of subtype A in *gag* and *pol*, but subtype E in *env*, whereas CRF02_AG is composed of subtype A in *gag* and *env* and subtype G in *pol* (**Figure 1** and **Supplementary Material**, pp. 4–6 and 8–10).

Subtype A contributed 25.4% of global HIV infections after CRF reassignment based on *gag*, 19.2% for *env*, and 16.9% for *pol* (**Figures 1C–E** and **Supplementary Material**, pp. 9–10). Subtype E contributed 5.3% of global infections after reassignment based on *env*, but none for *gag* and *pol*, as subtype E has never been identified for those genome segments. Subtype G constituted 12.7% of infections after reassignment based on *pol*, but only 5.0% for *env* and 4.4% for *gag*. The global contributions of other subtypes remained relatively stable following CRF reassignment (**Figure 1** and **Supplementary Material**, pp. 8–10).

In South-East Asia, where CRF01_AE plays an important role, the proportion of infections attributable to subtype A was 74.4% after CRF reassignment based on the full-length analysis (**Figure 2A** and **Supplementary Material**, pp. 8–10). Subtype A also constituted 74.4 and 74.2% of infections for *gag* and *pol* in this region, whereas subtype E constituted 67.8% for *env* (**Figures 2B–D** and **Supplementary Material**, pp. 8–10). In East Asia, subtype A constituted 47.0% of infections for full length and *gag*, and 46.7% for *pol*. However, subtype E constituted 46.8% of infections for *env* in this region. In West Africa, where CRF02_AG plays a major role, subtype A constituted 52.6% of infections for full length and similar percentages for *gag* and *env*. However, subtype G constituted 78.7% for *pol*. In the other regions, which had fewer CRF infections, there was less change in HIV-1 subtype proportions following reassignment of CRFs to "pure" HIV-1 subtypes (**Figure 2** and **Supplementary Material**, pp. 8–10).

## Therapeutic HIV-1 Vaccines

If HIV-infected people would be vaccinated against the most common subtype circulating in each country ("most common subtype" scenario), based on the full-length genome, 35.1 million vaccine doses would be required globally, of which 16.5 million were subtype C, 14.3 million subtype A, and 4.2 million subtype B, with much fewer doses for other subtypes (**Figure 3A** and **Table 1**). A vaccine based on *env* required 2.6 million subtype E doses, and a vaccine based on *pol* required 4.8 million subtype G doses (**Figure 3A** and **Table 1**). The global need for a therapeutic subtype C vaccine was largely driven by Southern Africa and South Asia (**Figure 4A** and **Supplementary Material**, p. 11). The need for a subtype A vaccine was largely driven by East and West Africa, as well as South-East Asia and Eastern Europe and Central Asia. Finally, the global need for a subtype B vaccine was driven by Western and Central Europe and North America and Latin America (**Figure 4A** and **Supplementary Material**, p. 11).

If HIV-infected people would be vaccinated against subtypes with a prevalence of >5% in people living with HIV in each country (">5% prevalence" scenario), based on the full-length genome, 58.2 million vaccine doses were estimated, of which 22.7 million would be subtype C, 15.9 million subtype A, 6.5 million subtype B, 5.6 million subtype G, and 5.3 million subtype D (**Figure 3B** and **Table 1**). A vaccine based on *env* required 2.9 million subtype E doses.

If HIV-infected people would be vaccinated against all circulating subtypes in each country ("all circulating subtypes" scenario), 141.3 million doses of vaccine would be required, based on the full-length genome, of which 33.3 million would be subtype C, 32.9 million subtype A, 20.8 million subtype B, 17.7 million subtype F, 15.7 million subtype G, 14.1 million subtype

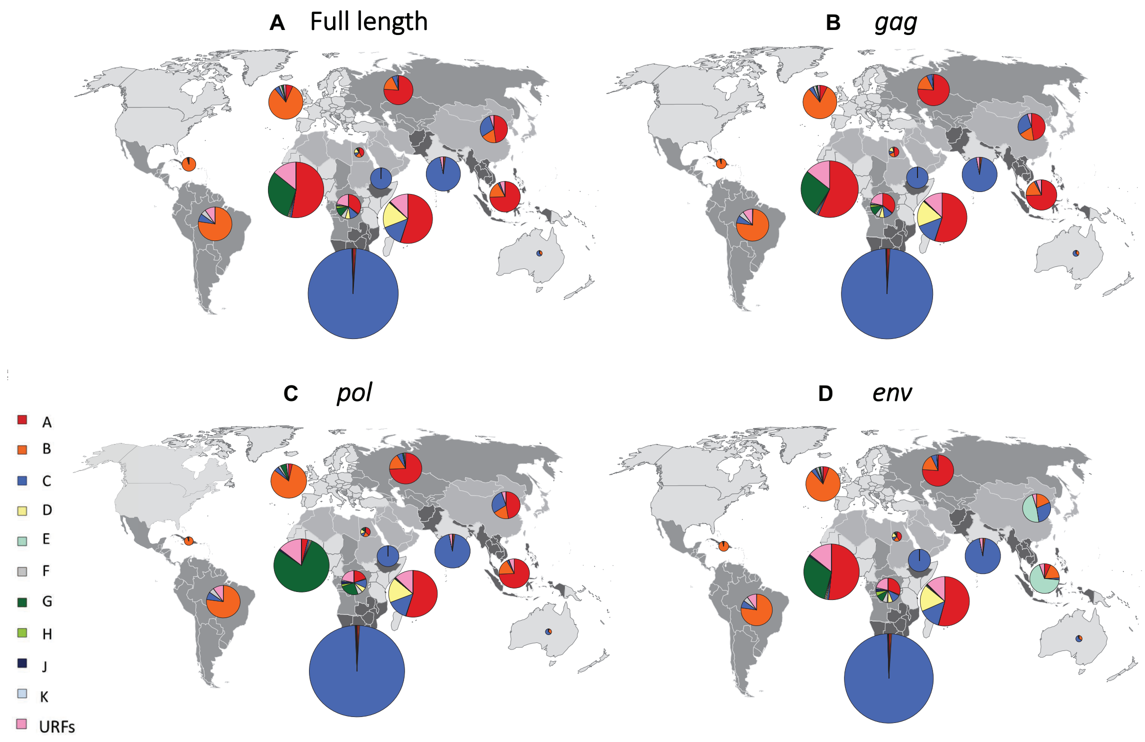


**FIGURE 2 |** Regional distribution of HIV-1 subtypes after reassignment of CRFs to "pure" HIV-1 subtypes. Regional proportions of HIV-1 subtypes after reassignment of CRFs to "pure" HIV-1 subtypes, based on full-length sequence **(A)**, *gag* **(B)**, *pol* **(C)**, and *env* **(D)**. We grouped all countries into 14 regions, as previously described (Hemelaar et al., 2019). Individual regions are shaded differently on the world map. The proportion of HIV infections attributed to each subtype in each region is shown in pie charts superimposed onto the regions. The sizes (surface area) of the pie charts for each region are proportional to the relative number of people living with HIV in each region. URF, unique recombinant form. Data underlying this figure is displayed in **Supplementary Material**, pp. 8–10.

D, 3.1 million subtype H, 2.7 million subtype J and 1.1 million subtype K (**Figure 3C** and **Table 1**). A vaccine based on *env* required 8.5 million subtype E doses.

In the final scenario in which each infected individual would be vaccinated only against the subtype with which they are already infected, 34.2 million vaccine doses were estimated, based on the full-length genome, of which 17.3 million would be subtype C, 9.3 million subtype A, 4.4 million subtype B, 1.9 million subtype G, and 1.0 million subtype D, reflecting the global distribution of HIV-1 variants (**Figures 1**, **3D**, **Table 1**, and **Supplementary Material**, 8–10).

### Prophylactic HIV-1 Vaccines

If all 10–49-year-old people would be vaccinated against the most common subtype circulating in each country ("most common subtype" scenario), based on the full-length genome, an estimated 4.1 billion doses of vaccines would be required globally, of which 1.9 billion were subtype A, 1.1 billion subtype C, and 1.0 billion subtype B (**Figure 3E** and **Table 1**). A vaccine based on *env* required 1.2 billion subtype E doses, and a vaccine based on *pol* required 262 million subtype G doses. The global need for a prophylactic subtype A vaccine was largely driven by East Asia and South-East Asia (**Figure 4B** and **Supplementary Material**, p. 11). This was due to their large populations as well as the endemic nature of CRF01_AE in these regions. The need for a subtype C vaccine was driven largely by South Asia, and the need for a subtype B vaccine was driven by Western and Central Europe and North America and Latin America.

In the ">5% prevalence" scenario, 8.2 billion doses of a vaccine based on the full-length genome were estimated, of which 2.5 billion would be subtype A, 2.5 billion subtype C, and 2.3 billion subtype B (**Figure 3F** and **Table 1**). A vaccine based on *env* required 1.6 billion subtype E doses.

Finally, in the "all circulating subtypes" scenario, for a vaccine based on the full-length genome, 17.5 billion vaccine doses would be required, of which 3.9 billion would be subtype A, 3.8 billion subtype C, 3.6 billion subtype B, 2.2 billion subtype G, 1.9 billion subtype F, 1.3 billion subtype D, and 0.4 billion for subtype H (**Figure 3G** and **Table 1**). A vaccine based on *env* required 2.1 billion subtype E doses.

## DISCUSSION

In this study, we estimated the global and regional need for subtype-specific therapeutic and prophylactic HIV-1 vaccines. When targeting the most common HIV-1 subtype in each country, we estimated the largest number of therapeutic vaccine doses were needed for a subtype C vaccine (16.5 million), followed by subtype A (14.3 million) and subtype B (4.2 million).

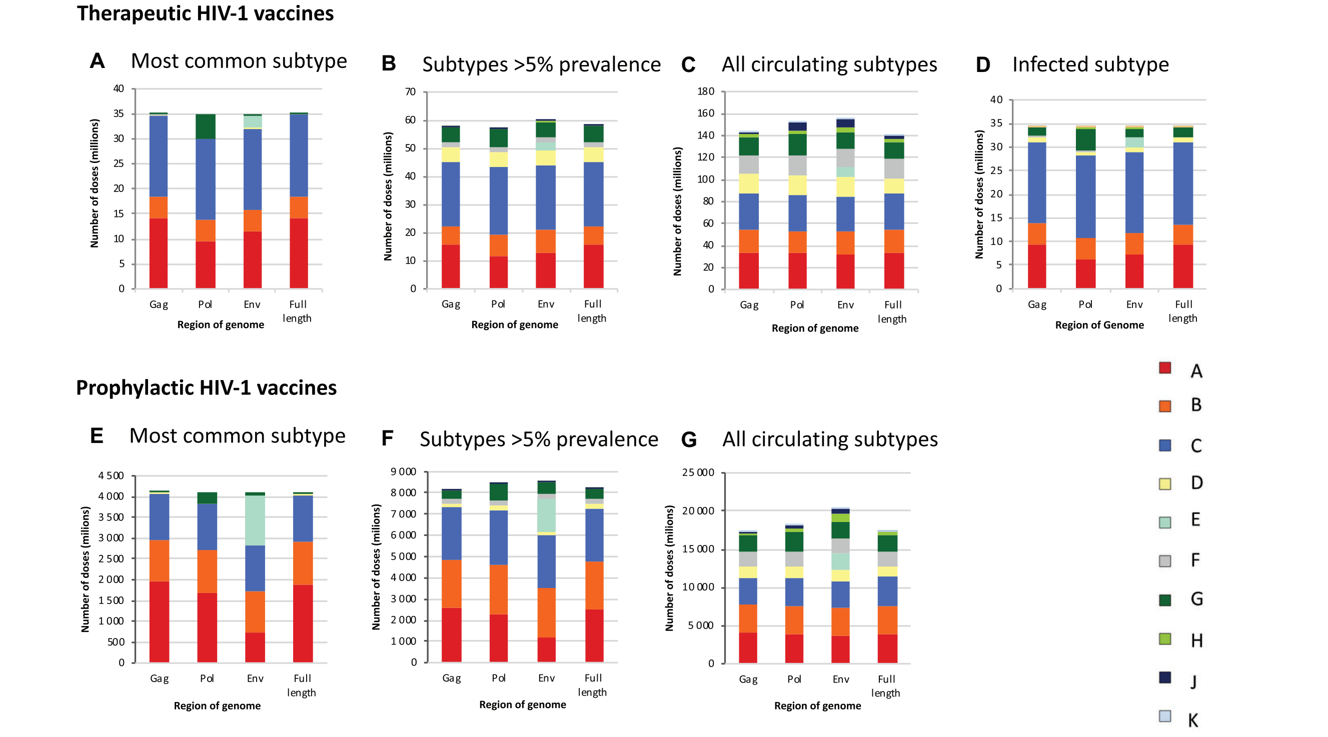


**FIGURE 3 |** Global estimates of the number of doses of subtype-specific therapeutic and prophylactic HIV-1 vaccines. Estimates of the number of doses of subtype-specific therapeutic **(A–D)** and prophylactic **(E–G)** HIV-1 vaccines. Estimates are stratified according to genome region (*gag*, *pol*, *env*, and full length) and the number of subtypes against which people in each country are vaccinated, according to different scenarios, i.e., vaccinating people against the most common subtype in each country **(A,E)**, subtypes with a prevalence of >5% in people living with HIV in each country **(B,F)**, all subtypes circulating in each country **(C,G)**, and the HIV-1 subtype with which HIV-positive people are infected **(D)**. Data underlying this figure is displayed in **Table 1**.

A vaccine based on *env* required 2.6 million subtype E doses, and a vaccine based on *pol* required 4.8 million subtype G doses. The need for therapeutic subtype C vaccines was largely driven by the endemicity of subtype C in Southern Africa and South Asia, the need for subtype A vaccines by East and West Africa, and the need for subtype B vaccines by Western and Central Europe and North America and Latin America.

For prophylactic vaccines targeting the most common HIV-1 subtype in each country, 1.9 billion doses of subtype A vaccine were estimated, followed by subtype C (1.1 billion) and subtype B (1.0 billion). A vaccine based on *env* required 1.2 billion subtype E doses, and a vaccine based on *pol* required 0.3 billion subtype G doses. The need for prophylactic subtypes A and E vaccines was largely driven by East Asia and South-East Asia, owing to their large populations as well as prevalence of CRF01_AE. The need for a prophylactic subtype C vaccine was largely driven by South Asia.

Employing vaccines against more than one HIV-1 subtype in each country, as estimated in the ">5% prevalence" and "all circulating subtypes" scenarios, dramatically increases the number of vaccine doses and number of different subtype-specific vaccines required for both therapeutic and prophylactic vaccines.

It is apparent that to provide global coverage against the most common HIV-1 subtype circulating in each country, subtype-specific therapeutic and prophylactic HIV-1 vaccines need to be directed against subtypes A, B, and C. Vaccines targeting the envelope protein would also need to include subtype E and those targeting Pol need to include subtype G. If subtype-specific vaccines are also directed against less common HIV-1 subtypes in each country, vaccines targeting subtypes D, F, H, and K also need to be considered and would require up to five times more vaccine doses in total.

This study has several strengths. We utilized the largest available global HIV-1 molecular epidemiology database and applied the novel approach of reassigning CRFs to "pure" HIV-1 subtypes based on their genetic composition. This enabled us to address the complexity posed by multiple distinct recombinants and to generate new estimates of the proportion of infections caused by each "pure" HIV-1 subtype. In addition, because HIV-1 vaccines could aim to elicit antibodies or T-cell responses or both, we generated estimates for subtype-specific HIV-1 vaccines based on *env*, *gag*, and *pol* as well as the full-length genome. Moreover, we conducted separate analyses for therapeutic and prophylactic vaccines. In addition, we examined a number of different scenarios according to the number of HIV-1 subtypes eligible for inclusion in a vaccine.

Our study also has some limitations. Although CRFs were reassigned to "pure" HIV-1 subtypes based on their genetic composition, we do not know the extent of their overlapping immunogenic properties. Although the target population for

**TABLE 1 |** Global estimates of numbers of doses (millions) of subtype-specific therapeutic and prophylactic HIV-1 vaccines.

| | | | HIV-1 Subtypes | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Vaccine Type** | **Scenario** | **Genome Region** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **J** | **K** | **Total** |
| Therapeutic | Most common | *gag* | 14.28 | 4.21 | 16.21 | 0.06 | 0.00 | 0.02 | 0.04 | 0.00 | 0.00 | 0.00 | 34.81 |
| | | *pol* | 9.54 | 4.22 | 16.21 | 0.06 | 0.00 | 0.02 | 4.78 | 0.00 | 0.00 | 0.00 | 34.81 |
| | | *env* | 11.65 | 4.23 | 16.21 | 0.06 | 2.60 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 34.81 |
| | | Full length | 14.25 | 4.22 | 16.50 | 0.06 | 0.00 | 0.02 | 0.06 | 0.00 | 0.00 | 0.00 | 35.10 |
| | >5% | *gag* | 15.94 | 6.47 | 22.73 | 5.34 | 0.00 | 1.62 | 5.11 | 0.28 | 0.20 | 0.00 | 57.69 |
| | | *pol* | 11.67 | 7.75 | 23.89 | 5.34 | 0.00 | 1.62 | 6.53 | 0.28 | 0.32 | 0.00 | 57.38 |
| | | *env* | 13.00 | 7.75 | 23.14 | 5.28 | 2.92 | 1.62 | 5.70 | 0.31 | 0.32 | 0.00 | 60.02 |
| | | Full length | 15.94 | 6.52 | 22.74 | 5.28 | 0.00 | 1.63 | 5.63 | 0.28 | 0.20 | 0.00 | 58.24 |
| | All subtypes | *gag* | 33.12 | 20.78 | 32.73 | 18.11 | 0.00 | 17.82 | 15.20 | 3.14 | 2.27 | 0.70 | 143.89 |
| | | *pol* | 32.71 | 20.78 | 32.73 | 18.11 | 0.00 | 17.89 | 18.70 | 3.05 | 8.05 | 1.09 | 153.12 |
| | | *env* | 31.20 | 20.89 | 32.11 | 17.96 | 8.54 | 17.81 | 14.22 | 4.02 | 8.10 | 1.09 | 155.94 |
| | | Full length | 32.86 | 20.78 | 33.31 | 14.08 | 0.00 | 17.69 | 15.66 | 3.14 | 2.68 | 1.07 | 141.27 |
| | Infected subtype | *gag* | 9.50 | 4.34 | 17.33 | 1.03 | 0.00 | 0.25 | 1.62 | 0.04 | 0.04 | 0.02 | 34.17 |
| | | *pol* | 6.37 | 4.36 | 17.33 | 1.03 | 0.00 | 0.24 | 4.66 | 0.04 | 0.12 | 0.02 | 34.17 |
| | | *env* | 7.33 | 4.38 | 17.33 | 1.02 | 1.85 | 0.25 | 1.81 | 0.07 | 0.12 | 0.02 | 34.17 |
| | | Full length | 9.26 | 4.36 | 17.33 | 1.02 | 0.00 | 0.25 | 1.85 | 0.04 | 0.04 | 0.02 | 34.17 |
| Prophylactic | Most common | *gag* | 1,968.11 | 1,004.82 | 1,095.35 | 23.24 | 0.00 | 2.97 | 1.34 | 0.00 | 0.00 | 0.00 | 4,095.84 |
| | | *pol* | 1,702.24 | 1,010.52 | 1,095.35 | 23.24 | 0.00 | 2.97 | 261.52 | 0.00 | 0.00 | 0.00 | 4,095.84 |
| | | *env* | 718.02 | 1,012.00 | 1,095.35 | 23.24 | 1,180.43 | 1.50 | 65.31 | 0.00 | 0.00 | 0.00 | 4,095.84 |
| | | Full length | 1,898.45 | 1,010.52 | 1,116.87 | 23.24 | 0.00 | 2.97 | 65.31 | 0.00 | 0.00 | 0.00 | 4,117.36 |
| | >5% | *gag* | 2,561.34 | 2,256.68 | 2,486.53 | 203.94 | 0.00 | 253.16 | 356.69 | 21.52 | 8.40 | 0.00 | 8,148.27 |
| | | *pol* | 2,298.97 | 2,345.43 | 2,567.32 | 203.94 | 0.00 | 253.16 | 748.03 | 21.52 | 19.02 | 0.00 | 8,457.38 |
| | | *env* | 1,160.36 | 2,345.43 | 2,481.26 | 196.47 | 1,557.26 | 253.16 | 489.63 | 27.85 | 19.02 | 0.00 | 8,530.44 |
| | | Full length | 2,522.69 | 2,269.40 | 2,491.46 | 180.70 | 0.00 | 258.09 | 477.17 | 21.52 | 8.40 | 0.00 | 8,229.43 |
| | All subtypes | *gag* | 4,033.34 | 3,627.84 | 3,660.23 | 1,443.14 | 2.46 | 1,952.32 | 2,096.18 | 349.77 | 84.54 | 71.15 | 17,320.97 |
| | | *pol* | 3,907.84 | 3,627.84 | 3,660.23 | 1,443.14 | 2.46 | 2,007.39 | 2,685.46 | 345.61 | 457.35 | 158.67 | 18,296.00 |
| | | *env* | 3,643.27 | 3,637.83 | 3,604.44 | 1,407.96 | 2,112.43 | 1,950.97 | 2,190.60 | 1,200.43 | 457.57 | 157.98 | 20,363.47 |
| | | Full length | 3,942.57 | 3,627.84 | 3,814.19 | 1,274.02 | 2.46 | 1,945.75 | 2,243.83 | 362.80 | 151.96 | 129.63 | 17,495.05 |

*Estimates are stratified according to genome region (gag, pol, env, and full length) and the number of subtypes against which people in each country are vaccinated, according to different scenarios, i.e., vaccinating people against the most common subtype in each country ("most common"), subtypes with a prevalence of > 5% in people living with HIV in each country (">5%"), all subtypes circulating in each country ("all subtypes"), and the HIV-1 subtype with which HIV-positive people are infected (infected subtype).*

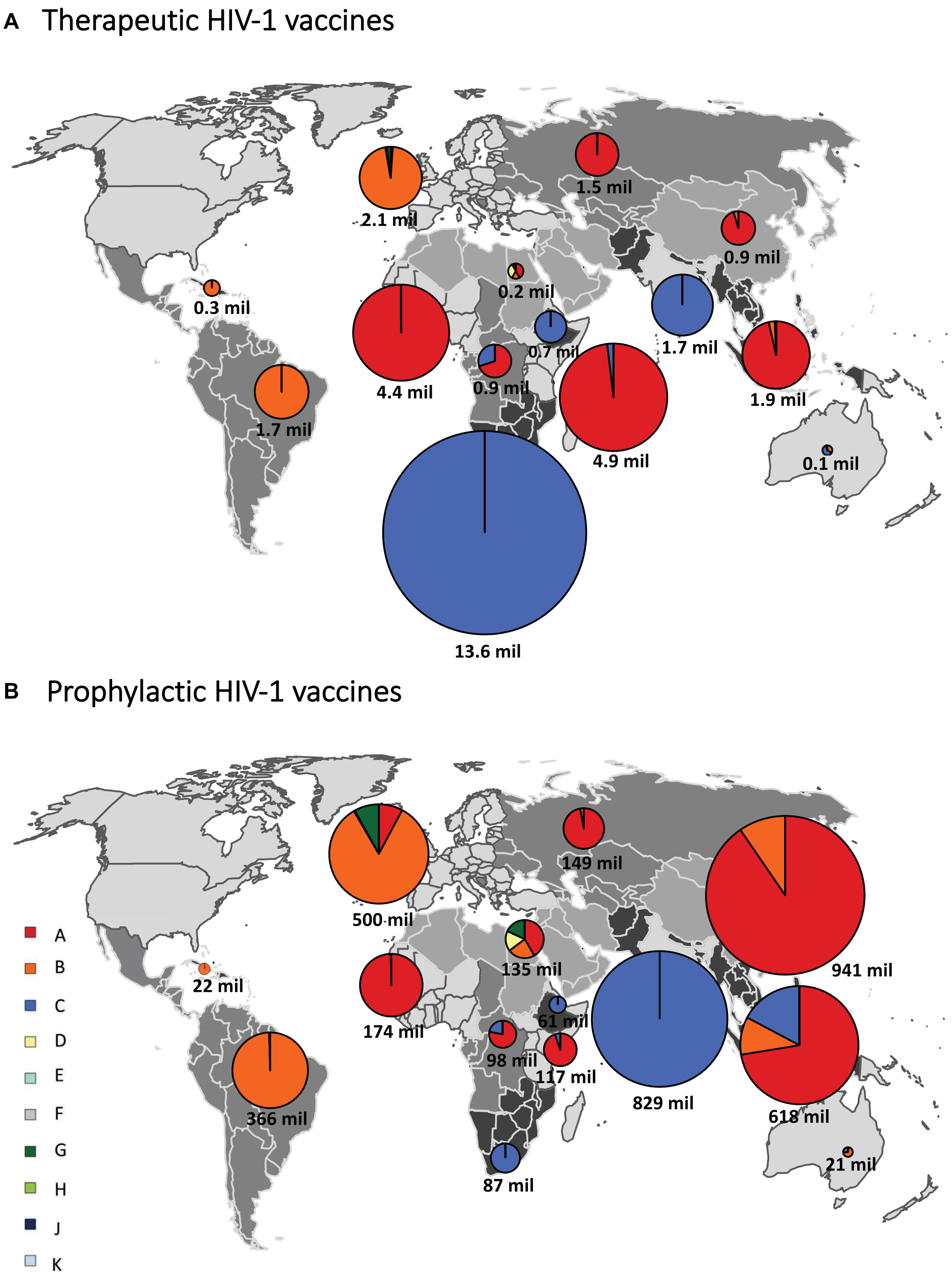


**FIGURE 4 |** Regional estimates of the number of doses of subtype-specific therapeutic and prophylactic HIV-1 vaccines. Regional estimates for therapeutic **(A)** and prophylactic **(B)** HIV-1 vaccines based on the most common subtype, based on the full-length sequence, in each country. We grouped all countries into 14 regions, as previously described (Hemelaar et al., 2019). Individual regions are shaded differently on the world map. The estimates for vaccines based on each HIV-1 subtype are shown in pie charts superimposed onto the regions. The sizes of the pie charts (surface area) correspond to the relative number of vaccine doses for each region and the total number of vaccine doses is shown below each pie chart. mil, million. Data underlying this figure is displayed in **Supplementary Material**, p. 11.

a therapeutic vaccine are people living with HIV, the target population for a prophylactic vaccine is less certain. We opted to estimate a one-off vaccination of all people aged 10–49 years old, to include most sexually active people and other risk groups (Marzetta et al., 2010). We made no distinction between routine and catch-up vaccinations (Marzetta et al., 2010). This comprehensive approach may have led to higher estimates of numbers of doses needed, but did allow us to gauge the relative importance of HIV-1 subtypes for regional and global vaccine development to enable prioritization of relevant HIV-1 subtypes. Vaccine efficacy may differ between subtype-specific vaccines and was not factored in Dimitrov et al. (2015). Moreover, a putative vaccine with high efficacy would likely be administered to larger populations whereas a vaccine with low/moderate efficacy would more likely be limited to high risk groups (Esparza et al., 2003). Duration of protection offered by a putative vaccine was not modeled and consequently revaccination was also not factored in. Furthermore, we aimed to estimate the need for subtype-specific vaccines and did not estimate the actual uptake or use, which depends on factors such as adoption time, accessibility and acceptability, which will vary by country (Marzetta et al., 2010). Lastly, cost and cost-effectiveness were not taken into account.

There are also limitations to the concept of subtype-specific HIV-1 vaccines. One issue is the need to generate multiple different vaccines for the different HIV-1 subtypes. This could be partially overcome by formulation of multivalent vaccines ("cocktails") incorporating multiple subtype-specific preparations (Korber et al., 2017). Another limitation is that a subtype-specific HIV-1 vaccine would need to be matched to locally circulating strains, which requires availability of up-to-date HIV-1 diversity data and the relevant subtype-specific HIV-1 vaccines. Furthermore, protection would be limited to a certain geographical region and thereby limit travel to other regions, while at the same time leave vaccinated people vulnerable to infection by newly imported strains of HIV, as HIV-1 subtype distribution is very dynamic (Hemelaar et al., 2019).

A crucial outstanding limitation of subtype-specific HIV-1 vaccines is the issue of intrasubtype genetic diversity (Korber et al., 2001; Gaschen et al., 2002). The HIV-1 vaccine recently tested, and proven ineffective, in South Africa consisted of a recombinant canarypox vaccine, which contained a subtype C *env* gp120 isolate sequence (96ZM651 from Zambia), and a bivalent subtype C gp120 consisting of two distinct subtype C recombinant monomeric Env gp120 proteins (derived from isolates TV1.C from South Africa and 1086.C from Malawi) (Zambonelli et al., 2016; Bekker et al., 2018). Utilization of subtype C isolate sequences in the vaccine, matching subtype C dominating in South Africa, was hoped to lead to protective immunogenicity. However, intrasubtype diversity is considerable, with median percentage amino acid differences within HIV-1 subtypes estimated at 17% (range 4–30%) for Env and 8% (2–15%) for Gag (Korber et al., 2001), thereby limiting the potential for eliciting cross-reactive protective immune responses. One way to reduce the genetic distance between vaccine immunogens and circulating strains is the inclusion of artificial centralized sequences, such as consensus, ancestral or center-of tree sequences (Gaschen et al., 2002; Nickle et al., 2003). For example, subtype C isolate sequences are around 5–15% different to other subtype C isolate sequences, whereas, a subtype C consensus amino acid sequence is only around 3–8% different from individual subtype C isolates (Gaschen et al., 2002). Of note, a group M consensus sequence (i.e., a consensus of all subtype consensus sequences) would be around 5–15% different to individual circulating HIV-1 isolates and therefore not better than a subtype-matched isolate sequence (Gaschen et al., 2002). Indeed, the use of isolate sequences in all candidate HIV vaccines tested in phase 3 trials to date may have limited cross-reactivity and hence limited efficacy (HIV Vaccine Study Group et al., 2005; Pitisuttithum et al., 2006; Buchbinder et al., 2008; Rerks-Ngarm et al., 2009; Gray et al., 2011; Hammer et al., 2013; Bekker et al., 2018). Thus, the use of subtype consensus (or ancestral or center-of-tree) sequences may be a more successful approach in the future.

Ideally, a globally effective HIV vaccine will need to confer protection against all diverse HIV-1 subtypes and recombinants. There are multiple HIV vaccine efforts on-going, utilizing a number of different approaches to address HIV-1 diversity. One approach is to use mosaic vaccines, which have been shown to elicit increased breadth and depth of immune responses (Barouch et al., 2010). The HVTN 705/Imbokodo trial currently underway in southern Africa is evaluating a tetravalent vaccine composed of adenovirus serotype 26 vector expressing mosaic *gag, pol* and *env* inserts combined with subtype C gp140 Env protein, with the intention of eliciting responses against a wide range of HIV subtypes, but still matching subtype C predominant in the region. The HVTN 706/Mosaico trial taking place in North America, Western Europe, and Latin America evaluates a nearly identical mosaic vaccine, which also includes a mosaic gp140 glycoprotein (Baden et al., 2020). Other approaches in preclinical development include focussing on conserved or structurally important regions of HIV (Letourneau et al., 2007; Gaiha et al., 2019). For all HIV-1 vaccine approaches, it is crucial to have up-to-date knowledge of HIV-1 genetic diversity to allow prioritization and development of vaccine concepts that are likely to provide the greatest benefit to specific countries, regions, and the world.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by local ethics committees of contributing studies. The participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

RE and MJ conducted the analyses, prepared the figures and tables, interpreted the data, and wrote the first draft of the manuscript. RE, JY, LD-T, SK, and JH collected the data. JH conceived, designed and coordinated the study, designed the analysis and figures, interpreted the data, and wrote the manuscript. All authors read and approved the final version of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.690647/full#supplementary-material

## REFERENCES

Adepoju, P. (2020). Moving on from the failed HIV vaccine clinical trial. *Lancet HIV* 7:e161. doi: 10.1016/s2352-3018(20)30047-3

Baden, L. R., S9itieh, D. J., Sarnecki, M., Walsh, S. R., Tomaras, G. D., Kublin, J. G., et al. (2020). Safety and immunogenicity of two heterologous HIV vaccine regimens in healthy, HIV-uninfected adults (TRAVERSE): a randomised, parallel-group, placebo-controlled, double-blind, phase 1/2a study. *Lancet HIV* 7, e688–e698. doi: 10.1016/s2352-3018(20)30229-0

Barouch, D. H. (2008). Challenges in the development of an HIV-1 vaccine. *Nature* 455, 613–619. doi: 10.3109/9781420060744-59

Barouch, D. H., O'Brien, K. L., Simmons, N. L., King, S. L., Abbink, P., Maxfield, L. F., et al. (2010). Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. *Nat. Med.* 16, 319–323. doi: 10.1038/nm.2089

Bekker, L.-G., Moodie, Z., Grunenberg, N., Laher, F., Tomaras, G. D., Cohen, K. W., et al. (2018). Subtype C ALVAC-HIV and bivalent subtype C gp120/MF59 HIV-1 vaccine in low-risk, HIV-uninfected, South African adults: a phase 1/2 trial. *Lancet HIV* 5, e366–e378.

Buchbinder, S. P., Buchbinder, S. P., Mehrotra, D. V., Duerr, A., Fitzgerald, D. W., Mogg, R., et al. (2008). Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet* 372, 1881–1893. doi: 10.1016/s0140-6736(08)61591-3

Dimitrov, D., Kublin, J. G., Ramsey, S., and Corey, L. (2015). Are clade specific HIV vaccines a necessity? An analysis based on mathematical models. *EBioMedicine* 2, 2062–2069. doi: 10.1016/j.ebiom.2015.11.009

Dorrell, L. (2005). Therapeutic immunization strategies for the control of HIV-1. *Expert Rev. Vaccines* 4, 513–520. doi: 10.1586/14760584.4.4.513

Esparza, J., Chang, M. L., Widdus, R., Madrid, Y., Walker, N., and Ghys, P. D. (2003). Estimation of "needs" and "probable uptake" for HIV/AIDS preventive vaccines based on possible policies and likely acceptance (a WHO/UNAIDS/IAVI study). *Vaccine* 21, 2032–2041. doi: 10.1016/s0264-410x(02)00775-2

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346, 56–61. doi: 10.1126/science.1256739

Fauci, A. S. (2017). An HIV vaccine is essential for ending the HIV/AIDS pandemic. *J. Am. Med. Assoc.* 318, 1535–1536. doi: 10.1001/jama.2017.13505

Gaiha, G. D., Rossin, E. J., Urbach, J., Landeros, C., Collins, D. R., Nwonu, C., et al. (2019). Structural topology defines protective CD8(+) T cell epitopes in the HIV proteome. *Science* 364, 480–484. doi: 10.1126/science.aav5095

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., et al. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 397, 436–441. doi: 10.1038/17130

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., et al. (2002). Diversity considerations in HIV-1 vaccine selection. *Science* 296, 2354–2360. doi: 10.1126/science.1070441

Goulder, P. J., Brander, C., Annamalai, K., Mngqundaniso, N., Govender, U., Tang, Y., et al. (2000). Differential narrow focusing of immunodominant human immunodeficiency virus gag-specific cytotoxic T-lymphocyte responses in infected African and caucasoid adults and children. *J. Virol.* 74, 5679–5690. doi: 10.1128/jvi.74.12.5679-5690.2000

Gray, G. E., Allen, M., Moodie, Z., Churchyard, G., Bekker, L. G., Nchabeleng, M., et al. (2011). Safety and efficacy of the HVTN 503/Phambili study of a clade-B-based HIV-1 vaccine in South Africa: a double-blind, randomised, placebo-controlled test-of-concept phase 2b study. *Lancet Infect. Dis.* 11, 507–515. doi: 10.1016/s1473-3099(11)70098-6

Hammer, S. M., Sobieszczyk, M. E., Janes, H., Karuna, S. T., Mulligan, M. J., Grove, D., et al. (2013). Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *N. Engl. J. Med.* 369, 2083–2092. doi: 10.1056/NEJMoa1310566

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* 18, 182–192. doi: 10.1016/j.molmed.2011.12.001

Hemelaar, J. (2013). Implications of HIV diversity for the HIV-1 pandemic. *J. Infect.* 66, 391–400. doi: 10.1016/j.jinf.2012.10.026

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* 19, 143–155.

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020). Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781.

HIV Vaccine Study Group, Flynn, N. M., Forthal, D. N., Harro, C. D., Judson, F. N., Mayer, K. H., et al. (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J. Infect. Dis.* 191, 654–665. doi: 10.1086/428404

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.* 58, 19–42. doi: 10.1093/bmb/58.1.19

Korber, B., Hraber, P., Wagh, K., and Hahn, B. H. (2017). Polyvalent vaccine approaches to combat HIV-1 diversity. *Immunol. Rev.* 275, 230–244. doi: 10.1111/imr.12516

Letourneau, S., Im, E. J., Mashishi, T., Brereton, C., Bridgeman, A., Yang, H., et al. (2007). Design and pre-clinical evaluation of a universal HIV-1 vaccine. *PLoS One* 2:e984. doi: 10.1371/journal.pone.0000984

Los Alamos National Laboratory *HIV Sequence Database*. Los Alamos, NM: Los Alamos National Laboratory. Available online at: http://www.hiv.lanl.gov/

Marzetta, C. A., Lee, S. S., Wrobel, S. J., Singh, K. J., Russell, N., and Esparza, J. (2010). The potential global market size and public health value of an HIV-1 vaccine in a complex global market. *Vaccine* 28, 4786–4797.

Nickle, D. C., Jensen, M. A., Gottlieb, G. S., Shriner, D., Learn, G. H., Rodrigo, A. G., et al. (2003). Consensus and ancestral state HIV vaccines. *Science* 299, 1515–1518.

Pitisuttithum, P9., Gilbert, P., Gurwith, M., Heyward, W., Martin, M., van Griensven, F., et al. (2006). Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. *J. Infect. Dis.* 194, 1661–1671. doi: 10.1086/508748

Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., et al. (2009). Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N. Engl. J. Med.* 361, 2209–2220.

Robertson, D. L. et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56.

Tebit, D. M., and Arts, E. J. (2011). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infectious Diseases* 11, 45–56. doi: 10.1016/s1473-3099(10)70186-9

UNAIDS (2017). *Global AIDS update 2017. Ending AIDS: Progress Towards the 90-90-90 Targets*. Geneva: UNAIDS.

UNAIDS (2020). *Global AIDS Update*. Geneva: UNAIDS.

United Nations (2017). *Department of Economic and Social Affairs, Population Division (2017): World Population Prospects 2017 – Data Booklet (ST/ESA/SER.A/401)*. New York, NY: United Nations.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., et al. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455, 661–664. doi: 10.1038/nature07390

Yamaguchi, J., Vallari, A., McArthur, C., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2020). Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *J. Acquir. Immune Defic. Syndr.* 83, 319–322. doi: 10.1097/qai.0000000000002246

Zambonelli, C., Dey, A. K., Hilt, S., Stephenson, S., Go, E. P., Clark, D. F., et al. (2016). Generation and characterization of a bivalent HIV-1 subtype C gp120 protein boost for proof-of-concept HIV vaccine efficacy trials in southern Africa. *PLoS one* 11:e0157391. doi: 10.1371/journal.pone.0157391

Zhou, J., Li, M., Min, C., Ma, Y., Shao, Y., and Xing, H. (2020). Near full-length genomic characterization of a novel HIV-1 circulating recombinant form (CRF106_cpx) identified among heterosexuals in China. *AIDS Res. Hum. Retroviruses* 36, 875–880. doi: 10.1089/aid.2020.0101

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 07 October 2021
doi: 10.3389/fmicb.2021.734929

# Increased Frequency of Inter-Subtype HIV-1 Recombinants Identified by Near Full-Length Virus Sequencing in Rwandan Acute Transmission Cohorts

**Gisele Umviligihozo[1†], Erick Muok[1†], Emmanuel Nyirimihigo Gisa[1], Rui Xu[2], Dario Dilernia[2], Kimberley Herard[2], Heeyah Song[2], Qianhong Qin[2], Jean Bizimana[1], Paul Farmer[2], Jonathan Hare[3], Jill Gilmour[4], Susan Allen[5], Etienne Karita[1], Eric Hunter[2,5‡*] and Ling Yue[2‡*]**

*[1] Centre for Family Health Research, Kigali, Rwanda, [2] Emory Vaccine Center at Yerkes National Primate Research Center, Atlanta, GA, United States, [3] IAVI, New York, NY, United States, [4] Faculty of Medicine, Imperial College London, London, United Kingdom, [5] Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, United States*

Most studies of HIV-1 transmission have focused on subtypes B and C. In this study, we determined the genomic sequences of the transmitted founder (TF) viruses from acutely infected individuals enrolled between 2005 and 2011 into IAVI protocol C in Rwanda and have compared these isolates to viruses from more recent (2016–2019) acute/early infections in three at risk populations – MSM, high risk women (HRW), and discordant couples (DC). For the Protocol C samples, we utilized near full-length single genome (NFLG) amplification to generate 288 HIV-1 amplicons from 26 acutely infected seroconverters (SC), while for the 21 recent seroconverter samples (13 from HRW, two from DC, and six from MSM), we PCR amplified overlapping half-genomes. Using PacBio SMRT technology combined with the MDPseq workflow, we performed multiplex sequencing to obtain high accuracy sequences for each amplicon. Phylogenetic analyses indicated that the majority of recent transmitted viruses from DC and HRW clustered within those of the earlier Protocol C cohort. However, five of six sequences from the MSM cohort branched together and were greater than 97% identical. Recombination analyses revealed a high frequency (6/26; 23%) of unique inter-subtype recombination in Protocol C with 19% AC and 4% CD recombinant viruses, which contrasted with only 6.5% of recombinants defined by sequencing of the *pol* gene previously. The frequency of recombinants was significantly higher (12/21; 57%) in the more recent isolates, although, the five related viruses from the MSM cohort had identical recombination break points. While major drug resistance mutations were absent from Protocol C viruses, 4/21 of recent isolates exhibited transmitted nevirapine resistance. These results demonstrate the ongoing evolution and increased prevalence of recombinant and drug resistant transmitted viruses in Rwanda and highlight the importance of defining NFLG sequences to fully understand the nature

OPEN ACCESS

***Edited by:***
*Kok Keng Tee,*
*University of Malaya, Malaysia*

***Reviewed by:***
*Sudhanshu Shekhar Pandey,*
*Trivitron healthcare Ltd., India*
*Marcel Tongo,*
*Center for Research on Emerging and Re-Emerging Diseases (CREMER), Cameroon*

****Correspondence:***
*Ling Yue*
*lyue2@emory.edu*
*Eric Hunter*
*ehunte4@emory.edu*

*[†] These authors have contributed equally to this work and share first authorship*

*[‡] These authors have contributed equally to this work and share senior authorship*

***Specialty section:***
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *01 July 2021*
***Accepted:*** *07 September 2021*
***Published:*** *07 October 2021*

***Citation:***
*Umviligihozo G, Muok E, Nyirimihigo Gisa E, Xu R, Dilernia D, Herard K, Song H, Qin Q, Bizimana J, Farmer P, Hare J, Gilmour J, Allen S, Karita E, Hunter E and Yue L (2021) Increased Frequency of Inter-Subtype HIV-1 Recombinants Identified by Near Full-Length Virus Sequencing in Rwandan Acute Transmission Cohorts. Front. Microbiol. 12:734929. doi: 10.3389/fmicb.2021.734929*

of TF viruses and in particular the prevalence of unique recombinant forms (URFs) in transmission cohorts.

**Keywords: HIV-1 transmission, genetic bottleneck, HIV subtype, inter-subtype recombination, antiretroviral drug resistance mutations, near full-length genome sequencing**

## INTRODUCTION

Worldwide 37 million people are living with HIV, two-thirds of these infected individuals are found in Sub-Saharan Africa (UNAIDS, 2020). Even though more than half are receiving ART, a significant fraction of treated patients are not virally suppressed (Hamers et al., 2012; Hauser et al., 2019), and HIV prevention remains a major problem in the fight against HIV. A global effort to design and develop an effective HIV-1 vaccine has been carried out over the last 30 years, but one of its major challenges is the enormous diversity of HIV-1. This can be attributed to several factors: the high error rate of the viral reverse transcriptase, since it lacks a proof-reading function; host immune responses that constantly apply selection pressure for less susceptible virus; and the propensity for the virus to undergo recombination (Lukashov and Goudsmit, 1998; Mansky, 1998; Korber et al., 2001; Song et al., 2018). A number of investigators have studied the interplay between HIV-1 and host immunity, and have shown that viral adaptation, particularly to the cellular immune response during the course of infection, can be a major contributor to viral evolution (Moore et al., 2002; Brumme and Walker, 2009; Crawford et al., 2009; Kawashima et al., 2009; Carlson et al., 2012, 2016; Monaco et al., 2016). However, recombination between genetically distinct viruses has the greatest potential to generate diversity (McCutchan et al., 1996; Butler et al., 2007; Lau and Wong, 2013; Giovanetti et al., 2020). HIV-1 has been classified into four phylogenetic groups M, O, N, and P based on nucleic acid sequencing of the viral genomic RNA, with group M being by far the most widespread (Robertson et al., 2000; Desire et al., 2018). The latter is subdivided into nine different subtypes (A–D, F–H, J, K, and the newly identified L), with genetic variation between subtypes ranging from 20 to 35% depending on the genomic regions and the subtypes being compared (Korber et al., 2001; Desire et al., 2018; Yamaguchi et al., 2020). The process of recombination between viruses belonging to different subtypes and the ongoing spread of those recombinants is the basis for the emergence of circulating recombinant forms or CRFs. To date over 102 inter-subtype CRFs have been described (Hemelaar et al., 2020; LANL, 2021). Therefore, to develop a broadly effective prophylactic vaccine, there is a clear need to gain insight into the genotypic and phenotypic features of the viruses from various geographic locations against which a potential vaccine must act.

Rwanda is an East-Central African country bordered by the Democratic Republic of Congo, with a highly diverse HIV-1 population (Rodgers et al., 2017), Burundi, where subtype C is most prevalent (Delatorre and Bello, 2012) and Uganda where subtypes A1 and D predominate, but with a high percentage of unique recombinant forms (URFs; Lee et al., 2017; Grant et al., 2020). Thus, defining the nature of HIV-1 diversity over time in this geographically small, land-locked country will be relevant to ongoing HIV-1 vaccine efforts. Based on a region encompassing the *pol* gene, an earlier subtype analysis of over 90 incident infections enrolled under IAVI Protocol C in Rwanda identified 80% as subtype A1 and only 6.5% as recombinant viruses (Amornkul et al., 2013), while a second study of smaller sample size, where *gag*, *pol*, and *env* genes were sequenced, reported 13.5% recombinant forms in the same cohort (Kemal et al., 2013).

In the current study, we have amplified near full-length single genomes (NFLG) of viruses from the plasma of a total of 26 acutely HIV infected individuals from the Rwandan heterosexual acute infection cohort Protocol C and 21 recently infected individuals from high-risk cohorts. This allowed us to define the sequence of the infecting viruses and compare over two time periods (2005–2012 and 2016–2019) the frequency of inter-subtype recombination across the full genome. In addition, because these two periods define very different availability of anti-retroviral therapies, we were able to compare the prevalence of antiretroviral drug resistance. These data point to an increase in genetic mixing and prevalence of transmitted drug resistance in Rwanda over the last 15 years and highlight the importance of NFLG sequencing for assessing diversity in viral populations.

## MATERIALS AND METHODS

### Ethics Statement

Subjects in this study were enrolled in human subjects protocols approved by the Rwanda National Ethics Committee and Emory University Institutional Review Board. All study subjects have provided written informed consent.

### Study Subjects

In this study, we have studied HIV-1 Early infection subjects from two distinct time periods during last 15 years in Kigali Rwanda. During the first period from 2005 to 2011, plasma samples were collected under IAVI Protocol C (Price et al., 2020) from 26/97 (27%) seroconvertors from a heterosexual transmission cohort in Kigali. Individuals enrolled in this cohort were from HIV-1 discordant couples who underwent couples counseling and testing and who were followed, with additional counseling and testing, every 1–3 months to reduce the incidence of transmission. HIV infection was identified by p24 ELISA antigen testing or seroconversion. The 26 seroconverters were selected from the original 97 based on the availability of sample during the very early/acute period of HIV infection. During the second period from 2016 to 2019, 21 individuals who were followed in government clinics every 3–6 months in virtual cohorts comprised of high-risk women, men having sex with men and discordant couples, were enrolled immediately after seroconversion.

## Viral RNA Extraction and cDNA Synthesis

Viral RNA was extracted from patient's plasma using the QIAamp RNA mini kit (Qiagen, Valencia, CA). For near-full-length genome amplification, 140 µl plasma were used for vRNA isolation, and then the purified vRNA was converted to full-length cDNA with SuperScript III Reverse Transcriptase (Life Technologies) enzyme with a reverse HIV-1 primer that designed at the end of the R region in LTR (Yue et al., 2015). For 5' and 3' half genome amplifications, the amount of plasma sample was calculated according to VL and aliquoted based on 300 copies x # of reaction per each extraction, and then diluted to 140 µl with PBS if less than 140 µl.

## Near Full-Length Single HIV-1 Genome Amplification

cDNA was serially diluted to yield approximately 30% PCR positive to ensure the majority of amplicons were derived from single virus RNA molecule (Yue et al., 2015). A 9 kb PCR fragment extending from the 5' U5 to 3' R region of the genome was generated by using Q5 Hot Start High Fidelity DNA Polymerase (NEB; Deymier et al., 2014). The amplification primers are shown in **Table 1** and conditions were as described previously (Deymier et al., 2014).

## One-Step RT-PCR Population Half-Genome Amplifications

One-step PCR was conducted using the SuperScript™ III One-Step RT-PCR System with Platinum™ Taq High Fidelity DNA Polymerase (Invitrogen). Master mix I (MMI) contained 25 µm reverse first round primer and template vRNA (300 copies per reaction), adjusted to a total volume of 11 µl per reaction with $H_2O$. The MMI was incubated at 65°C for 5 min to melt secondary structures in the RNA then temperature was decreased to 4°C to anneal the 5' or 3' first round reverse primer (**Table 1**) with the RNA template. Master mix II (MMII) contains 2XReaction buffer, 5' or 3' first round forward primer (25 µm; **Table 1**), SuperScript III/Platinum Taq Mix in a total volume of 39 µl per each reaction. The 39 µl MMII was added to 11 µl MMI at 4°C, and then the entire 50 µl reaction was incubated at 55°C for 30 min in PCR cycler to synthesize cDNA.

After the cDNA was synthesized, the initial PCR step was 2 min at 94°C; followed by 30 cycles of 94°C 15 s, 52°C 30 s, 68°C 6 min; and an additional one step of 68°C for 10 min for stabilization.

Second round PCR was carried out by using Q5 Hot Start High Fidelity DNA Polymerase (NEB). The sequences and positions of the 5' half second round primers are shown in **Table 1**. PCR conditions were 98°C 30 s as the initial step, followed by 35 cycles of 98°C 10 s, 64°C 30 s and 72°C 4 min; plus, an additional step of 72°C 10 min prior to keeping the reaction at 4°C.

The sequences and positions of the 3' half second round forward primer and reverse primer OFM19 are shown in **Table 1**. PCR conditions were 98°C 30 s as the initial step, followed by 35 cycles at 98°C 10 s, 58°C 30 s for annealing and 72°C 4 min for extension; plus, an additional step of 72°C 10 min prior to keeping the reaction at 4°C. The second round PCR resulted in a 4,456 bp fragment from the 5'half, and a 4,742 bp fragment from the 3'half.

## PacBio DNA Sequencing Library Preparation

Four SMRTbell™ libraries of NFLSGA and four SMRTbell™ libraries of half genome amplicons were built to gain deep sequencing data. The PacBio sequencing method was described previously (Dilernia et al., 2015). In brief, we combined 75 NFLSGA amplicons for each RSII library; 10 patients' half genome PCR products were collected for each RSII library. The final library DNA concentration was more than 20 ng/µl,

**TABLE 1** | cDNA and PCR Primers.

| Primer name | Primer sequence (5'-3') | Position in HXB2 | Application |
|---|---|---|---|
| OFM19 | 5'-GCACTCAAGGCAAGCTTTATTGAGGCTTA-3' | 9,604–9,632 | cDNA synthesis for NFLSGA |
| 1.3'3'plCb | 5'-ACTACTTAAAGCACTCAAGGCAAGCTTTATTG-3' | 9,611–9,642 | cDNA synthesis for NFLSGA |
| 1U5Cc | 5'-CCTTGAGTGCTCTAAGTAGTGTGTGCCCGTCTGT-3' | 538–571 | First round NFLSGA forward primer |
| 1.3'3'plCb | 5'-ACTACTTAAAGCACTCAAGGCAAGCTTTATTG-3' | 9,611–9,642 | First round NFLSGA reverse primer |
| 2U5Cd | 5'-AGTAGTGTGTGCCCGTCTGTTGTGTGACTC-3' | 552–581 | Second round NFLSGA forward primer |
| 2.3'3'plCb | 5'-TAGAGCACTCAAGGCAAGCTTTATTGAGGCTTA-3' | 9,604–9,636 | Second round NFLSGA reverse primer |
| 1half_R1_For | 5'-TTTGACTAGCGGAGGCTAGAA-3' | 761–781 | 5'HF first round PCR forward primer |
| 1half_R1_Rev | 5'-TTCTATGGAGACYCCATGACCC-3' | 5,304–5,283 | 5'HF cDNA synthesis and first round PCR reverse primer |
| 2half_R1_For | 5'-GGGTTTATTACAGGGACAGCAGAG-3' | 4,900–4,923 | 3'HF first round PCR forward primer |
| OFM19 | 5'-GCACTCAAGGCAAGCTTTATTGAGGCTTA-3' | 9,604–9,632 | 3'HF cDNA synthesis and first and second round PCR reverse primer |
| 1half_R2_For | 5'-TTTGACTAGCGGAGGCTAGAAGGA-3' | 761–784 | 5'HF second round PCR forward primer |
| 1half_R2_Rev | 5'-TCCCCTARTGGGATGTGTACTTCTGAAC-3' | 5,195–5,222 | 5'HF second round PCR reverse primer |
| 2half_R2_For | 5'-GCAAAACTACTCTGGAAAGGTGAAGGG-3' | 4,944–4,970 | 3'HF second round PCR forward primer |

purity 260/280 ratio was greater than 1.8, and 260/230 ratio was greater than 2.0; total volume was 30 µl. SMRT sequencing was performed on a PacBio RSII at the University of Delaware DNA Sequencing & Genotyping Center.

## Sequence Analysis

Data derived from the PacBio RSII was run using the error correction algorithm MDPseq (Dilernia et al., 2015). Defining transmitted founder (TF) viral sequences and phylogenetical analysis were carried out through Geneious v9.1.8 (Biomatters Ltd). Codon-align for each HIV-1 protein was performed by Gene Cutter (LANL). Subtyping and recombinant identification were carried out by Recombinant analysis program (RIP) and jpHMM at GOBICS (LANL).

## GenBank Submission

Near full-length (NFL) sequences were submitted to GenBank. The accession numbers for the 26 Protocol C derived viruses are JX236678.1, JX236677.1, and MT942708-MT942972. Those for the 21 recent seroconverters are MZ642260-MZ642280.

# RESULTS

## Study Volunteers

This project was conducted in partnership with Projet San Francisco/Centre for Family Health Research which was established in Kigali, Rwanda in 1986. Two distinct groups of HIV-1 infected volunteers were studied: Group 1 represented acutely and very-early infected individuals from 2005 to 2011 (IAVI's Protocol C cohort), while Group 2 represented infected individuals with early HIV infection from 2016 to 2019. For Group 1, HIV discordant couples enrolled in a longitudinal prospective prevention study were provided with counseling, condoms and HIV testing of the seronegative partner during the study (Allen et al., 1992). Couples voluntary counseling and testing (CVCT) in high prevalence areas has been shown to reduce transmission incidence of HIV in cohabiting couples by more than two-thirds (Wall et al., 2019). When infection of the seronegative partner was identified as described in methods, they were enrolled in IAVI Protocol C, an acute infection, long-term follow-up study, and samples were obtained from both partners (Price et al., 2020). Originally, 94 volunteers with incident HIV infection were enrolled into Protocol C from Kigali; in the current study, we analyzed plasma viruses from the 26 seroconverters with the shortest estimated time from the date of infection (median time from EDI = 23 days) calculated as described in methods; with 16 out of 26 plasma collected less than 30 days post-EDI (**Table 2**). For Group 2, individuals were identified in collaboration with government clinics generally within 3 months of their last seronegative visit (median time from EDI = 91 days; **Table 3**), so that we could compare the phylogenetics and subtypes of contemporaneous viruses circulating in 2016 through 2019 with those from 2005 to 2011. We analyzed plasma viruses from 21 of these newly infected individuals who included partners of HIV-1 discordant couples (2), female sex workers (FSW; 13), and young MSM (6; **Table 3**).

## Severity of Genetic Bottleneck During HIV Transmission in the Protocol C Heterosexual Acute Transmission Cohort

In sub-Saharan Africa, heterosexual transmission remains the predominant mode of infection by HIV-1, and accounts for approximately 75% of newly infected cases worldwide (Monaco et al., 2017). We performed near full-length, single genome amplification (NFLSGA ~9,000 bp) on HIV-1 from the 26 acute plasma samples using a high-fidelity nested PCR approach described previously (Rousseau et al., 2006; Deymier et al., 2014; Yue et al., 2015; Kinloch et al., 2019). A total of 288 PCR amplicons from viral RNA were sequenced with an average of 11 NFLSGAs per individual using a multiplexed, highly accurate, DNA sequencing approach based on the PacBio Sequencing platform combined with the MDPseq work flow we have described previously (Dilernia et al., 2015). Phylogenetic analysis of the near full-length genome sequences showed that in 20 out of 26 Rwandan seroconverters a single virus variant established infection, with sequences for each individual clustering in a very homogeneous rake (**Figure 1**). In contrast, in six of 26 individuals, infection was initiated by more than one virus variant (**Figure 1**); as an example, for the acutely infected individual 175,071 three clearly distinct sequence branches were observed in the phylogenetic tree and three different populations were shown by the highlighter analysis (**Figure 2**).

## Inter-Subtype Recombinants Recognized by Near Full-Length Sequencing

Worldwide HIV-1 Group M is the major source of the global pandemic. In this group there are 9 subtypes, over 100 CRF, and many URF (Hemelaar et al., 2020; LANL, 2021). Globally, the proportions of recombinants has increased over time, reaching almost 23% of all infections in the period 2010 to 2015 (Hemelaar et al., 2020), and URF infections occur frequently in the regions and countries where more than one subtype circulate (Tebit and Arts, 2011). Based on previous subtyping of HIV-1 infection for Protocol C volunteers which employed *pol* gene sequences (Fabiani et al., 1998; Amornkul et al., 2013), we expected that the 26 viruses would comprise approximately 21 Subtype A1 (81%); three subtype C (11%); and two Recombinant (8%, 1 A1/C and 1 C/D). In contrast, utilizing the 9 kb near full-length genome sequences and the programs RIP and jpHMM[1] to detect recombinants, we observed a 3-fold higher percentage of recombinant viruses, with six (5A1/C and 1C/D) recombinants (23%) in addition to 18 subtype A1 (69%), and two subtype C (8%) infections (**Figures 1, 3A** and **Table 2**). Each of the recombinant genomes exhibited unique recombination breakpoints (URFs), with two viruses 175,011 and 175,017 resulting from multiple crossovers between their subtype A1 and C progenitors (**Figure 3A**). A majority of the A1/C recombinants retained a portion of the Env gene and/or

[1] www.hiv.lanl.gov

**TABLE 2 |** Transmission variants and subtypes in protocol C individuals.

| PCID | Coded ID | Sample date | EDI | VL | Transmission variants | Sequence type | Subtype by NFLG | Subtype by pol |
|---|---|---|---|---|---|---|---|---|
| 175,019 | R49M | 30-Mar-05 | 10 | 3,000,000 | Single | TF | A1 | A1 |
| 175,042 | R463F | 9-Mar-07 | 10 | 152,000,000 | Single | TF | A1 | A1 |
| 175,079 | R3103M | 19-Sep-08 | 13 | 750,001 | Single | TF | A1 | A1 |
| 175,089 | R3584M | 15-Dec-09 | 13 | 223,440 | Single | TF | A1C | A1 |
| 175,005 | R53F | 16-Jun-05 | 14 | 2,090,040 | Single | TF | A1 | A1 |
| 175,062 | R1135M | 11-Feb-08 | 14 | 1,256,940 | Single | TF | A1 | A1 |
| 175,065 | R269M | 24-Mar-08 | 14 | 31,700,000 | Single | Con | A1 | A1 |
| 175,072 | R3137M | 28-May-08 | 14 | 9,061,540 | Single | Con | C | C |
| 175,053 | R254F | 8-Aug-07 | 15 | 1,876,000 | Single | TF | A1C | C |
| 175,090 | R3469F | 15-Apr-10 | 15 | 2,920,000 | >2 | TF(P1) | A1 | A1 |
| 175,093 | R2302M | 9-Jul-10 | 15 | 3,760,000 | Single | TF | A1 | A1 |
| 175,094 | R3843F | 14-Oct-10 | 16 | 4,400,000 | Single | TF | A1 | A1 |
| 175,059 | R977F | 14-Dec-07 | 17 | 7,290,000 | Single | TF | A1 | A1 |
| 175,038 | R880F | 12-Jan-07 | 21 | 730,000 | Single | TF | A1 | A1 |
| 175,097 | R3894M | 18-Mar-11 | 23 | 6,934,997 | Single | TF | CD | CD |
| 175,092 | R3671F | 8-Jun-10 | 25 | 3,940,000 | Single | NA | A1 | A1 |
| 175,074 | R873F | 24-Jul-08 | 37 | 219,920 | 2 | TF(P1) | A1 | A1 |
| 175,071 | R1077F | 12-Jun-08 | 41 | 4,702,444 | 3 | TF(P2) | A1 | A1 |
| 175,008 | R50M | 25-May-05 | 45 | 117,608 | >2 | NA | C | C |
| 175,020 | R57F | 12-Oct-05 | 46 | 134,472 | >2 | NA | A1 | A1 |
| 175,014 | R59M | 21-Oct-05 | 46 | 806,290 | Single | TF | A1 | A1 |
| 175,017 | R44M | 3-May-05 | 49 | 123,560 | Single | TF | A1C | A1 |
| 175,011 | R63M | 14-Dec-05 | 50 | 184,270 | Single | Con | A1C | A1C |
| 175,012 | R40F | 31-Mar-05 | 53 | 1,398,004 | Single | TF | A1C | A1 |
| 175,027 | R72M | 10-Aug-06 | 67 | 425,000 | >2 | NA | A1 | A1 |
| 175,010 | R65M | 27-Jan-06 | 73 | 148,220 | Single | TF | A1 | A1 |

*PCID, protocol C identification number of the participant; Coded ID, coded identification number used in two previous publications (Haaland et al., 2009; Yue et al., 2015); Sample date, date of sample collection; EDI, time since estimated date of infection; VL, viral load of sample; Transmission variants, number of genetic variants transmitted from partner; Sequence type, sequence defined as TF: transmitted founder virus sequence amplicon identical to consensus identified in NFLG amplicons; (P1), (P2): TF sequence identified in subpopulation 1 or 2 of multiple variants, Con: consensus sequence of NFLG amplicons: Subtype by NFLG: subtype as deduced from the NFLG sequence, Subtype by pol: Subtype defined previously from pol gene sequencing.*

**TABLE 3 |** Summary of recent infection samples and derived near full-length sequences.

| Coded ID | Sample date | EDI | VL | Subtype | Risk group* |
|---|---|---|---|---|---|
| BUS71F | 10-July-19 | 91.5 | 328,000 | A1 | FSW |
| GIT84F | 1-August-19 | 131 | 102,000 | A1 | FSW |
| GWE47F | 10-October-18 | 111.5 | 150,000 | A1 | FSW |
| GWE68F | 25-June-19 | 216 | 120,000 | A1 | FSW |
| KAG34F | 10-November-17 | 88 | 191,000 | A1/C/D | FSW |
| KIN18F | 17-March-17 | 46 | 113,000 | A1 | FSW |
| MAS1F | 18-August-16 | 44.5 | 553,000 | A1 | FSW |
| MAS21F | 30-March-17 | 164.5 | 856,000 | A1 | DC |
| MAS6M | 22-November-16 | 174.5 | 202,000 | A1/C | DC |
| MAS7F | 23-November-16 | 46 | 47,200 | A1/C | FSW |
| MAT81F | 24-July-19 | 181 | 292,000 | A1/C/D | FSW |
| NGA76F | 17-July-19 | 90 | 141,000 | A1/C | FSW |
| NGA77F | 17-July-19 | 90 | 473,000 | A1/C | FSW |
| PSF24M | 2-June-17 | 96 | 306,000 | A1/C | MSM |
| PSF33M | 20-October-17 | 194 | 160,000 | A1/C | MSM |
| PSF36F | 4-January-18 | 94 | 702,000 | A1/C | FSW |
| PSF38M | 17-September-18 | 174 | 145,000 | A1/C | MSM |
| PSF39M | 17-February-18 | 72 | 100,000 | A1/C | MSM |
| PSF3M | 21-October-16 | 51 | 86,500 | A1/C | MSM |
| PSF80M | 19-July-19 | 17 | 683,000 | C | MSM |
| REM29F | 24-August-17 | 91 | 148,000 | A1 | FSW |

*Coded ID, coded identification number of the participant; sample date, date of sample collection; EDI, time since estimated date of infection; VL, viral load of sample; subtype, subtype defined by near full-length genomic sequence; risk group, risk group of the individual, FSW: female sex worker; DC, HIV discordant couple; and MSM, men having sex with men.*

Nef from their subtype A parent, but no single region was conserved in all five of these A1/C recombinants. Of these the five A1/C URFs and 1 C/D URF were single virus transmission cases: while one of two subtype C and five of 18 subtype A1 involved multi-variant transmission (**Figure 1**).

## Near Full-Length Sequencing of Plasma Virus in Recent Seroconverters

Near full-length (9 kb) single genome amplification is very inefficient, expensive and time consuming, therefore, for recent samples, we opted to utilize the more efficient population PCR amplification of 5' and 3' half-genome regions, with amplicons that overlapped by ~250 nucleotides. The population amplicons were then sequenced using next-generation PacBio single molecule real time sequencing and individual reads were analyzed using the MDPseq workflow. This allowed us to determine the consensus sequence for the early virus population and breakpoints in those determined to be recombinants. A phylogenetic analysis of the 21 recent seroconversion viruses in the context of the 26 viruses from Protocol C (Red; **Figure 4**) shows that overall, the two viruses from discordant couple transmissions (Blue) and a majority of the viruses from newly infected FSWs (Magenta) clustered within the diversity of the older viral isolates. In contrast five of the six newly infecting viruses from the cohort of young MSM (Cyan) clustered together on the phylogenetic tree with the sequences exhibiting very limited diversity (median 97.5% identity). This would be consistent with a recent transmission network within a risk group that otherwise has demonstrated low sero-incidence (Karita et al., in preparation).

An analysis for recombination using the RIP and jpHMM tools[2] revealed that 12 of the 21 recent seroconversions were initiated by unique A1/C and A1/C/D recombinant forms (**Figure 4**, denoted by blue circles). These represent, therefore, 57% of the samples analyzed, a significantly higher frequency than we observed in the Protocol C samples (23%; $p = 0.033$). Even if we define the five closely related viruses from MSM

[2] https://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html

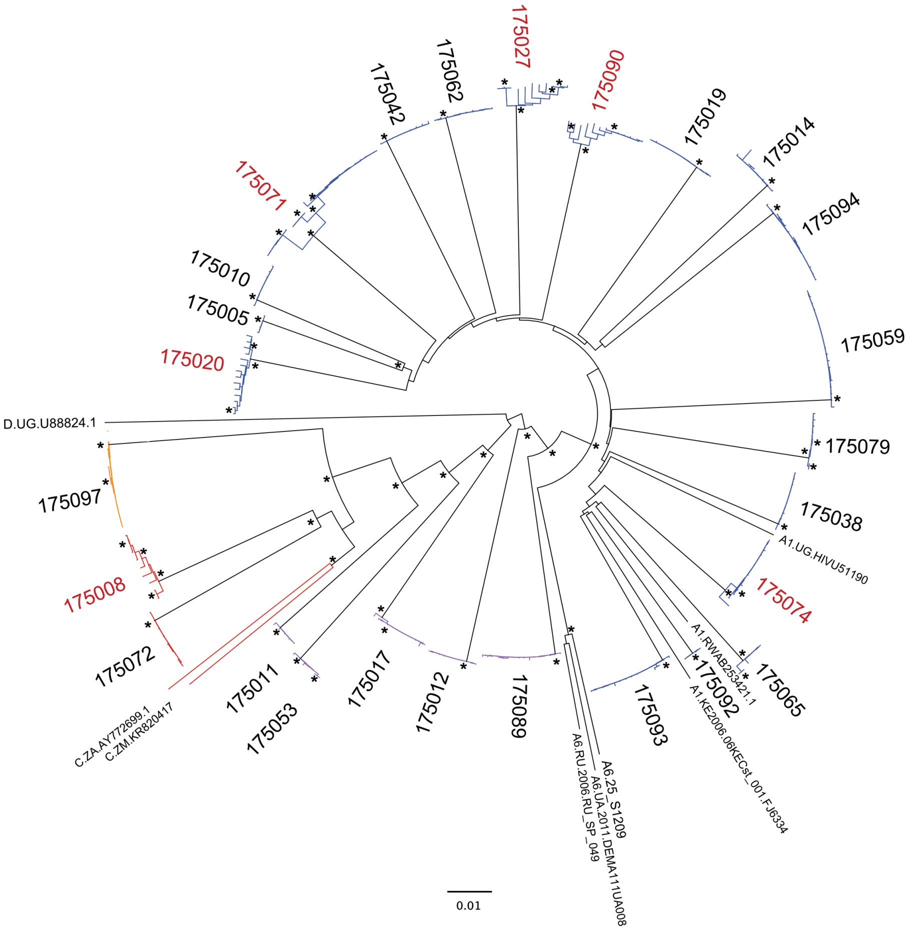


**FIGURE 1** | Phylogenetic analysis of virus sequences from Protocol C acute infections. Neighbor-joining tree representing the near full-length genome (NFLG) sequences of viruses from 26 acutely infected recipients with reference subtypes A1, A6, C, and D. Subtype and recombinants are indicated with specific colors (Subtype A, Blue; Subtype C, Red; Recombinant AC, Purple; Recombinant CD, Orange). The IDs of subjects infected with more than one viral variant from their partner are highlighted in red. Nodes with bootstrap values > 0.9 are denoted with an asterisk.

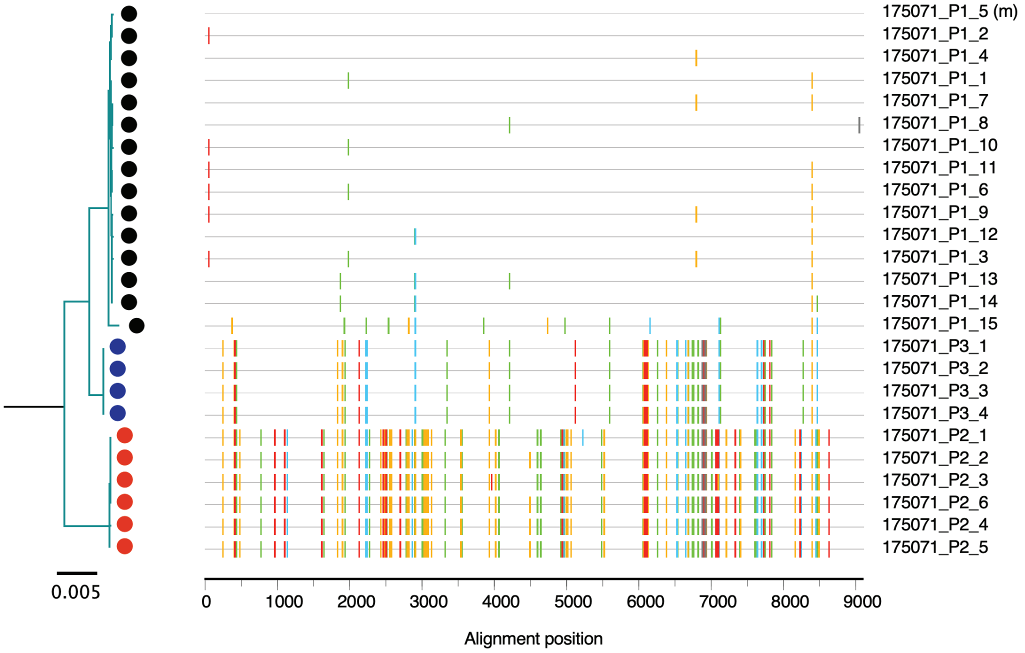


**FIGURE 2 |** Multi-variants transmission. The phylogenetic tree and aligned sequence highlighter analysis show 175,071 contains three viral populations at the first visit time-point (EDI = 41). Tick marks indicate nucleotide change from the master at same position according to the alignment. The color codes are as follows: A: green, T: red, G: yellow, C: blue, and gaps: gray.

as a single recombinant, the frequency (47.1%) while no longer significant ($p = 0.182$) remains double that of the earlier samples.

Interestingly, while each of the recent recombinants exhibited unique numbers and positions of crossovers (**Figure 3B**), a region of subtype A1 Env, extending from just before an 18-residue amphipathic alpha-helical region located on the outer domain of gp120, known as the alpha 2 helix, to the membrane-spanning domain of gp41, was conserved throughout. A similar, albeit shorter region that terminated in the C-terminal heptad repeat of gp41, was present in four of the five unique A1/C recombinants identified in the Protocol C samples (**Figure 3A**).

## Identification of Antiretroviral Drug Resistance Mutations

Finally, in order to determine whether, in the context of increased availability of ART, transmitted drug resistance was increasing, sequences encompassing the protease, reverse transcriptase, and integrase were extracted from all 47 near full-length genomes and submitted to the Stanford HIVdb analysis program, and any major drug resistance mutations (DRMs) documented. None of the Protocol C viral sequences analyzed for this study encoded any major DRMs. In contrast, four of the 21 recent isolates encoded DRMs in the reverse transcriptase. MAS-21F, MAS-1F, and PSF-80M encoded the K103N mutation that confers high level resistance to the non-nucleoside inhibitors nevirapine (NVP) and efavirenz (EFV). NGA-77F encoded K103S, which confers high level resistance to NVP and intermediate resistance to EFV, and E138A, which confers low level resistance to etravirine and rilpivirine.

## DISCUSSION

Here, we report on the NFLG sequences of HIV from 47 acute and early infected Rwandans. This more than quadruples the number of near full-length sequences previously reported for this East African country (Lamers et al., 2016). For the IAVI Protocol C samples, we amplified more than 10 NFLG single-genome amplicons per individual, a total of 288, from acute infections that allowed us to define a TF virus sequence in a majority of cases. This was facilitated by the use of PacBio single molecule real-time (SMRT) long-read sequencing combined with the MDPseq workflow, to define accurate sequences (Dilernia et al., 2015), We utilized the same workflow for the more efficient, overlapping half-genome population amplicons generated from recent infections. The long-read technology allowed the population PCR to be deconvoluted to yield the individual sequences of the corresponding amplicons in order to generate a consensus sequence of the infecting virus.

We have previously reported, based on sequencing of the V1-V4 region of Env that in a majority of transmission events in both a Zambian discordant couple cohort and the Rwandan Protocol C cohort infection was established by a single genetic variant from the transmitting partner (Derdeyn et al., 2004; Haaland et al., 2009). In general, multiple variant infections represented only 10–15% of the transmission pairs examined

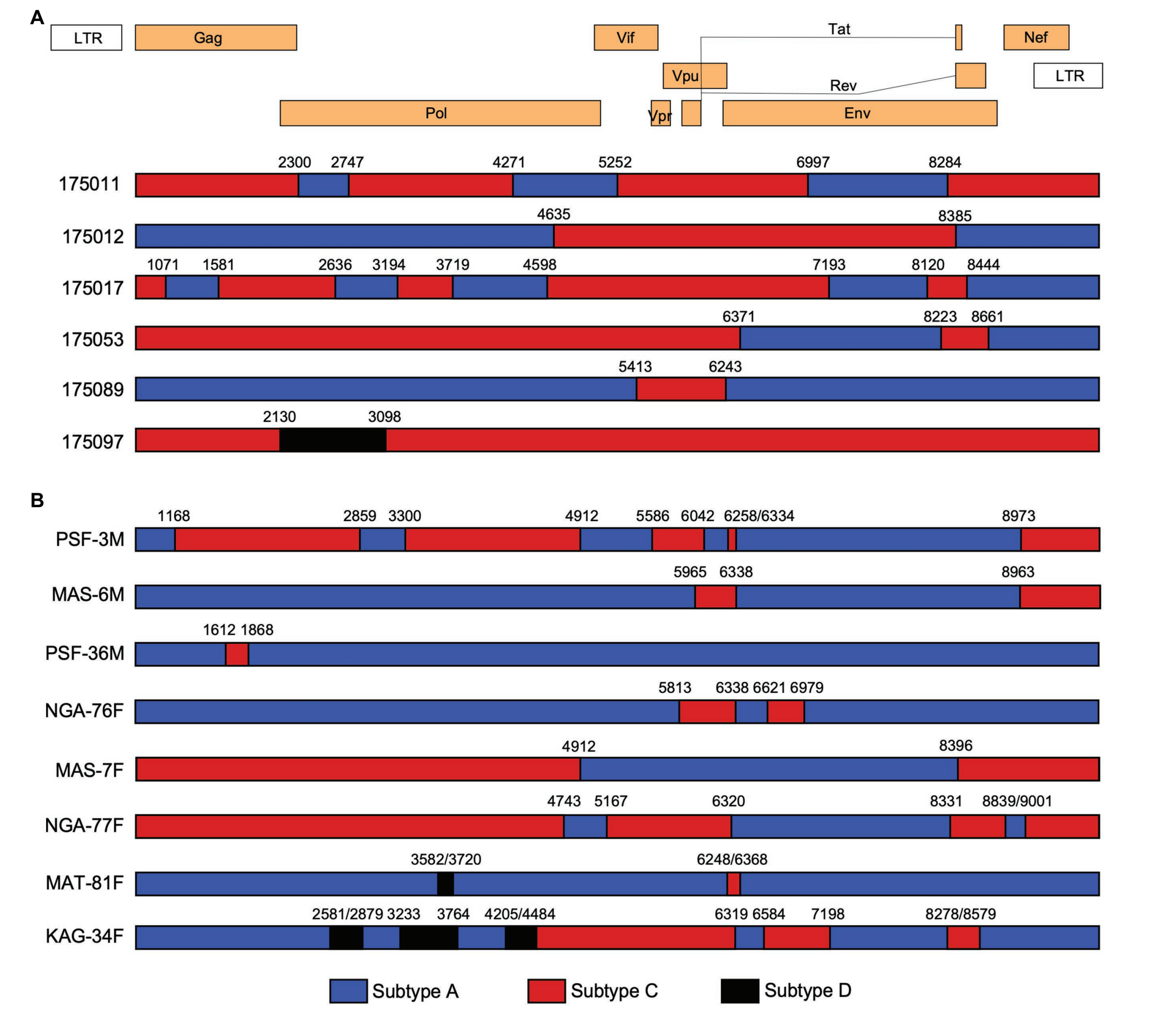


**FIGURE 3 |** Recombination analysis of Protocol C and recent infections. The positions of recombination cross-overs and the subtype origin of genomic regions are shown based on the jphMM analytical tool. Nucleotide positions refer to the equivlent positions on the HXB-2 HIV-1 genome. **(A)** The six unique recombinant forms (URFs) from the Protocol C acute infection cohort. **(B)** The recombination structure of recent Rwandan seroconvertor viruses. PSF-3M illustrates the identical recombinant structure found in all five early isolates from MSM (PSF-3M, -24M, -33M, -38M, and -39M). Subtype sequences are denoted by the same colors as in **Figure 1**.

(Haaland et al., 2009). In the current study, we observed a somewhat higher frequency (23%) of infections initiated by more than one partner-derived virus variant, similar to that reported in a South African women cohort (Abrahams et al., 2009) and a predominantly MSM cohort (Keele et al., 2008). We have reported previously that evidence of genital inflammation and ulcers can lower the barrier to transmission and increase multi-variant transmission (Haaland et al., 2009; Carlson et al., 2014), and a recent study in a Kenyan MSM cohort, where sexually transmitted infections were common reported 39% (15 out of 38) of the participants were infected with multiple founder viruses (Macharia et al., 2020). In the heterosexual protocol C cohort studied here we observed evidence of genital inflammation or ulcers in just six individuals in the 6 months prior to HIV-1 infection and these individuals were equally distributed between the single and multiple variant infections groups ($p = 0.67$; Chi-square with Yates correction).

Inter-subtype recombination represents a significant challenge to HIV-1 vaccine design since it provides the virus with a mechanism to rapidly diversify following co-infection of an individual with two different subtype viruses (Robertson et al., 1995). This gives rise to unique recombinant forms that can be prevalent in regions where more than one subtype circulates. An example of this is Uganda, Rwanda's neighbor, where subtypes A and D cocirculate, and recent studies have identified a frequency of recombinants between the two subtypes as high as 49% (Lee et al., 2017; Grant et al., 2020). In contrast, a previous

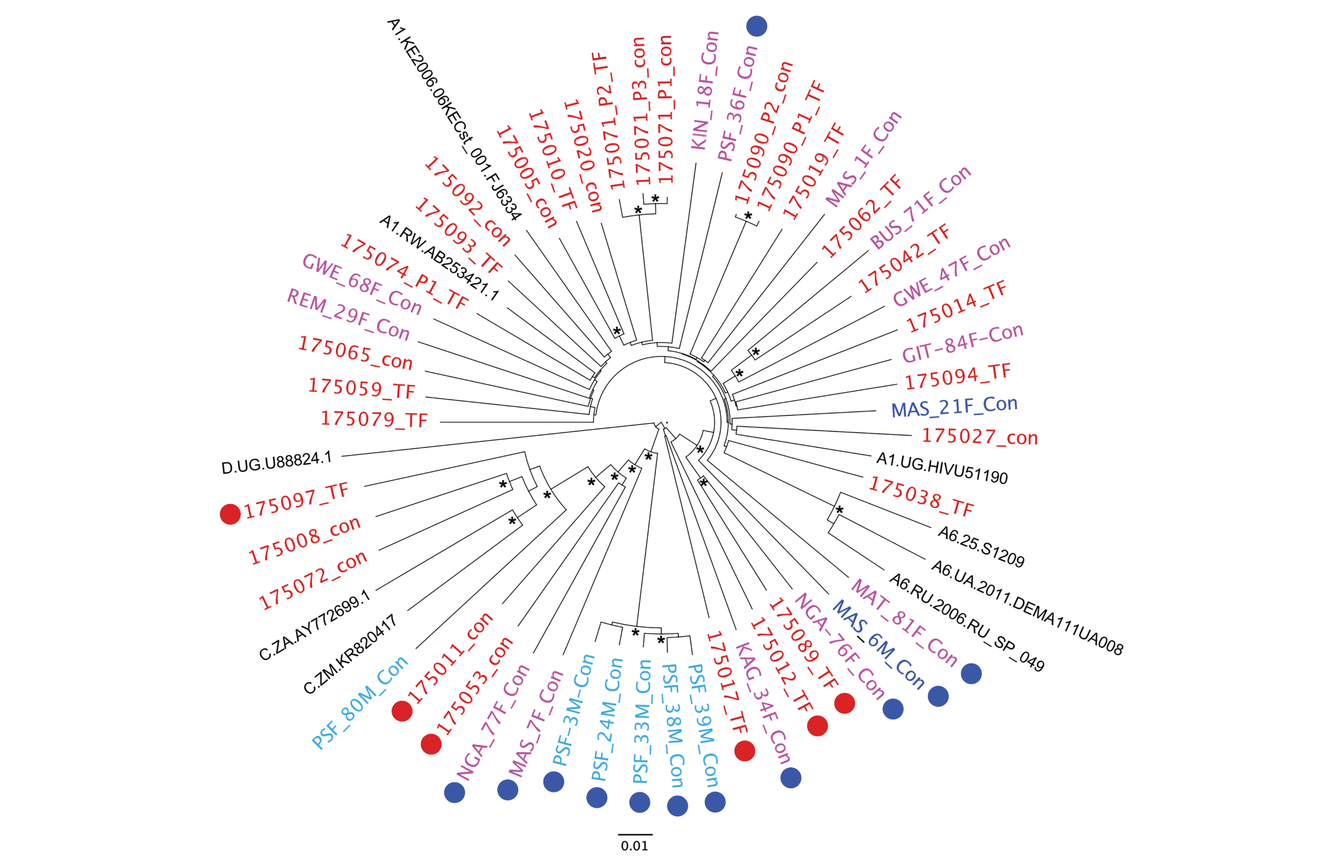


**FIGURE 4 |** Phylogenetic analysis of virus sequences from recent infections in the context of Protocol C acute infections. Neighbor-joining tree representing NFLG sequences of viruses from 21 recent infections and 26 acute Protocol C recipients. Origin of the viruses is denoted by color – Red, Protocol C; Blue, Discordant Couple; Magenta, FSW; Cyan, MSM. Recombinant viruses are highlighted by appended Red (Protocol C) and Blue (Recent Infection) circles. Nodes with bootstrap values > 0.9 are denoted with an asterisk.

analysis of viral subtypes from 30 women in Rwanda in 2013 reported 80% subtype A1, 3% subtype C and D, and 13% AC or AD recombinant forms based on sequencing of *gag*, *pol*, and *env* (Kemal et al., 2013). Similarly, an earlier subtype analysis of 92 IAVI Protocol C incident infections in Rwanda, which was based on a region encompassing the *pol* gene, defined 80% as subtype A1 and only 6.5% as recombinant viruses (Amornkul et al., 2013). By contrast, using NFLG sequences, we identified 23% (6/26) recombinants, a much higher frequency than even the 8% (2/26) previously defined for these same individuals through *pol* sequencing. This highlights the importance of NFLG sequencing to fully understand the complexity of virus populations circulating in multi-subtype countries. Indeed, sequencing of 21 more recent (2016–2019) incident infection viruses from different high-risk groups suggests that recombinant viruses are increasing in frequency, since we observed that 57% of these recent infections were A1/C, or A1/C/D recombinants. With the exception of the five viruses that appeared to represent a recent viral transmission network, the recombinants in these recent infections and those in Protocol C resulted from a series of unique recombination events. In contrast, five of the six viruses from the MSM cohort were highly related and had identical recombination patterns. Of interest, one common recombination breakpoint (4912) was shared by this MSM group (represented by PSF-3M, **Figure 3B**) and its nearest neighbor MAS-7F (**Figure 4**) raising the possibility that the former evolved from the latter following further recombination events. Although, the risk groups in Protocol C and the recent infection cohorts are different, the majority of infections in the latter represented heterosexual transmission in high-risk women and discordant couple partners (15/21), where recombinant viruses remained prevalent (47%). Nevertheless, this difference in risk groups should be considered a potential weakness of the comparison.

Although, a majority of the recombinant viruses from the recent seroconverters exhibited unique recombination break points, they all retained a common region of Env that was derived from a subtype A1 parent. This region extended minimally in KAG-34F from the last few residues of the third variable (V3) loop of gp120, just before the hydrophobic alpha 2 helix, to the beginning of the membrane spanning domain of gp41. We have shown previously, that, in subtype C viruses, the alpha 2 helix is under positive selection pressure and that variations in this region are in part linked to early neutralizing antibody escape (Rong et al., 2007). The conserved region also spans the fifth conserved domain (C5) of gp120 and the ectodomain of gp41, both critical for gp120 and gp41 interactions and trimer stability (Binley et al., 2000; Julien et al., 2013). Thus it is possible that this region from subtype A1 provides a fitness advantage to the recombinants.

We have recently reported on the presence of DRMs in the transmitted virus of newly infected partners of Rwandan couples, where the transmitting partner was on ART but carried drug resistant virus (Woodson et al., 2018). It was of interest, therefore, to compare the prevalence and nature of DRMs in the Protocol C cohort, from a time when antiretroviral drugs were not widely available, to those in more recently infected individuals, when treatment with the standard first-line combination of Tenofovir (TDF), Lamivudine (3TC), and EFV following diagnosis was routine. Although none of the 26 acutely infected individuals from Protocol C encoded DRMs, a previous study of 78 Rwandans from the same cohort did identify five individuals with NNRTI mutations (three with K103N, two with L100I) and one with the protease inhibitor mutation, M46L (Price et al., 2011). Our finding that four of the 21 recent infection viruses encoded K103N/S, which confers resistance to NVP and EFV, indicates that this mutation is becoming more prevalent ($p = 0.035$) within the population. Moreover, since all of these samples were collected prior to the initiation of ART, for the three females whose virus encoded this mutation, the resistance mutations must have been present in the virus of their male partners.

Overall, we demonstrate here, through NFLG sequence analysis that recombinant viruses are more prevalent in Rwanda than previously reported, and that the frequency of both recombinants and NNRTI DRMs appear to have increased between two sampling periods. It will be critical therefore to continue to monitor the nature of circulating viruses to ensure the validity of ongoing vaccine development efforts.

## DATA AVAILABILITY STATEMENT

NFL sequences were submitted to GenBank. The accession numbers for the 26 Protocol C derived viruses are JX236678.1, JX236677.1, MT942708-MT942972. Those for the 21 recent seroconverters are MZ642260-MZ642280.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Rwanda National Ethics Committee and Emory University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

LY, EH, JG, SA, and EK conceived and designed the experiments. GU, EM, EN, RX, DD, KH, HS, and QQ performed the experiments. LY, GU, EM, and EN analyzed the data. LY, EH, PF, JH, JB, EK, JG, and SA contributed reagents, materials, and analysis. EH and LY wrote the paper. All authors contributed to the article and approved the submitted version.

## FUNDING

This work is made possible in part by NIH grant R01AI051231 and by IAVI contract # 06-646-9933 through the support of the American People through the United States President's Emergency Plan for AIDS Relief (PEPFAR) through the United States Agency for International Development (USAID). The contents of this manuscript are the sole responsibility of the authors and do not necessarily reflect the views of PEPFAR, USAID, or the United States Government. This work was also supported, in part, by the Virology Core at the Emory Center for AIDS Research by performing viral load determinations (grant P30 AI050409); the Yerkes National Primate Research Center base grant through the Office of Research Infrastructure Programs /OD P51OD11132. EH is a Georgia Eminent Scholar.

## ACKNOWLEDGMENTS

The authors thank all the volunteers in Rwanda who participated in this study and all the staff in the Center for Family Health Research who made this study possible. The authors would like to thank Jon Allen and Paul Farmer for technical assistance, sample management, and database management, and Pat Fast and Matt Price for thoughtful comments on the manuscript.

## REFERENCES

Abrahams, M. R., Anderson, J. A., Giorgi, E. E., Seoighe, C., Mlisana, K., Ping, L. H., et al. (2009). Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J. Virol.* 83, 3556–3567. doi: 10.1128/JVI.02132-08

Allen, S., Tice, J., Van De Perre, P., Serufilira, A., Hudes, E., Nsengumuremyi, F., et al. (1992). Effect of serotesting with counselling on condom use and seroconversion among HIV discordant couples in Africa. *BMJ* 304, 1605–1609. doi: 10.1136/bmj.304.6842.1605

Amornkul, P. N., Karita, E., Kamali, A., Rida, W. N., Sanders, E. J., Lakhi, S., et al. (2013). Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. *AIDS* 27, 2775–2786. doi: 10.1097/QAD.0000000000000012

Binley, J. M., Sanders, R. W., Clas, B., Schuelke, N., Master, A., Guo, Y., et al. (2000). A recombinant human immunodeficiency virus type 1 envelope glycoprotein complex stabilized by an intermolecular disulfide bond between the gp120 and gp41 subunits is an antigenic mimic of the trimeric virion-associated structure. *J. Virol.* 74, 627–643. doi: 10.1128/JVI.74.2.627-643.2000

Brumme, Z. L., and Walker, B. D. (2009). Tracking the culprit: HIV-1 evolution and immune selection revealed by single-genome amplification. *J. Exp. Med.* 206, 1215–1218. doi: 10.1084/jem.20091094

Butler, I. F., Pandrea, I., Marx, P. A., and Apetrei, C. (2007). HIV genetic diversity: biological and public health consequences. *Curr. HIV Res.* 5, 23–45. doi: 10.2174/157016207779316297

Carlson, J. M., Brumme, C. J., Martin, E., Listgarten, J., Brockman, M. A., Le, A. Q., et al. (2012). Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *J. Virol.* 86, 13202–13216. doi: 10.1128/JVI.01998-12

Carlson, J. M., Du, V. Y., Pfeifer, N., Bansal, A., Tan, V. Y., Power, K., et al. (2016). Impact of pre-adapted HIV transmission. *Nat. Med.* 22, 606–613. doi: 10.1038/nm.4100

Carlson, J. M., Schaefer, M., Monaco, D. C., Batorsky, R., Claiborne, D. T., Prince, J., et al. (2014). HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* 345:1254031. doi: 10.1126/science.1254031

Crawford, H., Lumm, W., Leslie, A., Schaefer, M., Boeras, D., Prado, J. G., et al. (2009). Evolution of HLA-B*5703 HIV-1 escape mutations in HLA-B*5703-positive individuals and their transmission recipients. *J. Exp. Med.* 206, 909–921. doi: 10.1084/jem.20081984

Delatorre, E. O., and Bello, G. (2012). Phylodynamics of HIV-1 subtype C epidemic in East Africa. *PLoS One* 7:e41904. doi: 10.1371/journal.pone.0041904

Derdeyn, C. A., Decker, J. M., Bibollet-Ruche, F., Mokili, J. L., Muldoon, M., Denham, S. A., et al. (2004). Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science* 303, 2019–2022. doi: 10.1126/science.1093137

Desire, N., Cerutti, L., Le Hingrat, Q., Perrier, M., Emler, S., Calvez, V., et al. (2018). Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. *Retrovirology* 15:80. doi: 10.1186/s12977-018-0461-y

Deymier, M. J., Claiborne, D. T., Ende, Z., Ratner, H. K., Kilembe, W., Allen, S., et al. (2014). Particle infectivity of HIV-1 full-length genome infectious molecular clones in a subtype C heterosexual transmission pair following high fidelity amplification and unbiased cloning. *Virology* 468-470, 454–461. doi: 10.1016/j.virol.2014.08.018

Dilernia, D. A., Chien, J. T., Monaco, D. C., Brown, M. P., Ende, Z., Deymier, M. J., et al. (2015). Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. *Nucleic Acids Res.* 43:e129. doi: 10.1093/nar/gkv630

Fabiani, M., Blè, C., Grivel, P., Lukwiya, M., and Declich, S. (1998). 1989-1996 HIV-1 prevalence trends among different risk groups in Gulu District, North Uganda. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 18:514. doi: 10.1097/00042560-199808150-00015

Giovanetti, M., Ciccozzi, M., Parolin, C., and Borsetti, A. (2020). Molecular epidemiology of HIV-1 in African countries: a comprehensive overview. *Pathogens* 9:1072. doi: 10.3390/pathogens9121072

Grant, H. E., Hodcroft, E. B., Ssemwanga, D., Kitayimbwa, J. M., Yebra, G., Esquivel Gomez, L. R., et al. (2020). Pervasive and non-random recombination in near full-length HIV genomes from Uganda. *Virus Evol.* 6:veaa004. doi: 10.1093/ve/veaa004

Haaland, R. E., Hawkins, P. A., Salazar-Gonzalez, J., Johnson, A., Tichacek, A., Karita, E., et al. (2009). Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathog.* 5:e1000274. doi: 10.1371/journal.ppat.1000274

Hamers, R. L., Sigaloff, K. C., Wensing, A. M., Wallis, C. L., Kityo, C., Siwale, M., et al. (2012). Patterns of HIV-1 drug resistance after first-line antiretroviral therapy (ART) failure in 6 sub-Saharan African countries: implications for second-line ART strategies. *Clin. Infect. Dis.* 54, 1660–1669. doi: 10.1093/cid/cis254

Hauser, A., Kusejko, K., Johnson, L. F., Wandeler, G., Riou, J., Goldstein, F., et al. (2019). Bridging the gap between HIV epidemiology and antiretroviral resistance evolution: modelling the spread of resistance in South Africa. *PLoS Comput. Biol.* 15:e1007083. doi: 10.1371/journal.pcbi.1007083

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020). Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781. doi: 10.1016/S2352-3018(20)30252-6

Julien, J. P., Cupo, A., Sok, D., Stanfield, R. L., Lyumkis, D., Deller, M. C., et al. (2013). Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* 342, 1477–1483. doi: 10.1126/science.1245625

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., et al. (2009). Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458, 641–645. doi: 10.1038/nature07746

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* 105, 7552–7557. doi: 10.1073/pnas.0802203105

Kemal, K. S., Anastos, K., Weiser, B., Ramirez, C. M., Shi, Q., and Burger, H. (2013). Molecular epidemiology of HIV type 1 subtypes in Rwanda. *AIDS Res. Hum. Retrovir.* 29, 957–962. doi: 10.1089/aid.2012.0095

Kinloch, N. N., Lee, G. Q., Carlson, J. M., Jin, S. W., Brumme, C. J., Byakwaga, H., et al. (2019). Genotypic and mechanistic characterization of subtype-specific HIV adaptation to host cellular immunity. *J. Virol.* 93, e01502–e01518. doi: 10.1128/JVI.01502-18

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *Br. Med. Bull.* 58, 19–42. doi: 10.1093/bmb/58.1.19

Lamers, S. L., Barbier, A. E., Ratmann, O., Fraser, C., Rose, R., Laeyendecker, O., et al. (2016). HIV-1 sequence data coverage in Central East Africa from 1959 to 2013. *AIDS Res. Hum. Retrovir.* 32, 904–908. doi: 10.1089/aid.2016.0079

Lanl (2021). HIV Circulating Recombinant Forms (CRFs) [Online]. Los Alamos National Laboratory. Available at: https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html (Accessed Aug 24, 2021).

Lau, K. A., and Wong, J. J. (2013). Current trends of HIV recombination worldwide. *Infect. Dis. Rep.* 5:e4. doi: 10.4081/idr.2013.s1.e4

Lee, G. Q., Bangsberg, D. R., Mo, T., Lachowski, C., Brumme, C. J., Zhang, W., et al. (2017). Prevalence and clinical impacts of HIV-1 intersubtype recombinants in Uganda revealed by near-full-genome population and deep sequencing approaches. *AIDS* 31, 2345–2354. doi: 10.1097/QAD.0000000000001619

Lukashov, V. V., and Goudsmit, J. (1998). HIV heterogeneity and disease progression in AIDS: a model of continuous virus adaptation. *AIDS* 12(Suppl. A), S43–S52.

Macharia, G. N., Yue, L., Staller, E., Dilernia, D., Wilkins, D., Song, H., et al. (2020). Infection with multiple HIV-1 founder variants is associated with lower viral replicative capacity, faster CD4+ T cell decline and increased immune activation during acute infection. *PLoS Pathog.* 16:e1008853. doi: 10.1371/journal.ppat.1008853

Mansky, L. M. (1998). Retrovirus mutation rates and their role in genetic variation. *J. Gen. Virol.* 79, 1337–1345. doi: 10.1099/0022-1317-79-6-1337

Mccutchan, F. E., Salminen, M. O., Carr, J. K., and Burke, D. S. (1996). HIV-1 genetic diversity. *AIDS* 10(Suppl. 3), S13–S20.

Monaco, D. C., Dilernia, D. A., Fiore-Gartland, A., Yu, T., Prince, J. L., Dennis, K. K., et al. (2016). Balance between transmitted HLA preadapted and nonassociated polymorphisms is a major determinant of HIV-1 disease progression. *J. Exp. Med.* 213, 2049–2063. doi: 10.1084/jem.20151984

Monaco, D. C., Ende, Z., and Hunter, E. (2017). Virus-host gene interactions define HIV-1 disease progression. *Curr. Top. Microbiol. Immunol.* 407, 31–63. doi: 10.1007/82_2017_33

Moore, C. B., John, M., James, I. R., Christiansen, F. T., Witt, C. S., and Mallal, S. A. (2002). Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* 296, 1439–1443. doi: 10.1126/science.1069660

Price, M. A., Kilembe, W., Ruzagira, E., Karita, E., Inambao, M., Sanders, E. J., et al. (2020). Cohort profile: IAVI's HIV epidemiology and early infection cohort studies in Africa to support vaccine discovery. *Int. J. Epidemiol.* 50, 29–30. doi: 10.1093/ije/dyaa100

Price, M. A., Wallis, C. L., Lakhi, S., Karita, E., Kamali, A., Anzala, O., et al. (2011). Transmitted HIV type 1 drug resistance among individuals with recent HIV infection in East and Southern Africa. *AIDS Res. Hum. Retrovir.* 27, 5–12. doi: 10.1089/aid.2010.0030

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Robertson, D. L., Sharp, P. M., Mccutchan, F. E., and Hahn, B. H. (1995). Recombination in HIV-1. *Nature* 374, 124–126. doi: 10.1038/374124b0

Rodgers, M. A., Wilkinson, E., Vallari, A., Mcarthur, C., Sthreshley, L., Brennan, C. A., et al. (2017). Sensitive next-generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo. *J. Virol.* 91, e01841–e01816. doi: 10.1128/JVI.01841-16

Rong, R., Gnanakaran, S., Decker, J. M., Bibollet-Ruche, F., Taylor, J., Sfakianos, J. N., et al. (2007). Unique mutational patterns in the envelope alpha 2 amphipathic helix and acquisition of length in gp120 hypervariable domains are associated with resistance to autologous neutralization of subtype C human immunodeficiency virus type 1. *J. Virol.* 81, 5658–5668. doi: 10.1128/JVI.00257-07

Rousseau, C. M., Birditt, B. A., Mckay, A. R., Stoddard, J. N., Lee, T. C., Mclaughlin, S., et al. (2006). Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. *J. Virol. Methods* 136, 118–125. doi: 10.1016/j.jviromet.2006.04.009

Song, H., Giorgi, E. E., Ganusov, V. V., Cai, F., Athreya, G., Yoon, H., et al. (2018). Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nat. Commun.* 9:1928. doi: 10.1038/s41467-018-04217-5

Tebit, D. M., and Arts, E. J. (2011). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infect. Dis.* 11, 45–56. doi: 10.1016/S1473-3099(10)70186-9

UNAIDS (2020). Global HIV & AIDS statistics — 2020 Fact Sheet [Online]. https://www.unaids.org/en/resources/fact-sheet (Accessed April 05, 2021).

Wall, K. M., Inambao, M., Kilembe, W., Karita, E., Vwalika, B., Mulenga, J., et al. (2019). HIV testing and counselling couples together for affordable HIV prevention in Africa. *Int. J. Epidemiol.* 48, 217–227. doi: 10.1093/ije/dyy203

Woodson, E., Goldberg, A., Michelo, C., Basu, D., Tao, S., Schinazi, R., et al. (2018). HIV transmission in discordant couples in Africa in the context of antiretroviral therapy availability. *AIDS* 32, 1613–1623. doi: 10.1097/QAD.0000000000001871

Yamaguchi, J., Vallari, A., Mcarthur, C., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2020). Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *J. Acquir. Immune Defic. Syndr.* 83, 319–322. doi: 10.1097/QAI.0000000000002246

Yue, L., Pfafferott, K. J., Baalwa, J., Conrod, K., Dong, C. C., Chui, C., et al. (2015). Transmitted virus fitness and host T cell responses collectively define divergent infection outcomes in two HIV-1 recipients. *PLoS Pathog.* 11:e1004565. doi: 10.1371/journal.ppat.1004565

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fmicb.2021.753675

# Characterization of HIV-1 Epidemic in Kyrgyzstan

**_Mariya V. Sivay[1]*, Alexei V. Totmenin[1], Daria P. Zyryanova[1], Irina P. Osipova[1], Tatyana M. Nalimova[1], Mariya P. Gashnikova[1], Vladimir V. Ivlev[1], Ivan O. Meshkov[2], Umut Z. Chokmorova[3], Elmira Narmatova[4], Ulukbek Motorov[4], Zhyldyz Akmatova[3], Nazgul Asybalieva[3], Aybek A. Bekbolotov[3], Ulan K. Kadyrbekov[3], Rinat A. Maksutov[1] and Natalya M. Gashnikova[1]_**

*[1] Department of Retroviruses, State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia, [2] National Research Center for Hematology, Moscow, Russia, [3] Republican Center of AIDS, Ministry of Health of Kyrgyzstan, Bishkek, Kyrgyzstan, [4] Osh Regional Center of AIDS Treatment and Prevention, Osh, Kyrgyzstan*

OPEN ACCESS

**Edited by:**
*Kok Keng Tee,
University of Malaya, Malaysia*

**Reviewed by:**
*Ujjwal Neogi,
Karolinska Institute, Sweden
Monick Lindenmeyer Guimarães,
Oswaldo Cruz Foundation (Fiocruz),
Brazil*

***Correspondence:**
*Mariya V. Sivay
mariya.sivay@gmail.com*

**Specialty section:**
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *05 August 2021*
***Accepted:*** *14 September 2021*
***Published:*** *15 October 2021*

**Citation:**
*Sivay MV, Totmenin AV, Zyryanova DP, Osipova IP, Nalimova TM, Gashnikova MP, Ivlev VV, Meshkov IO, Chokmorova UZ, Narmatova E, Motorov U, Akmatova Z, Asybalieva N, Bekbolotov AA, Kadyrbekov UK, Maksutov RA and Gashnikova NM (2021) Characterization of HIV-1 Epidemic in Kyrgyzstan. Front. Microbiol. 12:753675. doi: 10.3389/fmicb.2021.753675*

Kyrgyzstan has one of the highest rates of HIV-1 spread in Central Asia. In this study, we used molecular–epidemiological approaches to examine the HIV-1 epidemic in Kyrgyzstan. Samples were obtained from HIV-positive individuals who visited HIV/AIDS clinics. Partial *pol* gene sequences were used to identify HIV-1 subtypes and drug resistance mutations (DRMs) and to perform phylogenetic analysis. Genetic diversity and history reconstruction of the major HIV-1 subtypes were explored using BEAST. This study includes an analysis of 555 HIV-positive individuals. The study population was equally represented by men and women aged 1–72 years. Heterosexual transmission was the most frequent, followed by nosocomial infection. Men were more likely to acquire HIV-1 during injection drug use and while getting clinical services, while women were more likely to be infected through sexual contacts ($p < 0.01$). Heterosexual transmission was the more prevalent among individuals 25–49 years old; individuals over 49 years old were more likely to be persons who inject drugs (PWID). The major HIV-1 variants were CRF02_AG, CRF63_02A, and sub-subtype A6. Major DRMs were detected in 26.9% of the study individuals; 62.2% of those had DRMs to at least two antiretroviral (ARV) drug classes. Phylogenetic analysis revealed a well-defined structure of CRF02_AG, indicating locally evolving sub-epidemics. The lack of well-defined phylogenetic structure was observed for sub-subtype A6. The estimated origin date of CRF02_AG was January 1997; CRF63_02A, April 2004; and A6, June 1995. A rapid evolutionary dynamic of CRF02_AG and A6 among Kyrgyz population since the mid-1990s was observed. We observed the high levels of HIV-1 genetic diversity and drug resistance in the study population. Complex patterns of HIV-1 phylogenetics in Kyrgyzstan were found. This study highlights the importance of molecular–epidemiological analysis for HIV-1 surveillance and treatment implementation to reduce new HIV-1 infections.

**Keywords: HIV phylogenetics, phylodynamic analysis, Kyrgyzstan, Central Asia, HIV molecular epidemiology**

## INTRODUCTION

Kyrgyzstan, or Kyrgyz Republic, is a small country in Central Asia bordering with Kazakhstan, Uzbekistan, Tajikistan, and China. In 2020, an estimated 9,200 (8,400–9,900) people lived with HIV-1 in Kyrgyzstan with a prevalence of 0.2%[1]. The HIV-1 epidemic concentrates in the key populations, mainly persons who inject drugs (PWID) and their sexual partners (Zhao et al., 2020). Access to HIV-1 care and treatment services is far behind the global target of 95-95-95 by 2030. In 2020, 76% of people were aware of their HIV-1 status, 48% of them received antiretroviral therapy (ART), and only 43% of those were viral suppressed; the ART coverage among pregnant women was 94% (see text footnote 1).

Molecular epidemiology and phylogenetic analysis are widely used to characterize HIV-1 epidemics. HIV-1 phylogenetics has been extensively used to characterize the virus transmission networks (Wilkinson et al., 2014; Grabowski et al., 2018) and to reconstruct the viral origin and spread history (Wilkinson et al., 2014). The combination of epidemiological and sociodemographic data with the phylogenetic approaches reveals data on the risk factors associated with the HIV-1 spread (Wolf et al., 2017; Fujimoto et al., 2021), identifies sub-epidemics (Peters et al., 2016), and informs prevention and care interventions (Brenner and Wainberg, 2013).

Molecular–epidemiological and phylogenetic studies of HIV-1 infection in Central Asia, particularly in Kyrgyzstan, are very limited. One of the most recent studies in Kyrgyzstan describes a high rate (range 39–50%) of HIV-1 drug resistance in individuals with treatment failure (Lapovok et al., 2020). The study of HIV-1 sub-subtype A6 [also known as Russian A1 or A-FSU (Foley et al., 2016)] describing the transmission networks in former Soviet Union (FSU) countries concludes that the major driving source of HIV-1 transmission is migrant workers, emphasizing an extensive mobility in the region (80–90% of migrants are from Central Asia) (Aibekova et al., 2018). The study also points at the emerging HIV-1 epidemic among hetero- and homosexual populations, surpassing the parenteral transmission (Aibekova et al., 2018). Drug trafficking from Afghanistan and Tajikistan to southern Kyrgyzstan significantly increased in the mid-1990s, resulting in predominant opiate injections over the homemade drugs (Wolf et al., 2008). Between 1991 and 1999, registered drug use rate in Kyrgyzstan increased sevenfold (Wolf et al., 2008). And in the following 5-year period from 2001 to 2006, Kyrgyzstan had a 15-fold increase of HIV-1 infections, with 76% of cases detected among PWID (Wolf et al., 2008). However, in the last decade, national programs on HIV/AIDS in cooperation with international programs (Thorne et al., 2010; Zhao et al., 2020) have significantly improved HIV-1 care services in the country. While in 2010, only 9% of HIV-infected people received ART, this indicator reached 48% in 2020. In 2010, only 17% HIV-infected people on ART were virally suppressed; this number almost tripled by 2020 (see text footnote 1). By the beginning of 2021, 86.5% of HIV-positive individuals on ART had undetectable viral loads[2].

In this study, we performed the HIV-1 molecular–epidemiological survey in Kyrgyzstan. To achieve that goal, we analyzed HIV-1 genetic diversity and HIV-1 drug resistance, identified potential transmission clusters, described epidemiological characteristics of studied individuals, and reconstructed the evolutionary history of the virus.

## MATERIALS AND METHODS

### Study Population

The blood samples were collected from HIV-positive adults and children who visited local HIV/AIDS clinics of the Ministry of Health of Kyrgyzstan. All the individuals in this study were diagnosed with HIV-1 infection a year prior to the sample collection and were on ART for at least a year. Samples were collected in four provinces (Bishkek, Osh, Jalal-Abad, and Batken) in Kyrgyzstan from 2016 to 2019. Data collected from individuals in Jalal-Abad and Batken were combined (further denoted as JAB) due to the small sample size and geographical closeness of these regions. HIV-1 testing was performed at the study sites according to the national guideline[3]. Samples were shipped to the Department of Retroviruses, State Research Center of Virology and Biotechnology "Vector" (Koltsovo, Novosibirsk region, Russia) for further testing. Demographic and HIV-related characteristics of individuals were collected at the local healthcare facilities from clinical records. A woman was assigned a "pregnant" status if she was pregnant at the time of her visit for the sample collection. In this study, we refer "children" to individuals 14 years of age and under; "young adults," between 15 and 24 years old; "adults," between 25 and 49 years old; and "older adults," individuals of 50 years of age and above. The study did not recruit individuals from any group; only individuals who visited HIV/AIDS clinics were included. However, samples from children and young adults who had in-hospital acquired HIV-1 infection were particularly collected to investigate those cases.

### Amplification of HIV-1 Pol Gene Fragment and Sequencing Analysis

Viral RNA or proviral DNA was extracted using the RealBest DeltaMag kit (Vector-Best, Novosibirsk, Russia) according to the manufacturer's manual. RNA/DNA was used for the amplification of HIV-1 partial *pol* gene region coding protease and reverse transcriptase (HXB2 #K03455 reference strain coordinates: 2249–3420). Amplification of the *pol* gene fragments and sequencing analysis were performed as previously described (Maksimenko et al., 2020).

### HIV-1 Subtyping and Drug Resistance Analysis

HIV-1 subtyping was performed using automated HIV-1 subtyping tools REGA v 3.0 (Pineda-Pena et al., 2013), COMET (Struck et al., 2014), and recombinant identification program (RIP) (Siepel et al., 1995). HIV-1 subtypes were also investigated using an approximately maximum-likelihood

[1] https://www.unaids.org/en/regionscountries/countries/kyrgyzstan

[2] http://aidscenter.kg/statistika/

[3] http://med.kg/images/MyFiles/KP/2018/sbornik_KP_VISH_prikaz_903_10102017.pdf

phylogenetic method using FastTree v2.1.9 (Price et al., 2010) with HIV-1 subtype references from Los Alamos National Laboratory (LANL) HIV Sequence Database, 2020[4]. HIV-1 subtype was assigned if three out of four methods agree. Drug resistance mutations (DRMs) were assessed using Stanford HIV drug resistance database (HIVdb Program) (Shafer, 2006). DRM was considered as major according to Stanford HIV drug resistance database.

## Phylogenetic and Cluster Analyses

Phylogenetic analysis was conducted for study and background sequences; background sequences were selected from BLAST as 100 most closely related (with a BLAST score < 1e–50) to each study sequence. Sequences were aligned using MAFFT software (Katoh et al., 2017). Recombination analysis was performed for study and background sequences using RDP4 (Martin et al., 2015). Sites of major DRMs (43 codon positions) were removed. Phylogenetic trees were constructed using IQ-TREE v2 (Minh et al., 2020) under the GTR + G4 + I substitution model. Phylogenetic trees were visualized using interactive Tree of Life (iTOL) (Letunic and Bork, 2021). Monophyletic groups of study sequences with branch support ≥ 80 were considered as a distinct viral lineage (clades). Transmission clusters (≥2 individuals infected by direct/indirect transmission of genetically related HIV-1 variants) were identified by Cluster Picker v1.2.3 (Ragonnet-Cronin et al., 2013) using thresholds of 0.045 of maximum pairwise genetic distance between sequences and a branch support of 90.

## Phylodynamic Analysis

Genetic diversity and history reconstruction of the major HIV-1 subtypes were explored using a Bayesian Markov chain Monte Carlo (MCMC) phylogenetic analysis using the BEAST v1.10.4 (Suchard et al., 2018). The temporal structure of the datasets was estimated using TempEst v1.5.3 (Rambaut et al., 2016). Analysis was performed using GTR + G4 + I substitution model, with different combinations of molecular clock models (strict and log-normal uncorrelated relaxed), and coalescent models [constant size, exponential growth, Bayesian Skyline, and Gaussian Markov random field (GMRF) Bayesian Skyride]. The adjustment to the data was estimated using the log marginal likelihood estimation (MLE) using path sampling/stepping-stone sampling (PS/SS). The best-fit model was selected based on the Bayes factor (BF; BF ≥ 3 was considered significant). Two independent MCMC runs were performed for 70 × $10^6$ generations. Convergence of the chains was estimated based on the effective sampling size (ESS; cutoff value over 200 for all the parameters) in Tracer v1.7.1 (Rambaut et al., 2018).

## Statistical Analysis

Categorical variables were analyzed using modified Fisher's test; quantitative variables were analyzed using the Kruskal–Wallis test. The Monte Carlo method ($10^6$ simulations) was used for p-value (P) estimation. P correction was performed to control the false discovery rate using the Benjamini–Yekutieli procedure. Statistical analysis was performed in RStudio Team (2020) v1.1.422[5].

## Nucleotide Sequence Accession Numbers

Study sequences were submitted to GenBank under accession numbers MG798935–MG799123, MK228729–MK228833, and MW303524–MW303757 and can be found in the online repository[6].

## Ethics Statement

The study was approved by the Ethical Committee of Research and Manufacturing Association "Prevention Medicine" of Ministry of Health of Kyrgyz Republic (2017). Written consent forms were provided by all the study individuals. For individuals 18 years old and younger, written consent forms were provided by their parents or legal guardians.

# RESULTS

## Study Population

The study population included 555 individuals who resided in four Kyrgyz provinces (202 individuals from Bishkek, 341 individuals from Osh, and 12 individuals from JAB). Detailed characteristics of the study individuals are presented in **Table 1** and **Supplementary Table**. Men and women were equally represented in the study population (50.8% men vs. 49.2% women), with a median age of 31 years (range 3–72 years old). HIV-1 infection prevailed among individuals in the 25–49 age group. The median time since HIV-1 diagnosis was 8.1 years (range 1–19 years); the median time since HIV-1 diagnosis was longer in Osh than in Bishkek (9 vs. 6.46 years, $p < 0.01$). Heterosexual transmission was the most prevalent (38.7%), followed by the nosocomial infection (21.4%). In Bishkek, over a half of individuals were infected through heterosexual contacts, and 30% were PWID. In Osh, most of the individuals acquired HIV-1 while getting clinical services. Men were more likely to acquire HIV-1 infection during drug injections (33.5%) and clinical services (25.3%), while women were more likely acquire HIV-1 infection through sexual contacts (72.6%, $p < 0.01$). When stratified by age group, heterosexual transmission was more prevalent among adults; older adults were more likely to acquire HIV-1 while injecting drugs. All the individuals with nosocomial HIV-1 infection were 24 years old and younger. HIV-1 DRMs to at least two ARV drug classes were more frequent in children and young adults compared with older groups ($p < 0.01$). Individuals 24 years of age and younger tended to have HIV-1 infection caused by CRF02_AG, while older individuals were more likely to be infected with sub-subtype A6 ($p < 0.01$). Nine (1.6%) individuals were identified as men who have sex with men (MSM), with a median age of 28 years (range 23–37 years old); seven of those individuals resided in Bishkek province. The median time since HIV-1 diagnosis was 4.6 years (range

[4] http://www.hiv.lanl.gov/

[5] http://www.rstudio.com/

[6] https://www.ncbi.nlm.nih.gov/genbank/

**TABLE 1 |** Demographic and HIV-related characteristics of the 555 study individuals.

| Characteristics | Total, *n* = 555 | Bishkek, *n* = 202 | Osh, *n* = 341 | JAB, *n* = 12 | *p* | Adj. *p* |
|---|---|---|---|---|---|---|
| **Gender, *n* (%)** | | | | | | |
| Male | 282 (50.8) | 98 (48.5) | 180 (52.8) | 4 (33.3) | 0.41 | 1 |
| Female | 273 (49.2) | 104 (51.5) | 161 (47.2) | 8 (66.7) | | |
| **Age (years), *n* (%)** | | | | | | |
| 3–14 | 101 (18.2) | 5 (2.5) | 95 (27.9) | 1 (8.3) | **<0.01** | **<0.01** |
| 15–24 | 120 (21.6) | 17 (6.9) | 103 (30.2) | – | | |
| 25–49 | 271 (48.8) | 143 (71.1) | 119 (34.9) | 7 (66.7) | | |
| 50–72 | 63 (11.4) | 36 (17.9) | 24 (7) | 3 (25) | | |
| Years since HIV-1 diagnosis, median | 8.1 (range: 1–19) | 6.46 (range: 1–18) | 9 (range: 2–19) | 8.92 (range: 3–15) | **<0.01** | **<0.01** |
| **Transmission mode, n (%)** | | | | | | |
| Heterosexual | 215 (38.7) | 117 (57.9) | 90 (26.4) | 8 (66.7) | **<0.01** | **<0.01** |
| MSM | 9 (1.6) | 7 (3.5) | 2 (0.6) | – | | |
| Vertical | 70 (12.6) | 8 (4) | 60 (17.9) | 1 (8.3) | | |
| PWID | 103 (18.6) | 61 (30.2) | 40 (11.7) | 3 (25) | | |
| Nosocomial | 119 (21.4) | – | 119 (34.9) | – | | |
| Unknown/No data | 39 (7) | 9 (4.5) | 30 (8.8) | – | | |
| Pregnancy status, *n* (%) Pregnant | 33 (12.1) | 33 (31.7) | – | – | – | – |
| **HIV-1 drug resistance mutations, *n* (%)** | | | | | | |
| Yes | 149 (26.9) | 50 (24.8) | 97 (28.5) | 2 (16.7) | 0.3 | 0.91 |
| No | 406 (73.1) | 152 (75.2) | 245 (71.5) | 9 (83.3) | | |
| **Dual- and multi-class HIV-1 drug resistance, *n* (%)** | | | | | | |
| Yes | 99 (66.4) | 29 (56.9) | 70 (72.2) | 0 | 0.06 | 0.68 |
| No | 50 (33.6) | 21 (43.1) | 27 (27.8) | 11 (100) | | |
| **HIV-1 subtyping, *n* (%)** | | | | | | |
| CRF02_AG | 332 (59.8) | 95 (47) | 232 (68) | 5 (41.7) | **<0.01** | **<0.01** |
| A6 | 184 (33.2) | 90 (44.6) | 89 (26.1) | 5 (41.7) | | |
| CRF63_02A | 10 (1.8) | 4 (2) | 4 (1.2) | 2 (16.6) | | |
| Minor subtypes and recombinants | 29 (5.2) | 13 (6.4) | 16 (4.7) | – | | |

*Numbers in bold indicate statistically significant associations. JAB, Jalal-Abad and Batken provinces; Adj. p, adjusted p-value; MSM, men who have sex with men; PWID, persons who inject drugs; CRF, circulating recombinant form.*

4–6 years). Four (44.4%) of nine MSM were infected by minor HIV-1 variants (subtype B, subtype G, URF B/CRF02_AG, and URF B/G); the remaining five MSM had sub-subtype A6 (*n* = 4) and CRF02_AG (*n* = 1) HIV-1 infection. None of MSM had HIV-1 DRMs detected.

## HIV-1 Subtyping and Drug Resistance

Genotyping results were successfully obtained for 555 (89.4%) of 621 individuals. The most frequent HIV-1 subtypes were CRF02_AG [*n* = 332 (59.8%)], CRF63_02A [*n* = 10 (1.8%)], and sub-subtype A6 [*n* = 184 (33.2%)]. The remaining 29 (5.2%) sequences represent the minor HIV-1 subtypes and unique recombinant forms (URFs). Detailed HIV-1 subtype distribution is shown in **Table 2**.

Major DRMs were identified in 149 of 555 (26.9%) sequences (**Table 2** and **Figure 1**). Nucleoside reverse-transcriptase inhibitor (NRTI)-resistance mutations were detected in 107 individuals. Non-NRTI (NNRTI)-resistance mutations were detected in 138 individuals. Protease inhibitor (PI)-resistance mutations were detected in four individuals. Resistance mutations to at least two ARV drug classes were identified in 99/149 (66.4%) sequences.

## Phylogenetic and Cluster Analyses

Phylogenetic trees were constructed for each major HIV-1 subtype separately (**Figure 1**); sequences of CRF02_AG and CRF63_02A were combined, and a single phylogenetic tree was constructed. Nineteen study sequences (CRF02_AG, *n* = 11 and A6, *n* = 8) were excluded from the analysis due to short sequence length (less than 1,000 base pairs). Final datasets included 325 study and 486 background sequences of CRF02_AG/CRF63_02A, and 173 study and 1,056 background sequences of sub-subtype A6. Phylogenetic analyses of CRF02_AG/CRF63_02A and sub-subtype A6 datasets revealed that Kyrgyz sequences intermingle with

**TABLE 2 |** Distribution of HIV-1 subtypes by the study provinces.

| Subtype, n (%) | Bishkek, *n* = 202 | Osh, *n* = 341 | JAB, *n* = 12 |
|---|---|---|---|
| CRF02_AG | 95 (47%) | 232 (68%) | 5 (41.7%) |
| A6 | 90 (44.6%) | 89 (26.1%) | 5 (41.7%) |
| CRF63_02A | 4 (2%) | 4 (1.8%) | 2 (16.6%) |
| A6 URFs | 3 (1.5%) | 5 (1.5%) | |
| G | 3 (1.5%) | 2 (0.6%) | |
| B | 3 (1.5%) | 2 (0.6%) | |
| CRF02_AG/A6 | 2 (1%) | 4 (1.8%) | |
| CRF02_AG/B | 1 (0.5%) | 1 (0.3%) | |
| A6/G | – | 1 (0.3%) | |
| B/G | 1 (0.5%) | 1 (0.3%) | |

*CRF, circulating recombinant form; URF, unique recombinant form; JAB, Jalal-Abad and Batken provinces.*

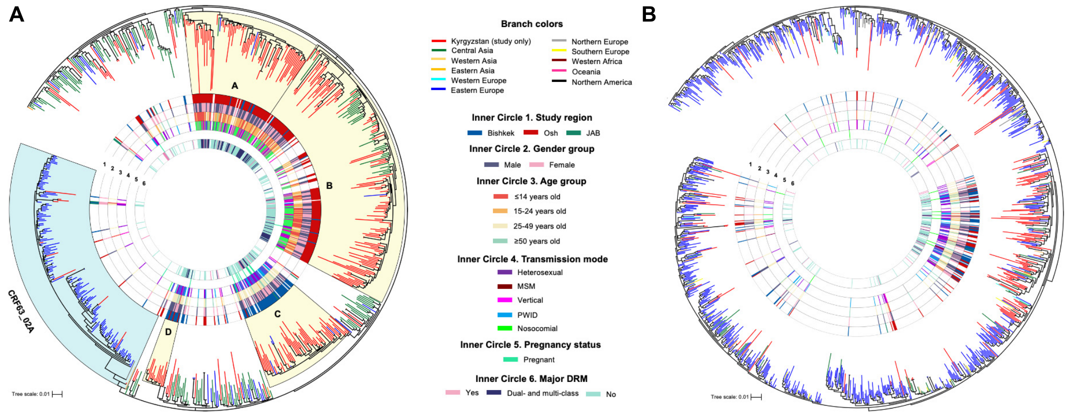


**FIGURE 1 |** Maximum-likelihood phylogenetic trees of HIV-1 *pol* sequences of CRF02_AG/CRF63_02A **(A)** and sub-subtype A6 **(B)**. Monophyletic clades (A–D) with the branch support ≥ 80 and containing for over 70% of study sequences are shaded in yellow. JAB, Jalal-Abad and Batken provinces; CRF, circulating recombinant form; MSM, men who have sex with men; PWID, persons who inject drugs; DRM, drug resistance mutation.

sequences from the neighboring countries, indicating genetic similarity and the potential common origins with corresponding variants in the Eastern Europe and Central Asia regions. The CRF02_AG/CRF63_02A tree shows the presence of the four well-defined clades composed of the local study sequences; two of those clades (A and B, 97 and 100 branch support values) are predominantly represented by the study sequences from Bishkek (**Figure 1A**). Two other clades (C and D, branch support values of 84 and 87) primarily include study sequences from Osh. This possibly represents the four distinct locally evolving sub-epidemics of CRF02_AG in Kyrgyzstan. Clades A and B were mainly represented by individuals aged 25–49 years infected by heterosexual and PWID transmission modes. Clades C and D are predominantly presented by children and young adults with the nosocomial HIV-1 infection. No noticeable clades were revealed among sub-subtype A6 study sequences; study samples scattered across the background sequences, indicating multiple independent introduction events of sub-subtype A6 to Kyrgyzstan from the neighboring countries (**Figure 1B**).

Fifty-seven putative transmission clusters representing 140 individuals (28.1%; 140/498) were detected; 34 clusters were identified among CRF02_AG sequences, 23 clusters among A6 sequences, and one cluster of CRF63_02A sequences (**Figure 2**). Most of the clusters were pairs ($n$ = 38) and triplets ($n$ = 12). Five clusters of four and one cluster of five sequences were also found. Nine clusters were identified among individuals from different study regions. Major DRMs were detected in 33 (23.6%) clustered individuals; 17 of those individuals had DRMs to at least two ARV dug classes. No statistically significant association between demographic and HIV-related characteristics of the study individuals and phylogenetic clustering was found.

## Phylodynamic Analysis

Log-normal uncorrelated relaxed molecular clock model outperforms a strict model based on the BFs in both CRF02_AG/CRF63_02A and sub-subtype A6 datasets. BFs also indicate an advantage of the Bayesian Skyline coalescent model in CRF02_AG/CRF63_02A, and a small advantage of the Gaussian Markov random field (GMRF) Bayesian Skyride coalescent model in the A6 dataset. The estimated median substitution rate for CRF02_AG/CRF63_02A was 2.79 × $10^{-3}$ (95% highest posterior density [HPD]: 1.93 × $10^{-3}$–3.69 × $10^{-3}$), and 2.55 × $10^{-3}$ (95% HPD: 1.69 × $10^{-3}$–3.43 × $10^{-3}$) for sub-subtype A6. The demographic history of HIV-1 CRF02_AG/CRF63_02A and sub-subtype A6 in Kyrgyzstan is presented in **Figure 3**. CRF02_AG/CRF63_02A Skyline plot reveals a rapid growth phase in ESS between the late 1990s and 2010 followed by a stable phase (**Figure 3A**). The estimated date of origin of the CRF02_AG epidemic in Kyrgyzstan was January 1997 (95% HPD: February 1986–May 2004). The origin date of CRF63_02A in Kyrgyzstan was estimated as April 2004 (95% HPD: January 1997–April 2009). The origin dates for the four major CRF02_AG clades were estimated as September 2000 (95% HPD: May 1993–August 2005) for clade D, May 2003 (95% HPD: May 1966–March 2008) for clade C, 2005 (95% HPD: March 1999–February 2009) for clade A, and September 2006 (95% HPD: July 2001–May 2010) for clade B. The date of origin for sub-subtype A6 was estimated as June 1995 (95% HPD: August 1985–August 2002). Sub-subtype A6 Skyride plot had a growth phase in ESS in the mid-2010s; the growth rate began to decline somewhat thereafter (**Figure 3B**).

## DISCUSSION

In this study, we performed molecular–epidemiological analysis of HIV-1 in Kyrgyzstan in four provinces, two of which (Bishkek and Osh) are most severely affected by HIV-1. Our results showed the predominant circulation of two major HIV-1 variants—CRF02_AG and sub-subtype A6. A third of HIV-infected individuals in the study had major HIV-1 DRMs: two-third of those had DRMs to at least two ARV drug classes. Phylogenetic analysis revealed a well-defined structure of CRF02_AG indicating the locally evolving sub-epidemics. The lack of a well-defined phylogenetic structure was observed for the

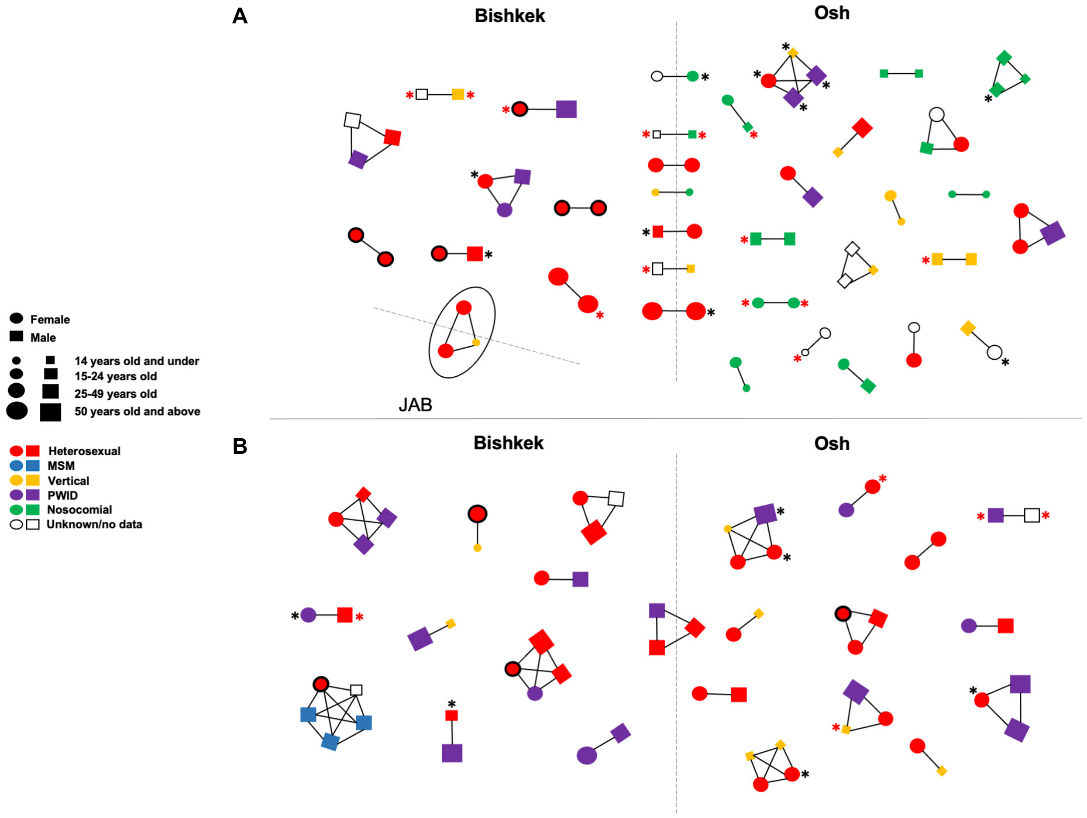


**FIGURE 2 |** Transmission clusters of study HIV-1 *pol* sequences ofCRF02_AG/CRF63_02A **(A)** and sub-subtype A6 **(B)**. Clusters were detected in IQ-TREE at maximum 0.045 genetic distances and 90 branch support thresholds. Each figure (circle/square) corresponds to a node on the phylogenetic tree; figure shape, size, and color correspond to an individual's gender, age, and HIV-1 transmission mode according to the left center footnote. The line drawn between figures indicates that HIV-1 sequences from the respective individuals in phylogenetic tree fell within the clustering thresholds. The black asterisk indicates the HIV-1 drug resistance; the red asterisk indicates the HIV-1 drug resistance to at least two ARV drug classes. Circles with the bold black border correspond to the pregnant women. The solid oval corresponds to a cluster of three individuals infected by CRF63_02A. The dashed lines divide figures corresponding to individuals from the different study regions. JAB, Jalal-Abad and Batken provinces; MSM, men who have sex with men; PWID, persons who inject drugs.

A6 sub-subtype. Small distinct HIV-1 transmission clusters were detected among study sequences. Clustering was not associated with any individual characteristics. DRMs were detected in 26.9% of clustered individuals.

HIV-1 epidemic in Kyrgyzstan is characterized by an overall low HIV-1 infection prevalence rate, a high HIV-1 prevalence rate among key populations (Deryabina et al., 2019), and a high proportion of HIV-positive children (Mansfeld et al., 2015). Our study found that most of the HIV-positive older adults were PWID; heterosexual transmission was more common in the 25–49 age group. This may indicate the shift of the HIV-1 epidemic to the general population from key groups such as PWID or change of substance consumption to non-injection drugs. Our study also found that over a half of children and young adults in Osh acquired healthcare-associated HIV-1 infection. Earlier studies described several outbreaks of hospital-acquired HIV-1 among children in Osh province in 2007 (Thorne et al., 2010) and 2011–2012 (Mansfeld et al., 2015), with the total number of over 300 reported cases. Although healthcare-associated HIV-1 infections in high-income countries are extremely low, risk of nosocomial HIV-1 infections in the resource-limited countries remains high (Ganczak and Barss, 2008; Myburgh et al., 2020). We also identified significant differences in the transmission modes between genders. Women were more likely to be infected thought the sexual contacts; PWID was a dominant transmission mode among men, followed by nosocomial infection. Similar epidemiological characteristics of HIV-positive individuals were described in the previous report (Zhao et al., 2020).

The first HIV-1 case in Kyrgyzstan was identified in 1987 in a foreigner [34]. The first local HIV-1 case was registered in 1996 (Roth et al., 2012), and HIV-1 cases in Kyrgyzstan are steadily increasing since then, like in other countries of Central Asia and Eastern Europe. CRF02_AG predominantly circulates in West Africa (Bbosa et al., 2019), but since 1999, this recombinant form is constantly detected in Central Asian countries (Laga et al., 2015; Aibekova et al., 2018). CRF02_AG circulating in Central Asia is genetically distant from that from African countries and phylogenetically represents well-supported distant clade (Mir et al., 2016). Our study showed the presence of the four main CRF02_AG region-specific lineages circulating in Kyrgyzstan. Phylodynamic reconstruction of HIV-1 epidemics in Kyrgyzstan revealed the date of origin of CRF02_AG as

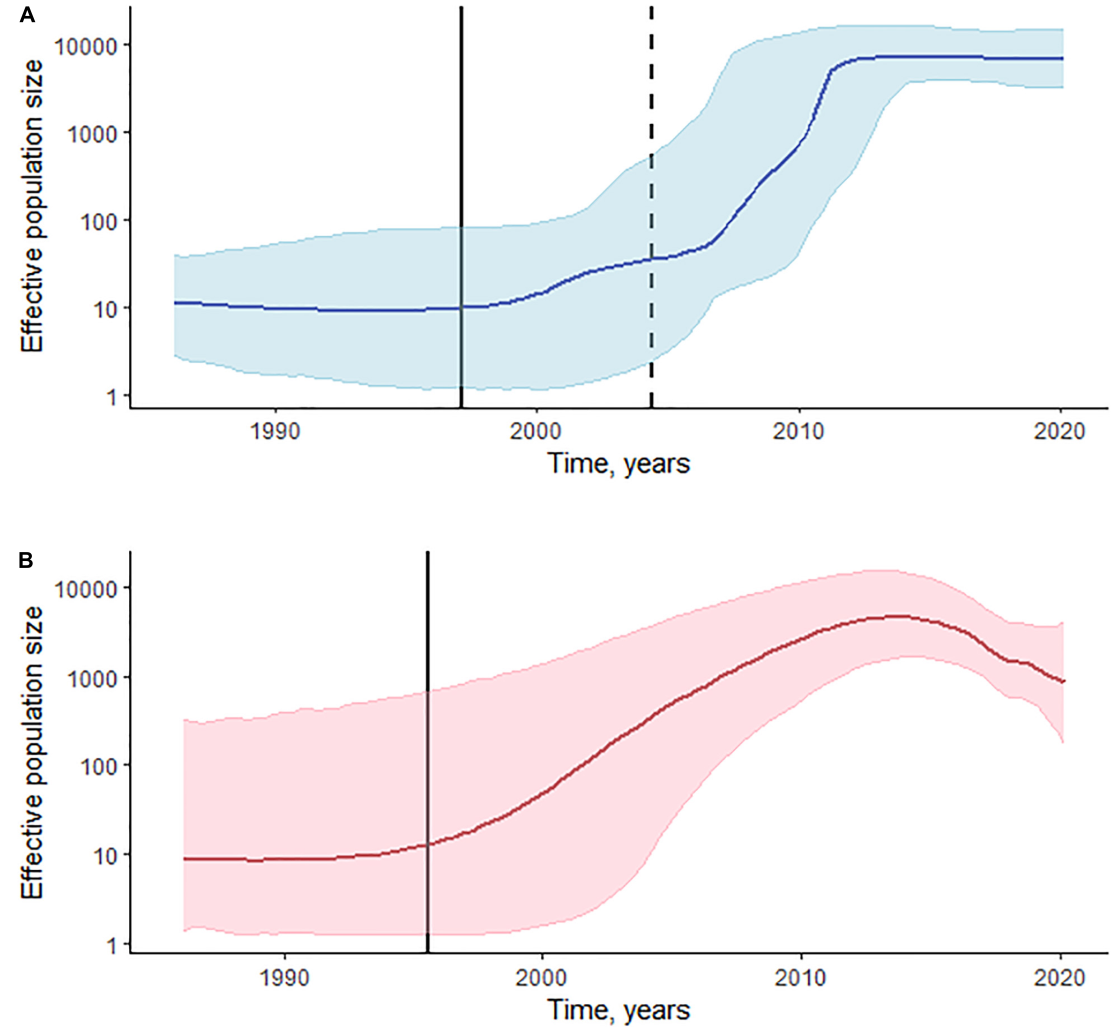


**FIGURE 3 |** Phylodynamic reconstruction of HIV-1 sCRF02_AG/CRF63_02A **(A)** and sub-subtype A6 **(B)** in Kyrgyzstan. **(A)** The Bayesian Skyline plot was reconstructed for the 325 *pol* gene sequences of CRF02_AG and 10 sequences of CRF63_02A. The horizontal bold line indicates effective population size through time; blue-shaded area represents the 95% highest posterior density (HPD). The vertical bold black line indicates the estimated origin date for CRF02_AG [January 1997 (95% HPD: February 1986–May 2004)]; the vertical dashed line indicates the estimated origin date for CRF63_02A [April 2004 (95% HPD: January 1997–April 2009)]. **(B)** The Gaussian Markov random field (GMRF) Bayesian Skyride plot was reconstructed for the 173 *pol* gene sequences of sub-subtype A6. The horizontal bold line indicates effective population size through time; pink-shaded area represents the 95% HPD. The vertical bold black line indicates the estimated origin date for sub-subtype A6 [June 1995 (95% HPD: August 1985–August 2002)].

January 1997. Osh-specific clades increased earlier than did the Bishkek-specific clades. Analysis of the characteristics of the study individuals also showed that people from Osh had HIV-1 infection for a longer period than those from Bishkek. The CRF02_AG rise in Osh coincided with the active drug trafficking from Afghanistan through the so-called "North route." The main traffic routes go to the South Kyrgyzstan (Osh province) through the Tajikistan and Uzbekistan and then spread across Kyrgyzstan and further to Kazakhstan, Russian, and other European countries[7]. Most likely, this viral variant was introduced to Kyrgyzstan by drug trafficking [Mir et al., 2016]. Bishkek-specific clades were predominantly represented by adults who acquired HIV-1 infection through heterosexual contacts and while injection of drugs, indicating an extensive mixing between high-risk and heterosexual populations. Osh-specific clades are represented by children and young adults who acquired healthcare-associated HIV-1 infection and represented by in-hospital HIV-1 outbreak sequences. These age groups play an important role in the local HIV-1 sub-epidemic, and targeted HIV-1 care interventions should be implemented to limit HIV-1 spread and improve healthcare services. HIV-1 sub-subtype A6 epidemic in FSU countries is characterized by monophyletic phylogenetic tree structure, suggesting a common ancestor for all the viruses (Díez-Fuertes et al., 2015). Our results revealed a significant geographic dissemination of the sub-subtype A6 across the FSU countries with the lack of

[7] https://www.unodc.org/documents/publications/NR_Report_21.06.18_low.pdf

country-specific clades. HIV-1 sub-subtype A6 phylogenetics discovered multiple introductions of the virus to Kyrgyzstan, which are most likely associated with the intense migration between Kyrgyzstan and the neighboring countries. The origin date for sub-subtype A6 in Kyrgyzstan was estimated as June 1995. The HIV-1 effective population size of these subtypes had an initial growing phase with the further stable phase for CRF02_AG/CRF63_02A and decline phase for sub-subtype A6. CRF63_02A (formally known as CRF02_AG recombinant) was initially described in 2012 as a viral variant that caused an outbreak in the Asian part of Russia in 2006 (Baryshev et al., 2012). In our study, a very limited number of CRF63_02A sequences were detected. A small number (1.6%) of HIV-positive MSM was found in the study. Almost a half of those individuals were infected by minor HIV-1 variants (subtypes B and G, and URFs B/CRF02_AG and B/G); the remaining five MSM had sub-subtype A6 or CRF02_AG HIV-1 infection. Despite the very limited number of MSM in our study, these data indicate the mixing of HIV-1 transmissions between MSM and general population. Also, the high HIV-1 genetic diversity among MSM could potentially be one of the driving forces for further increase of HIV-1 variety in Kyrgyzstan. Data of HIV-1 prevalence among MSM in Central Asia, and particularly in Kyrgyzstan, are very limited due to criminalization of same-sex relationships (Thorne et al., 2010; Wirtz et al., 2013). However, previous studies reported HIV-1 prevalence rate among MSM at around 1% (Thorne et al., 2010; El-Bassel et al., 2013). Further studies are needed to better understand HIV-1 epidemiology and HIV-related risk behavior among this high-risk group in Kyrgyzstan. This study showed that groups of at-risk people could contribute to the HIV-1 epidemic progression by maintaining new infections within the risk group or by linking infections between different populations.

Our study has several limitations. Children and young adults were sampled more intensely than older individuals, which could have biased some of the observed numbers. In this study, samples from children and young adults were deliberately collected from those who potentially acquired HIV-1 infection during in-hospital outbreaks. Also, this study provides information on HIV-positive pregnant women in Bishkek province only. Data on pregnancy status from other provinces were not available. In this study, we detected a small number of transmission clusters sized two to four individuals due to a small sampling fraction. The study dataset represents only around 6% (555 HIV-positive study individuals of an estimated 9,200 HIV-infections in 2020 (see text footnote 1) of the total HIV-1 detected cases in Kyrgyzstan. Major DRMs were detected in 23.6% of individuals in the HIV-1 transmission clusters. The study as performed does not power to reveal whether DRMs occurred due to the therapy failure or whether originating from the transmission of the drug-resistant virus. HIV-positive individuals not linked to care were not included in the study, since only individuals who visited healthcare facilities were included.

To our knowledge, this is the first study that provides comprehensive data on the HIV-1 epidemic in Kyrgyzstan. We identified complex patterns of HIV-1 phylogenetics among the key and general populations and observed high levels of HIV-1 genetic diversity and drug resistance. Further studies are needed, accompanied by extensive public health interventions to limit HIV-1 spread and improve future HIV-1 care services.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

## ETHICS STATEMENT

The study was approved by Ethical Committee of Research and Manufacturing Association "Prevention Medicine" of Ministry of Health of Kyrgyz Republic. Written consent forms were provided by all the study individuals. For individuals 18 years old and younger, written consent forms were provided by their parent or legal guardian.

## AUTHOR CONTRIBUTIONS

MS, AT, and NG planned and designed the study. DZ, IO, TN, MG, and VI performed HIV genotyping and collected the sequences. AT submitted the sequences. MS performed the analysis of the epidemiological data, performed the phylogenetic and phylodynamic analyses, produced the illustrations, and wrote the manuscript. NG supervised the project and edited the manuscript. IM performed the statistical analysis. UC, EN, UM, ZA, NA, AB, and UK collected the epidemiological data and assisted with sample collection at the study sites. All authors participated in the final review of the manuscript.

## FUNDING

This study was supported by the Directives of the Russian Government Nos. 2314-p dated November 14, 2015 and No. 2656-p dated December 1, 2018.

## ACKNOWLEDGMENTS

We thank the study individuals for their contributions to the study. We also thank the staff at the study HIV/AIDS clinics for sample management and laboratory testing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.753675/full#supplementary-material

## REFERENCES

Aibekova, L., Foley, B., Hortelano, G., Raees, M., Abdraimov, S., Toichuev, R., et al. (2018). Molecular epidemiology of HIV-1 subtype A in former Soviet Union countries. *PLoS One* 13:e0191891. doi: 10.1371/journal.pone.0191891

Baryshev, P. B., Bogachev, V. V., and Gashnikova, N. M. (2012). Genetic characterization of an isolate of HIV type 1 AG recombinant form circulating in Siberia, Russia. *Arch. Virol.* 15, 2335–2341. doi: 10.1007/s00705-012-1442-4

Bbosa, N., Kaleebu, P., and Ssemwanga, D. (2019). HIV subtype diversity worldwide. *Curr. Opin. HIV AIDS* 14, 153–160. doi: 10.1097/COH.0000000000000534

Brenner, B. G., and Wainberg, M. A. (2013). Future of phylogeny in HIV prevention. *J. Acquir. Immune Defic. Syndr.* 63 Suppl 2, 248–254. doi: 10.1097/QAI.0b013e3182986f96

Deryabina, A. P., Patnaik, P., and El-Sadr, W. M. (2019). Underreported injection drug use and its potential contribution to reported increase in sexual transmission of HIV in Kazakhstan and Kyrgyzstan. *Harm Reduct. J.* 16:1. doi: 10.1186/s12954-018-0274-2

Díez-Fuertes, F., Cabello, M., and Thomson, M. M. (2015). Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries. *Infect. Genet. Evol.* 33, 197–205. doi: 10.1016/j.meegid.2015.05.003

El-Bassel, N., Strathdee, S. A., and El Sadr, W. M. (2013). HIV and people who use drugs in central Asia: confronting the perfect storm. *Drug Alcohol Depend* 132 Suppl 1, S2–S6. doi: 10.1016/j.drugalcdep.2013.07.020

Foley, B. T., Leitner, T., Paraskevis, D., and Peeters, M. (2016). Primate immunodeficiency virus classification and nomenclature: review. *Infect. Genet. Evol.* 46, 150–158. doi: 10.1016/j.meegid.2016.10.018

Fujimoto, K., Bahl, J., Wertheim, J. O., Del Vecchio, N., Hicks, J. T., Damodaran, L., et al. (2021). Methodological synthesis of Bayesian phylodynamics, HIV-TRACE, and GEE: HIV-1 transmission epidemiology in a racially/ethnically diverse Southern U.S. context. *Sci. Rep.* 11:3325. doi: 10.1038/s41598-021-82673-8

Ganczak, M., and Barss, P. (2008). Nosocomial HIV infection: epidemiology and prevention – a global perspective. *AIDS Rev.* 10, 47–61.

Grabowski, M. K., Herbeck, J. T., and Poon, A. F. Y. (2018). Genetic cluster analysis for HIV prevention. *Curr. HIV/AIDS Rep.* 15, 182–189. doi: 10.1007/s11904-018-0384-1

HIV Sequence Database (2020). *HIV Sequence Database*. Available online at: http://www.hiv.lanl.gov/ (accessed November 15, 2020).

Katoh, K., Rozewicki, J., and Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* 20, 1160–1166. doi: 10.1093/bib/bbx108

Laga, V., Lapovok, I., Kazennova, E., Ismailova, A., Beisheeva, N., Asybalieva, N., et al. (2015). The genetic variability of HIV-1 in Kyrgyzstan: the spread of CRF02_AG and subtype A1 recombinants. *J. HIV AIDS* 1. doi: 10.16966/2380-5536.106

Lapovok, I. A., Saleeva, D. V., Lopatukhin, A. E., Kirichenko, A. A., Murzakova, A. V., Bekbolotov, A. A., et al. (2020). HIV-1 genetic characteristics in patients having experienced virological failure of therapy in the Kyrgyz Republic in 2017–2018. *Epidemiol. Infect. Dis.* 3, 105–111. doi: 10.18565/epidem.2020.10.3.105-11

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Maksimenko, L. V., Totmenin, A. V., Gashnikova, M. P., Astakhova, E. M., Shudarnov, S. E., Ostapova, T. S., et al. (2020). Genetic diversity of HIV-1 in Krasnoyarsk Krai: area with high levels of HIV-1 recombination in Russia. *Biomed. Res. Int.* 10:9057541. doi: 10.1155/2020/9057541

Mansfeld, M., Ristola, M., and Likatavicius, G?. (2015). *HIV Programme Review in Kyrgyzstan*. Copenhagen: WHO Regional Office for Europe.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 1:vev003. doi: 10.1093/ve/vev003

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Ministry of Health of Kyrgyz Republic (2017). *Clinical Protocols on HIV Infection for Outpatient and Inpatient Services. Bishkek*. Available online at: http://med.kg/images/MyFiles/KP/2018/sbornik_KP_VISH_prikaz_903_10102017.pdf (accessed May 20, 2021).

Mir, D., Jung, M., Delatorre, E., Vidal, N., Peeters, M., and Bello, G. (2016). Phylodynamics of the major HIV-1 CRF02_AG African lineages and its global dissemination. *Infect. Genet. Evol.* 46, 190–199. doi: 10.1016/j.meegid.2016.05.017

Myburgh, D., Rabie, H., Slogrove, A. L., Edson, C., Cotton, M. F., and Dramowski, A. (2020). Horizontal HIV transmission to children of HIV-uninfected mothers: a case series and review of the global literature. *Int. J. Infect. Dis.* 98, 315–320. doi: 10.1016/j.ijid.2020.06.081

Peters, P. J., Pontones, P., Hoover, K. W., Patel, M. R., Galang, R. R., Shields, J., et al. (2016). HIV infection linked to injection use of oxymorphone in Indiana, 2014–2015. *N. Engl. J. Med.* 375, 229–239. doi: 10.1056/NEJMoa1515195

Pineda-Pena, A. C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., et al. (2013). Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* 19, 337–348. doi: 10.1016/j.meegid.2013.04.032

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Leigh Brown, A. J., et al. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinformatics* 6:317. doi: 10.1186/1471-2105-14-317

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using tempest (formerly Path-O-Gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

Roth, S., Elfving, R., and Khan, A. R. (2012). *HIV/AIDS Vulnerabilities in Regional Transport Corridors in the Kyrgyz Republic and Tajikistan*. Manila: Asian Development Bank.

RStudio Team (2020). *RStudio: Integrated Development for R. RStudio*. Boston, MA: RStudio Team.

Shafer, R. W. (2006). Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* 194(Suppl. 1), 51–58. doi: 10.1086/505356

Siepel, A. C., Halpern, A. L., Macken, C., and Korber, B. T. (1995). A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* 11, 1413–1416. doi: 10.1089/aid.1995.11.1413

Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C., and Bercoff, D. P. (2014). COMET: adaptive context-390based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res.* 39142:e144. doi: 10.1093/nar/gku739

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

Thorne, C., Ferencic, N., Malyuta, R., Mimica, J., and Nieniec, T. (2010). Central Asia: hotspot in the worldwide HIV epidemic. *Lancet Infect. Dis.* 10, 479–488. doi: 10.1016/S1473-3099(10)70118-3

Wilkinson, E., Engelbrecht, S., and de Oliveira, T. (2014). Detection of transmission clusters of HIV-1 subtype C over a 21-year period in Cape Town, South Africa. *PLoS One* 9:e109296. doi: 10.1371/journal.pone.0109296

Wirtz, A. L., Kirey, A., Peryskina, A., Houdart, F., and Beyrer, C. (2013). Uncovering the epidemic of HIV among men who have sex with men in Central Asia. *Drug Alcohol Depend* 132(Suppl. 1), S17–S24. doi: 10.1016/j.drugalcdep.2013.06.031

Wolf, E., Herbeck, J. T., Van Rompaey, S., Kitahata, M., Thomas, K., Pepper, G., et al. (2017). Phylogenetic evidence of HIV-1 transmission between adult and adolescent men who have sex with men. *AIDS Res. Hum. Retroviruses* 33, 318–322. doi: 10.1089/AID.2016.0061

Wolf, D., Elovich, R., Boltaev, A., and Pulatov, D. (2008). *HIV in Central Asia: Tajikistan, Uzbekistan and Kyrgyzstan. Public Health Aspects of HIV/AIDS in Low- and Middle-Income Countries*. New York, NY: Springer.

Zhao, F., Benedikt, C., and Wilson, D. (2020). *Tackling the World's Fastest-Growing HIV Epidemic: More Efficient HIV Responses in Eastern Europe and Central Asia. Human Development Perspectives*. Washington, DC: World Bank.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 15 November 2021
doi: 10.3389/fmicb.2021.774386

# Identification of CRF66_BF, a New HIV-1 Circulating Recombinant Form of South American Origin

**Joan Bacqué[1], Elena Delgado[1], Sonia Benito[1], María Moreno-Lorenzo[1], Vanessa Montero[1], Horacio Gil[1], Mónica Sánchez[1], María Carmen Nieto-Toboso[2], Josefa Muñoz[2], Miren Z. Zubero-Sulibarria[2], Estíbaliz Ugalde[2], Elena García-Bodas[1], Javier E. Cañada[1], Jorge del Romero[3], Carmen Rodríguez[3], Iciar Rodríguez-Avial[4], Luis Elorduy-Otazua[5], José J. Portu[6], Juan García-Costa[7], Antonio Ocampo[8], Jorge J. Cabrera[8] and Michael M. Thomson[1]***

[1] HIV Biology and Variability Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain, [2] Hospital Universitario de Basurto, Bilbao, Spain, [3] Centro Sanitario Sandoval, IdISSC, Madrid, Spain, [4] Hospital Clínico San Carlos, Madrid, Spain, [5] Hospital Universitario de Cruces, Barakaldo, Spain, [6] Hospital Universitario de Araba, Vitoria, Spain, [7] Complejo Hospitalario Universitario de Ourense, Ourense, Spain, [8] Complejo Hospitalario Universitario de Vigo, Vigo, Spain

OPEN ACCESS

**Edited by:**
Miguel Angel Martinez,
IrsiCaixa, Spain

**Reviewed by:**
Gonzalo Jose Bello Bentancor,
Oswaldo Cruz Foundation (FIOCRUZ),
Brazil
Marcel Tongo,
Center for Research on Emerging
and Re-Emerging Diseases
(CREMER), Cameroon

***Correspondence:**
Michael M. Thomson
mthomson@isciii.es

**Specialty section:**
This article was submitted to
Virology,
a section of the journal
Frontiers in Microbiology

**Received:** 11 September 2021
**Accepted:** 25 October 2021
**Published:** 15 November 2021

**Citation:**
Bacqué J, Delgado E, Benito S, Moreno-Lorenzo M, Montero V, Gil H, Sánchez M, Nieto-Toboso MC, Muñoz J, Zubero-Sulibarria MZ, Ugalde E, García-Bodas E, Cañada JE, del Romero J, Rodríguez C, Rodríguez-Avial I, Elorduy-Otazua L, Portu JJ, García-Costa J, Ocampo A, Cabrera JJ and Thomson MM (2021) Identification of CRF66_BF, a New HIV-1 Circulating Recombinant Form of South American Origin. Front. Microbiol. 12:774386. doi: 10.3389/fmicb.2021.774386

Circulating recombinant forms (CRFs) are important components of the HIV-1 pandemic. Among 110 reported in the literature, 17 are BF1 intersubtype recombinant, most of which are of South American origin. Among these, all 5 identified in the Southern Cone and neighboring countries, except Brazil, derive from a common recombinant ancestor related to CRF12_BF, which circulates widely in Argentina, as deduced from coincident breakpoints and clustering in phylogenetic trees. In a HIV-1 molecular epidemiological study in Spain, we identified a phylogenetic cluster of 20 samples from 3 separate regions which were of F1 subsubtype, related to the Brazilian strain, in protease-reverse transcriptase (Pr-RT) and of subtype B in integrase. Remarkably, 14 individuals from this cluster (designated BF9) were Paraguayans and only 4 were native Spaniards. HIV-1 transmission was predominantly heterosexual, except for a subcluster of 6 individuals, 5 of which were men who have sex with men. Ten additional database sequences, from Argentina (*n* = 4), Spain (*n* = 3), Paraguay (*n* = 1), Brazil (*n* = 1), and Italy (*n* = 1), branched within the BF9 cluster. To determine whether it represents a new CRF, near full-length genome (NFLG) sequences were obtained for 6 viruses from 3 Spanish regions. Bootscan analyses showed a coincident BF1 recombinant structure, with 5 breakpoints, located in $p17^{gag}$, integrase, gp120, gp41-*rev* overlap, and *nef*, which was identical to that of two BF1 recombinant viruses from Paraguay previously sequenced in NFLGs. Interestingly, none of the breakpoints coincided with those of CRF12_BF. In a maximum likelihood phylogenetic tree, all 8 NFLG sequences grouped in a strongly supported clade segregating from previously identified CRFs and from the CRF12_BF "family" clade. These results allow us to identify a new HIV-1 CRF, designated CRF66_BF. Through a Bayesian coalescent analysis, the most recent common ancestor of CRF66_BF was estimated around 1984 in South America, either in Paraguay or

Argentina. Among Pr-RT sequences obtained by us from HIV-1-infected Paraguayans living in Spain, 14 (20.9%) of 67 were of CRF66_BF, suggesting that CRF66_BF may be one of the major HIV-1 genetic forms circulating in Paraguay. CRF66_BF is the first reported non-Brazilian South American HIV-1 CRF_BF unrelated to CRF12_BF.

**Keywords: HIV-1, circulating recombinant form, molecular epidemiology, phylogeny, phylodynamics**

## INTRODUCTION

One of the most distinctive features of HIV-1 evolution is its high recombinogenic potential, possibly the greatest among human pathogens, which is reflected in the high frequency of unique recombinant forms (URFs), each generated in a dually- or multiply-infected individual, found wherever different genetic forms circulate in the same population (Nájera et al., 2002). Some of the HIV-1 recombinant forms have spread beyond a group of epidemiologically linked individuals, in which case they are designated circulating recombinant forms (CRFs) (Robertson et al., 2000). Currently, 110 CRFs have been reported in the literature and their number is increasing incessantly, due to both the generation of new CRFs and the identification of old previously undocumented CRFs. The proportion of CRFs in the HIV-1 pandemic has increased over time, representing around 17% infections in 2010-2015 (Hemelaar et al., 2020). Among CRFs, the most numerous are those derived from subtype B and subusbtype F1, 17 of which have been identified, most of them originated in South America, derived from the F1 variant circulating in Brazil (Louwagie et al., 1994). The first CRF_BF identified in South America was CRF12_BF, which circulates widely in Argentina and Uruguay, where URFs related to CRF12_BF are frequently found (Thomson et al., 2000, 2002; Carr et al., 2001). Subsequently, 4 CRF_BFs related to CRF12_BF, as evidenced by shared breakpoints and phylogenetic clustering, were identified in the Southern Cone of South America or neighboring countries, CRF17_BF (Aulicino et al., 2012), CRF38_BF (Ruchansky et al., 2009), CRF44_BF (Delgado et al., 2010), and CRF89_BF (Delgado et al., 2021), the last three having clear country associations, with Uruguay, Chile, and Bolivia, and Peru, respectively. Due to their common ancestry, these 5 CRFs and related URFs have been proposed to constitute a "family" of recombinant viruses (Thomson and Nájera, 2005; Zhang et al., 2010; Delgado et al., 2021). By contrast, all CRF_BFs identified in Brazil are unrelated to CRF12_BF (De Sá Filho et al., 2006; Guimarães et al., 2008; Sanabani et al., 2010; Pessôa et al., 2014; Reis et al., 2017, 2019). Here we report the identification of a new CRF_BF, found mainly in Paraguayan immigrants in Spain and also identified in Paraguay and Argentina. Interestingly, unlike all South American CRF_BFs identified to date outside of Brazil, it has no relationship with CRF12_BF.

## MATERIALS AND METHODS

### Samples

Plasma samples from HIV-1-infected individuals were collected in 14 Spanish regions for antiretroviral drug resistance tests and for a molecular epidemiological study. The study was approved by the Committee of Research Ethics of Instituto de Salud Carlos III, Majadahonda, Madrid, Spain. The study did not require written informed consent by the study participants, as it used samples and data collected as part of routine clinical practice and patients' data were anonymized without retaining data allowing individual identification.

### RNA Extraction, Reverse Transcription-Polymerase Chain Reaction Amplification, and Sequencing

An ∼1.4 kb *pol* fragment in protease-reverse transcriptase (Pr-RT) was amplified from plasma RNA by Reverse Transcription-Polymerase Chain Reaction (RT-PCR) followed by nested PCR as described previously (Delgado et al., 2015) and sequenced with the Sanger method using a capillary automated sequencer. Some samples were also subject to amplification and sequencing of integrase. Near full-length genome (NFLG) sequences were obtained for selected samples by RT-PCR/nested PCR amplification from plasma RNA in four overlapping segments and sequenced by the Sanger method, as described (Delgado et al., 2002; Sierra et al., 2005; Cañada et al., 2021). Newly derived sequences are deposited in GenBank under accessions MK298150, OK011530-OK011552.

### Phylogenetic Sequence Analyses

Sequences were aligned with MAFFT v7 (Katoh and Standley, 2013). Initial phylogenetic trees with all Pr-RT sequences obtained by us were constructed via approximate maximum likelihood with FastTree v2.1.10 (Price et al., 2010), using the general time reversible evolutionary model with CAT approximation to account for among-site rate heterogeneity, with assessment of node support with Shimodaira-Hasegawa (SH)-like local support values (Guindon et al., 2010). A second phylogenetic tree with the Pr-RT sequences of interest and all Pr-RT sequences from the Los Alamos HIV Sequence Database (Los Alamos National Laboratory, 2021) labeled as being from F1 subsubtype or BF1 recombinant viruses, excluding those sequences identified as BF1 recombinant within Pr-RT, according to the analyses with REGA HIV-1 Subtyping Tool v3 (Pineda-Peña et al., 2013), was also constructed with FastTree, as described above. Subsequent maximum likelihood (ML) trees with sequences of interest were constructed with W-IQ-Tree (Trifinopoulos et al., 2016), using the best-fit substitution model selected by ModelFinder program (Kalyaanamoorthy et al., 2017), with assessment of node support with the ultrafast bootstrap approximation approach (Hoang et al., 2018). Trees were visualized with MEGA v7.0 (Kumar et al., 2016) or FigTree v1.4.2 (Rambaut[1]).

Mosaic structures were analyzed by bootscanning (Salminen et al., 1995) with SimPlot v1.3.5 (Lole et al., 1999). In these analyses, trees were constructed using the neighbor-joining method with the Kimura 2-parameter model and a window width of 400 nucleotides. Recombinant segments identified with SimPlot were further phylogenetically analyzed via ML with W-IQ-Tree. Intersubtype breakpoint locations were also determined with jpHMM (Schultz et al., 2009).

## Temporal and Geographical Estimations

The time and the location of the most recent common ancestor (MRCA) of the identified CRF was estimated using Pr-RT sequences with the Bayesian Markov chain Monte Carlo (MCMC) coalescent method implemented in BEAST v1.10.4 (Suchard et al., 2018), using a discrete trait approach. Prior to the BEAST analysis, the existence of temporal signal in the dataset was assessed with TempEst v1.5.3 (Rambaut et al., 2016), which determines the correlation of genetic divergence among sequences (measured as root-to-tip distance) with time. The BEAST analysis was performed using the SRD06 codon-based evolutionary model (with two codon position partitions: 1st+$2^{nd}$, and 3rd) (Shapiro et al., 2006). A uniform prior distribution ($2 \times 10^{-4} - 2 \times 10^{-2}$ subs/site/year) was used for the substitution rate. We also specified an uncorrelated lognormal relaxed clock and a Bayesian SkyGrid coalescent tree prior (Gill et al., 2013). In the SkyGrid analysis, the number of grid points was set at 50 and the time at last transition point at 60 years. The MCMCs were run for 50 million generations. We performed runs in duplicate, combining the posterior tree files with LogCombiner v1.10.4. Mixing and convergence were checked with Tracer v1.6, ensuring that effective sample size values of all parameters were $>200$. The posterior distribution of trees was summarized in a maximum clade credibility (MCC) tree with TreeAnnotator v1.10.4, after removal a 10% burn-in. MCC trees were visualized with FigTree. Parameter uncertainty was summarized in 95% highest posterior density (HPD) intervals.

# RESULTS

## Identification of a BF Intersubtype Recombinant Cluster and Epidemiological Associations

In a molecular epidemiology study on HIV-1 in Spain we identified a cluster of 20 viruses of F1 subsubtype in Pr-RT, that in integrase, sequenced in 4 samples, were of subtype B, which was designated BF9. Inclusion in the phylogenetic analyses of Pr-RT sequences from all viruses in the Los Alamos HIV Sequence Database (Los Alamos National Laboratory, 2021) classified as being of F1 subsubtype or BF1 recombinant, excluding those sequences that were BF1 recombinant within Pr-RT, according to REGA HIV-1 Subtyping Tool, identified 10 additional viruses belonging to BF9, from Argentina (*n* = 4), Spain (*n* = 3), Paraguay (*n* = 1), Brazil (*n* = 1), and Italy (*n* = 1) (**Figure 1** and **Supplementary Figure 1**). Pr-RT sequences of the BF9 cluster were most closely related to F1 viruses of the Brazilian variant (**Figure 1**). Epidemiological data of the 20 samples of the BF9 cluster from Spain processed by us are shown in **Table 1**. Remarkably, 14 individuals were from Paraguay [with the remaining 6 being from Spain (*n* = 4), Argentina, and Equatorial Guinea] and all 3 other database sequences from samples collected in Spain were from Latin Americans, one each from Paraguay, Argentina, and an unspecified country. Transmission was predominantly heterosexual, but 7 were men who have sex with men (MSM), the sequences of 5 of whom branched in a subcluster (**Figure 1**).

## Analyses of Near Full-Length Genome Sequences and Identification of a New Circulating Recombinant Form

To determine whether viruses from the BF9 cluster represent a new CRF, we obtained NFLG sequences from 6 samples from 3 Spanish regions and analyzed their mosaic structures by bootscanning. Two additional NFLG sequences of BF recombinant viruses from databases were also analyzed by bootscanning, both from Paraguay: 02PY_PSP0094, that branched in the BF9 cluster in Pr-RT, and 02PY_PSP0093, that showed high similarity to NFLGs of the BF9 cluster in BLAST searches of the Los Alamos database. All 8 analyzed genomes showed coincident mosaic structures, with 5 breakpoints, located in $p17^{gag}$, integrase, gp120, gp41-*rev* overlap, and *nef* (**Figure 2**). Breakpoints were more precisely located using the midpoint of B-F1 transitions, according to the positions where 75% consensuses of subtype B and the F1 Brazilian strain genomes differ, in HXB2 positions 950, 4327, 6486, 8498, and 9161. Breakpoint locations were also determined with jpHMM (**Supplementary Table**), which also found 5 breakpoints for each virus in intervals overlapping those of the other analyzed viruses and the 75% consensus B-F1 transition intervals in all cases except the breakpoint interval in *nef* of MD497796, that did not overlap the consensus B-F1 transition interval, and that in $p17^{gag}$ of PV106451, that was not detected by jpHMM. ML phylogenetic trees constructed with each interbreakpoint fragment confirmed the subtype assignation determined with bootscanning (**Figure 3**).

In an ML tree constructed with the 7 NFLG genomes of the BF9 cluster and 02PY_PSP0093, all 8 genomes grouped in a strongly supported clade segregating away from all other CRF_BFs and of the clade formed by the 5 CRF_BFs of the CRF12_BF family (**Figure 4**). It should be pointed out that 02PY_PSP0093 did not branch in the BF9 cluster in the tree of Pr-RT (**Supplementary Figure 1**), which suggests that the Pr-RT segment of this virus could derive from secondary recombination with an F1 strain different from the parental F1 strain of all other BF9 viruses.

These results, therefore, allow to define a new CRF, which was designated CRF66_BF, whose mosaic structure is shown in **Figure 5**.

[1] http://tree.bio.ed.ac.uk/software/figtree/

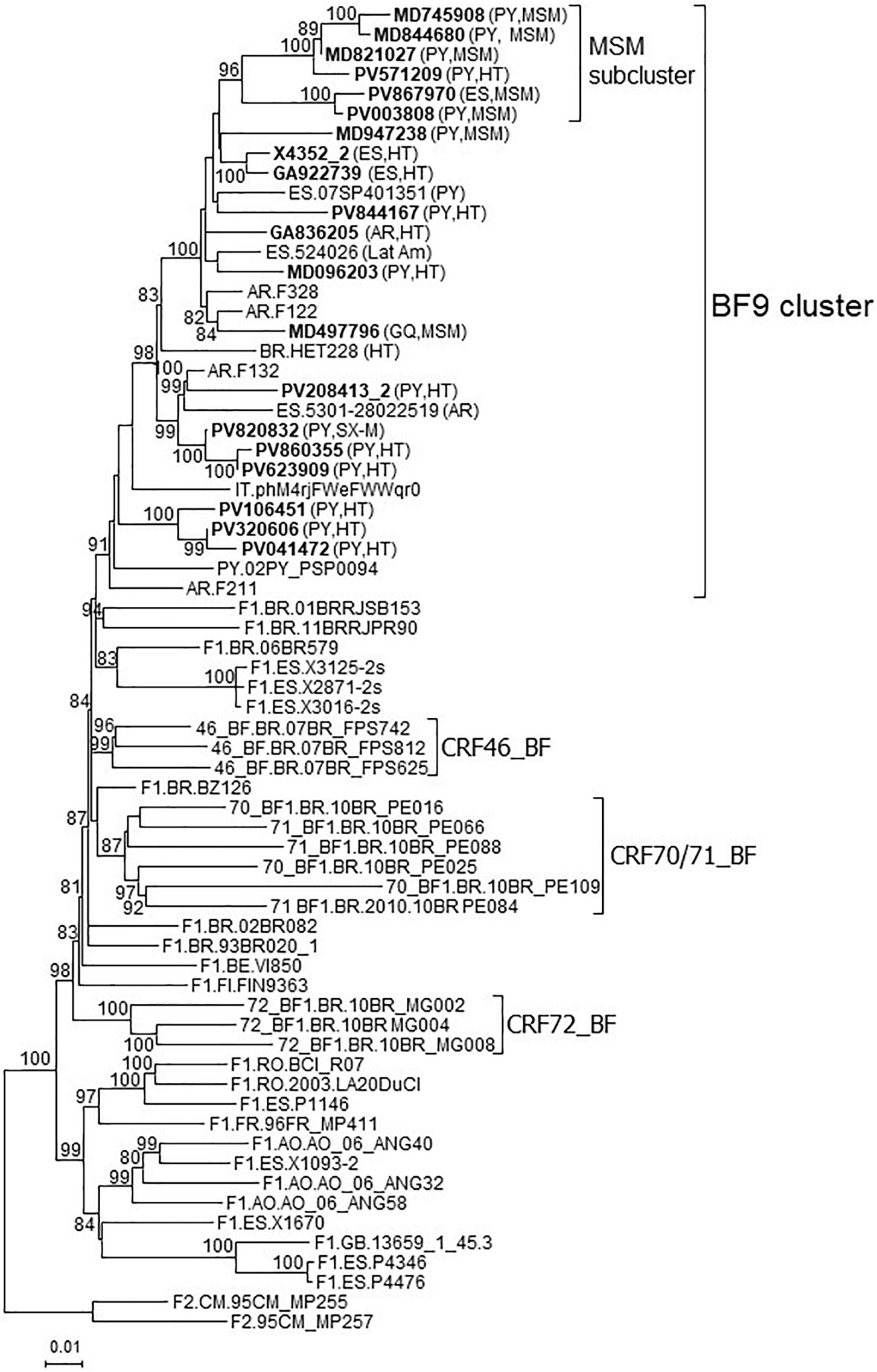


**FIGURE 1 |** Maximum likelihood tree of Pr-RT sequences of BF9 cluster constructed with IQ-Tree. Sequences used in this analysis were those from the BF9 cluster shown in the tree of **Supplementary Figure 1**, constructed with FastTree, plus sequences of F1 subsubtype from different countries and of CRF_BFs that are of nonrecombinant F1 subsubtype in Pr-RT, plus two F2 sequences used as outgroups. Names of sequences obtained by us, all collected in Spain, are in bold type. In database sequences, the country of sample collection is indicated before the virus name with the 2-letter ISO country code. After the names of viruses of the BF9 cluster, the 2-letter ISO code of country of origin of the patient and/or the transmission route, when known, are shown in parentheses. Only ultrafast bootstrap values ≥80% are shown. PY: Paraguay; AR: Argentina; ES: Spain; BR: Brazil; IT: Italy; GQ: Equatorial Guinea; Lat Am: Latin America (unknown country); MSM: man who has sex with men; HT: heterosexual; SX-M: male with unspecified sexual acquisition of HIV-1.

**TABLE 1 |** Epidemiological data of patients and GenBank accessions of sequences.

| Sample ID | City of sample collection | Region of sample collection | Year of sample collection | Year of HIV diagnosis | Gender | Transmission route* | Country of origin | GenBank accessions |
|---|---|---|---|---|---|---|---|---|
| X4352_2 | Vigo | Galicia | 2017 | 2017 | M | HT | Spain | MK298150 (NFLG) |
| GA836205 | Vigo | Galicia | 2020 | 2020 | M | HT | Argentina | OK011531 (Pr-RT)<br>OK011530 (integrase) |
| GA922739 | Ourense | Galicia | 2018 | 2017 | M | HT | Spain | OK011532 (NFLG) |
| MD096203 | Madrid | Madrid | 2017 | 2011 | F | HT | Paraguay | OK011534 (Pr-RT)<br>OK011533 (integrase) |
| MD497796 | Madrid | Madrid | 2017 | 2017 | M | MSM | Equatorial Guinea | OK011534 (NFLG) |
| MD745908 | Madrid | Madrid | 2019 | 2019 | M | MSM | Spain | OK011536 (Pr-RT) |
| MD821027 | Madrid | Madrid | 2018 | 2018 | M | MSM | Paraguay | OK011537 (Pr-RT) |
| MD844680 | Madrid | Madrid | 2020 | 2020 | M | MSM | Paraguay | OK011538 (Pr-RT) |
| MD947238 | Madrid | Madrid | 2018 | 2016 | M | MSM | Paraguay | OK011539 (Pr-RT) |
| PV003808 | Bilbao | Basque Country | 2020 | 2020 | M | MSM | Paraguay | OK011541 (Pr-RT)<br>OK011540 (integrase) |
| PV041472 | Bilbao | Basque Country | 2014 | 2014 | F | HT | Paraguay | OK011542 (Pr-RT) |
| PV106451 | Bilbao | Basque Country | 2010 | 2010 | F | HT | Paraguay | OK011543 (NFLG) |
| PV208413_2 | Bilbao | Basque Country | 2009 | 2009 | M | HT | Paraguay | OK011544 (Pr-RT) |
| PV320606 | Bilbao | Basque Country | 2014 | 2014 | M | HT | Paraguay | OK011545 (Pr-RT) |
| PV571209 | Bilbao | Basque Country | 2013 | 2013 | M | HT | Paraguay | OK011546 (Pr-RT) |
| PV623909 | Bilbao | Basque Country | 2011 | 2011 | F | HT | Paraguay | OK011547 (NFLG) |
| PV820832 | Bilbao | Basque Country | 2008 | 2008 | M | Sexual | Paraguay | OK011548 (Pr-RT) |
| PV844167 | Vitoria | Basque Country | 2016 | 2016 | M | HT | Paraguay | OK011549 (NFLG) |
| PV860355 | Bilbao | Basque Country | 2011 | 2011 | M | HT | Paraguay | OK011550 (Pr-RT) |
| PV867970 | Bilbao | Basque Country | 2020 | 2020 | M | MSM | Spain | OK011552 (Pr-RT)<br>OK011551 (integrase) |

**HT: heterosexual; MSM: man who has sex with men.*

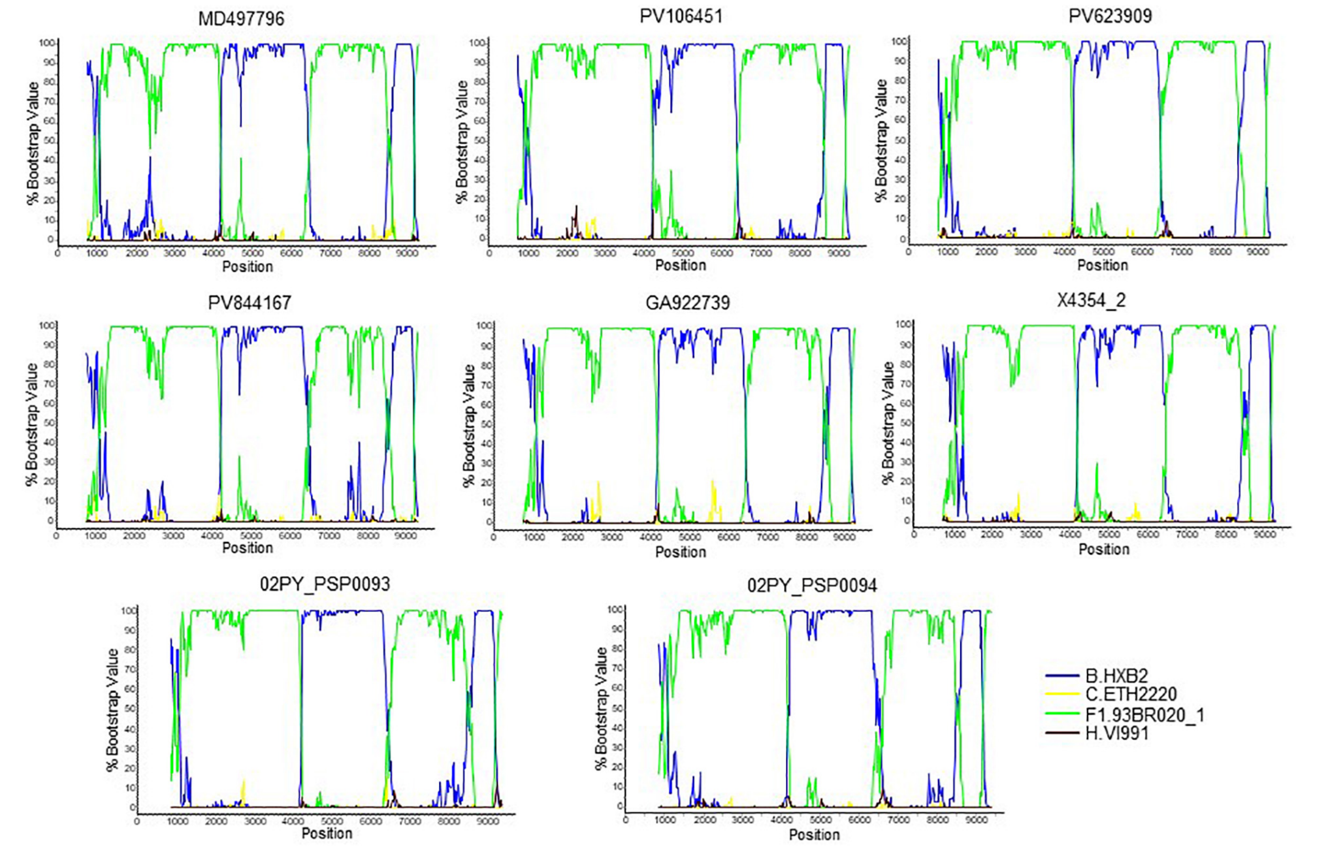


**FIGURE 2 |** Bootscan analyses of 6 NFLG sequences of viruses of the BF9 cluster obtained by us and of two BF1 database NFLG sequences from Paraguay, 02PY_PSP0093 and 02PY_PSP0094. The horizontal axis represents the position in the HXB2 genome of the midpoint of a 400 nt window moving in 20 nt increments and the vertical axis represents bootstrap values supporting clustering with subtype reference sequences.

## Prevalence of CRF66_BF Among HIV-1-Infected Paraguayans Residing in Spain and Among Sequences From Samples Collected in Paraguay Deposited at the HIV-1 Sequence Database

Among 67 HIV-1-infected Paraguayans residing in Spain studied by us, CRF66_BF was the most common non-subtype B genetic form, representing 20.9% (14 of 67) of total infections, 48.3% (14 of 29) of non-subtype B infections, and 60.1% (14 of 23) of F1/BF1 infections.

At the Los Alamos HIV Sequence Database, there are HIV-1 sequences from samples collected in Paraguay from only 23 individuals, of which 12 are NFLG sequences from samples collected in 2002 or 2003 and 11 are env V3 region sequences from samples with unknown collection years. In phylogenetic analyses, 2 of 12 NFLG and 1 of 11 V3 sequences [combined, 3 (13%) of 23 viruses] branched in the clade formed by CRF66_BF viruses (**Supplementary Figure 2**). However, due to the short length of the Paraguayan V3 sequences, the reliability of the tree of this segment for identifying CRF66_BF viruses is uncertain, since several CRF_BF and the subtype G references branched apart from other references of the same genetic form, and one CRF72_BF reference branched within the CRF66_BF clade.

## Temporal and Geographical Estimations of CRF66_BF Origin

To estimate the time and place of origin of CRF66_BF, Pr-RT sequences where analyzed with a Bayesian coalescent method with BEAST 1.10.4. Prior to this analysis, TempEst analysis determined that there was an adequate temporal signal in the dataset ($r^2$ = 0.5871). In the BEAST analysis, for the sequences corresponding to South American individuals residing in Spain, the assigned location trait was their country of origin, rather than their place of residence. This was done because most individuals harboring CRF66_BF identified in Spain were of South American origin (mostly from Paraguay) and because we found no definitive evidence of the local circulation of CRF66_BF in Spain, as reflected in clusters mainly comprising Spanish individuals. Therefore, we assumed that South Americans harboring CRF66_BF viruses had acquired HIV-1 in their countries of origin. In this analysis, the substitution rate was estimated at 1.987 x $10^{-3}$ subs/site/year (95% HPD, 8.885 x $10^{-4}$ – 3.282 x $10^{-3}$ subs/site/year) and the time of the MRCA of CRF66_BF was estimated around

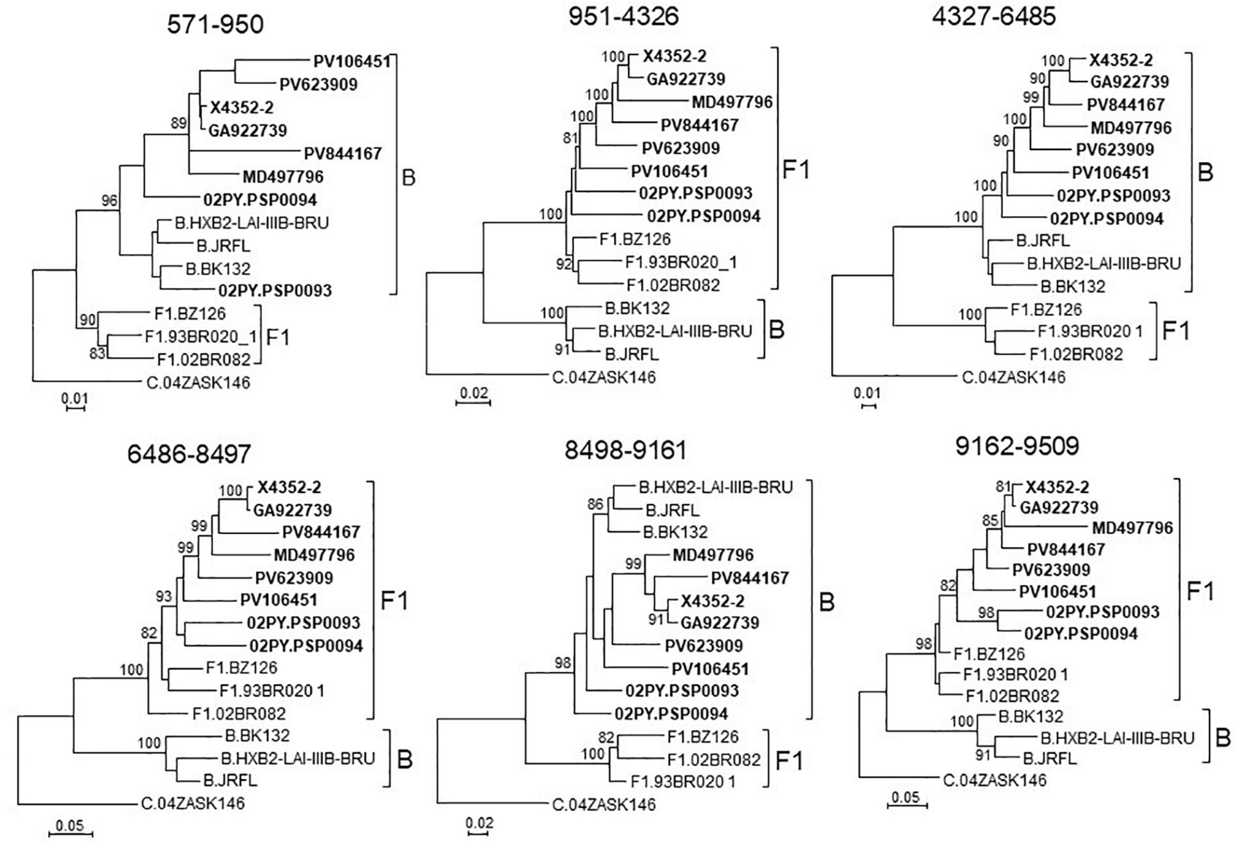


**FIGURE 3 |** Phylogenetic trees of interbreakpoint genome segments of the BF recombinant viruses analyzed by bootscanning, constructed with IQ-Tree. HXB2 positions delimiting the analyzed segments are indicated on top of the trees. Sequence names of BF viruses are in bold type. Names of subtype references are preceded by the corresponding subtype name. Only ultrafast bootstrap values ≥80% are shown.

1984 (95% HPD, 1970–1996), with its most probable location in Paraguay (PP = 0.76), with Argentina in second place (PP = 0.22) (**Figure 6**). Considering the possibility that local subclusters each found in one city could represent local transmissions, we performed a second analysis in which we assigned the country of location of the most recent diagnoses of such clusters to Spain, irrespective of the countries of origin of the individuals. In this analysis, Paraguay was also estimated as the most probable location of the MRCA of CRF66_BF, although with a lower support (PP = 0.55), and the support for Argentina increased to a PP = 0.42 (**Supplementary Figure 3**).

## DISCUSSION

The results of this study allow to define a new HIV-1 CRF, designated CRF66_BF, which is the 18$^{th}$ reported CRF derived from subtypes B and F. Samples harboring CRF66_BF were collected in 5 countries, in South America (Argentina, Paraguay, and Brazil) and Western Europe (Spain and Italy), with a majority collected in Spain. However, of samples collected in Spain, a great majority were from Paraguayan individuals. Bayesian coalescent analyses (performed with the assumption that South American individuals living in Spain harboring CRF66_BF viruses had acquired them in their countries of origin), pointed to a most probable origin of CRF66_BF in Paraguay (PP = 0.76), with Argentina being the second most probable location (PP = 0.22). When the analysis was performed assigning the most recently diagnosed samples of clusters found in a single Spanish city to Spain as the location trait, irrespective of the country of origin of the individual, the PPs for a MRCA in Paraguay or Argentina were not very different (0.55 vs. 0.42, respectively). Therefore, the results point to a South American origin of CRF66_BF, either in Paraguay or Argentina, without a definitive support for either country. However, given the great predominance of Paraguayans among CRF66_BF-infected individuals living in Spain, we cannot rule out that the same could happen in Argentina, where Paraguayans represent the largest immigrant national group (Instituto Nacional de Estadísticas y Censos, República Argentina, 2021). If this was the case, and information on country of origin of the sampled individuals living in Argentina was included in the analyses, it is possible that the support for a root of the CRF66_BF tree in Paraguay would increase.

The estimated date of origin of CRF66_BF around 1984 is consistent with the published estimated origin of the Brazilian F1 strain (around 1977) (Bello et al., 2007) and is similar to those

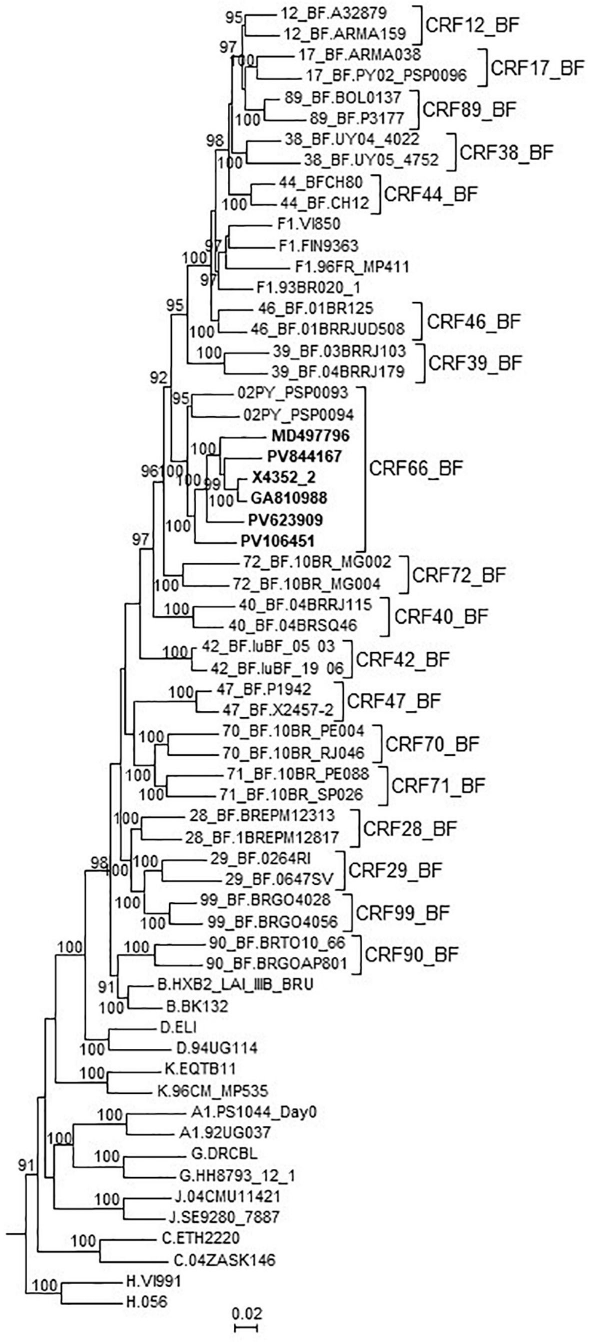


**FIGURE 4 |** Maximum likelihood tree of NFLG sequences of viruses of the BF9 cluster and PY02_PSP0094, constructed with IQ-Tree. References of all published CRF_BFs and of HIV-1 subtypes are also included in the analysis. The tree is rooted with SIVcpz virus MB66. Names of sequences obtained by us are in bold type. In reference sequences, the subtype or CRF is indicated before the virus name. Only ultrafast bootstrap values ≥90% are shown.

of other South American CRF_BFs (CRF12, CRF28/29, CRF38, CRF89, and CRF90) reported in the literature (Bello et al., 2010; Ristic et al., 2011; Reis et al., 2017; Delgado et al., 2021), although younger than some other estimates for CRF12_BF in the 1970s (Dilernia et al., 2011; Delgado et al., 2021) and older than the estimate for CRF99_BF, around 1993 (Reis et al., 2017).

Among HIV-1-infected Paraguayans residing in Spain studied by us, there was relatively high prevalence (21%) of CRF66_BF infections, which suggests that CRF66_BF could be one of the major HIV-1 genetic forms circulating in Paraguay. A better knowledge of the current prevalence of CRF66_BF in Paraguay would require sequencing a representative sample of recent HIV-1 diagnoses in the country. However, HIV-1 sequences from only 23 patients sampled in Paraguay are available at the Los Alamos HIV Sequence database, and the most recent molecular epidemiological study published to date involves the analysis of sequences from 55 samples collected 18 to 19 years ago (Aguayo et al., 2008), which are not available in public databases.

Notably, CRF66_BF, unlike all other non-Brazilian CRF_BFs identified to date in South America (CRF12_BF, CRF17_BF, CRF38_BF, CRF44_BF, and CRF89_BF, all circulating in the Southern Cone or neighboring countries), is unrelated to CRF12_BF, as deduced from the lack of breakpoint coincidence and of phylogenetic clustering with CRF12_BF. This implies that CRF66_BF originated independently from viruses of the CRF12_BF family, with a presumable ancestry in Brazil, where B and F1 viruses are circulating (Louwagie et al., 1994).

CRF66_BF is the 5th CRF of South American ancestry originally identified in Western Europe [after CRF42_BF (Struck et al., 2015), CRF47_BF (Fernández-García et al., 2010), CRF60_BC (Simonetti et al., 2014), and CRF89_BF (Delgado et al., 2021)], which, together with the reported propagation of HIV-1 variants of South America origin among the European population (de Oliveira et al., 2010; Collaço Verás et al., 2012; Thomson et al., 2012; Lai et al., 2014; Carvalho et al., 2015; Delgado et al., 2015; Fabeni et al., 2015, 2020; Vinken et al., 2019), points to an increasing relationship between the HIV-1 epidemics in both continents. This reflects migratory fluxes, most notably in Spain, where around 2.5 million South Americans live, representing nearly 40% of the migrant population (Instituto Nacional de Estadística, 2021a), and immigration from South America has increased greatly in recent years (Instituto Nacional de Estadística, 2021b) (**Supplementary Figure 4**). Considering the large and increasing South American immigrant population in Europe and the relative scarcity of HIV-1 sequences available in public databases from some South American countries (such as Colombia, Guyana, Ecuador, Bolivia, Paraguay, Chile, and Uruguay) (**Supplementary Figure 5**), studies on HIV-1 genetic diversity and molecular epidemiology among South American immigrants living in Europe could provide novel insights into the HIV-1 epidemics in their countries of origin, although the acquisition of some HIV-1 infections in their European countries of residence, reflected in branching in European clusters, should be taken into account (Osorno et al., 11th Conference on HIV Science, abstract PEC252, 18-21 July 2021[2]), as well as on the diffusion of South American HIV-1 variants in Europe.

It is interesting to note that although transmission of CRF66_BF is predominantly via heterosexual contact, most individuals in a cluster are MSM (**Figure 1**), which suggests

[2] https://ias2021.org/wp-content/uploads/2021/07/IAS2021_Abstracts_web.pdf

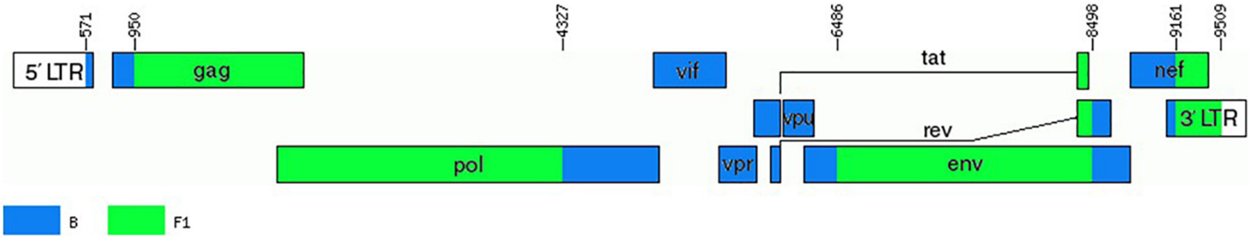


**FIGURE 5 |** Mosaic structure of CRF66_BF. Breakpoint positions are numbered as in the HXB2 genome. The drawing was made using the Recombinant HIV-1 Drawing Tool https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html.

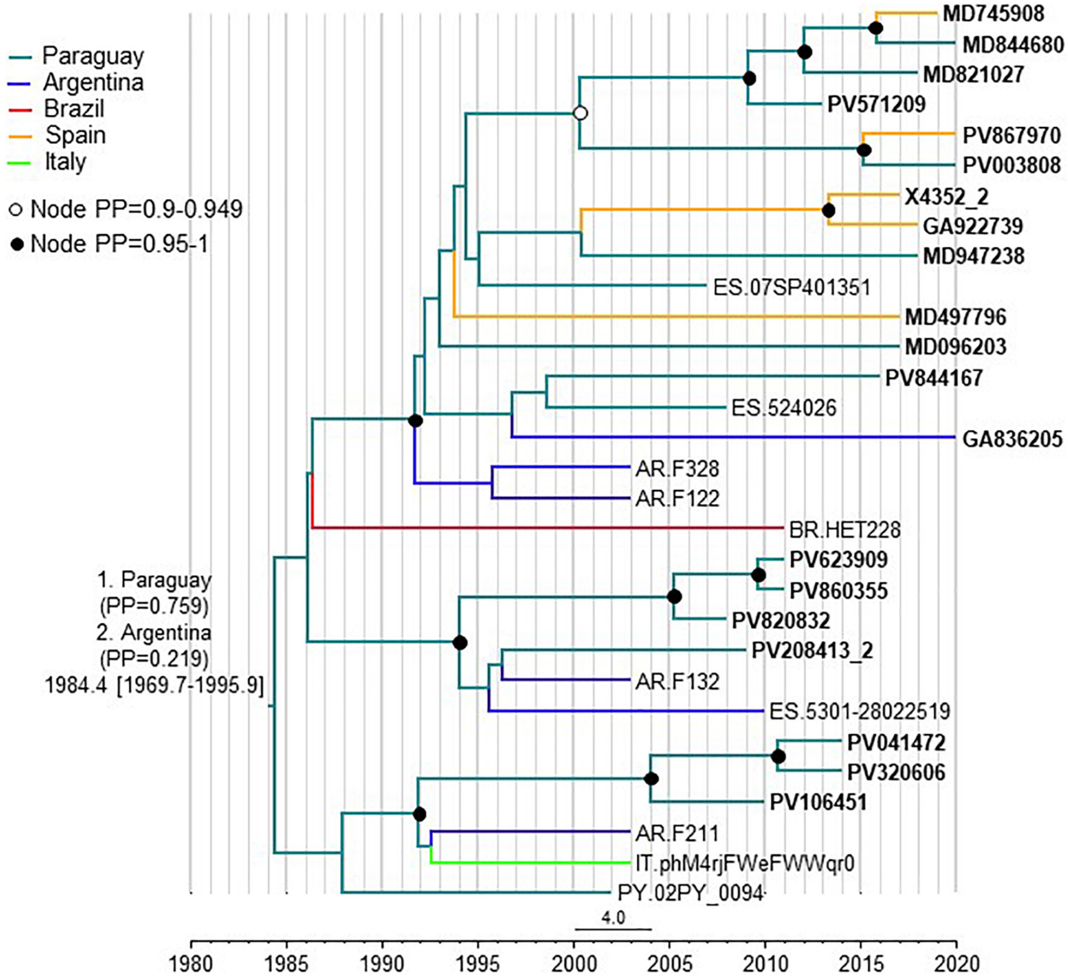


**FIGURE 6 |** Maximum clade credibility tree of CRF66_BF Pr-RT sequences. Branch colors indicate, for terminal branches, country of sample collection or, for South American individuals residing in Spain, of origin of the individual, which was used as location trait (see Methods), and for internal branches, the most probable location country of the subtending node, according to the legend on the upper left. For database sequence 524026, from a sample collected in Spain, location was assigned to Paraguay as the most probably country of origin, although the only available information in the GenBank entry is that the individual was from Latin America, because 15 (88.2%) of 17 Latin Americans with CRF66_BF sampled in Spain were from Paraguay. Nodes supported by PP ≥ 0.95 and PP 0.9–0.949 are indicated with filled and unfilled circles, respectively. The two most probable countries at the root of the tree are indicated, together with the PPs supporting each location and the time of the MRCA (mean value, with 95% HPD interval in brackets).

diffusion from a heterosexual to a MSM network. HIV-1 propagation between heterosexual and MSM networks has also been reported for CRF89_BF (Delgado et al., 2021) and for a large CRF02_AG cluster in Spain (Delgado et al., 2019), although in the latter case the direction of propagation was from MSM to heterosexuals.

One of the essential tasks of Biology is naming and classifying organisms. In this work, we have accomplished this task by identifying a new HIV-1 circulating recombinant form, derived from subtypes B and F1, named CRF66_BF. CRF66_BF most likely originated in South America, either in Paraguay or Argentina, and, unlike all non-Brazilian South American CRFs identified to date, is unrelated to CRF12_BF. The identification and genetic characterization of HIV-1 variants is the first and necessary step for molecular epidemiological studies examining their geographic dissemination, growth dynamics, and epidemiological associations, as well as for analyzing their biological properties, such as pathogenic and transmissibility potentials, response to antiretroviral therapies, and susceptibility to immune responses inducible by vaccines. Such studies on CRF66_BF and other South American CRFs may be the subject of future work.

## DATA AVAILABILITY STATEMENT

The names of the repository and accession numbers can be found in the article (Materials and Methods and **Table 1**).

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Committe of Research Ethics, Instituto de Salud Carlos III, Madrid, Spain. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

MT and ED conceived the study and supervised the experimental work. JB, MT, and ED processed sequences and performed phylogenetic analyses. MT performed phylodynamic analyses. HG performed data curation. JB, SB, MM-L, VM, MS, EG-B, and JEC performed experimental work. MN-T, JM, MZ-S, EU, JR, CR, IR-A, LE-O, JP, JG-C, AO, and JJC obtained samples and epidemiological data from patients. MT, ED, and HG wrote the manuscript. All authors read and approved the manuscript.

## FUNDING

This work was funded through Acción Estratégica en Salud Intramural (AESI), Instituto de Salud Carlos III, projects PI16CIII/00033 and PI19CIII/00042; Red de Investigación en SIDA (RIS), Instituto de Salud Carlos III, Subdirección General de Evaluación y Fondo Europeo de Desarrollo Regional (FEDER), Plan Nacional I+D+I, project RD16ISCIII/0002/0004; and scientific agreements with Consellería de Sanidade, Government of Galicia (MVI 1004/16) and Osakidetza-Servicio Vasco de Salud, Government of Basque Country (MVI 1001/16).

## ACKNOWLEDGMENTS

We thank José Antonio Taboada, from Consellería de Sanidade, Xunta de Galicia, and Daniel Zulaika, from Osakidetza-Servicio Vasco de Salud, for their support of this study, and the personnel at the Genomic Unit, Instituto de Salud Carlos III, for technical assistance in sequencing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.774386/full#supplementary-material

## REFERENCES

Aguayo, N., Laguna-Torres, V. A., Villafane, M., Barboza, A., Sosa, L., Chauca, G., et al. (2008). Epidemiological and molecular characteristics of HIV-1 infection among female commercial sex workers, men who have sex with men and people living with AIDS in Paraguay. *Rev. Soc. Bras. Med. Trop.* 41, 225–231. doi: 10.1590/s0037-86822008000300001

Aulicino, P. C., Gómez-Carrillo, M., Bello, G., Rocco, C., Mangano, A., Carr, J., et al. (2012). Characterization of full-length HIV-1 CRF17_BF genomes and comparison to the prototype CRF12_BF strains. *Infect. Genet. Evol.* 12, 443–447.

Bello, G., Aulicino, P. C., Ruchansky, D., Guimarães, M. L., López-Galíndez, C., Casado, C., et al. (2010). Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. *Retrovirology* 7:22. doi: 10.1186/1742-4690-7-22

Bello, G., Eyer-Silva, W. A., Couto-Fernandez, J. C., Guimarães, M. L., Chequer-Fernandez, S. L., Teixeira, S. L., et al. (2007). Demographic history of HIV-1 subtypes B and F in Brazil. *Infect. Genet. Evol.* 7, 263–270.

Cañada, J. E., Delgado, E., Gil, H., Sánchez, M., Benito, S., García-Bodas, E., et al. (2021). Identification of a new HIV-1 BC intersubtype circulating recombinant form (CRF108_BC) in Spain. *Viruses* 13:93. doi: 10.3390/v13010093

Carr, J. K., Avila, M., Gomez Carrillo, M., Salomon, H., Hierholzer, J., Watanaveeradej, V., et al. (2001). Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America. *AIDS* 15, F41–F47. doi: 10.1097/00002030-200110190-00002

Carvalho, A., Costa, P., Triunfante, V., Branca, F., Rodrigues, F., Santos, C. L., et al. (2015). Analysis of a local HIV-1 epidemic in Portugal highlights established transmission of non-B and non-G subtypes. *J. Clin. Microbiol.* 53, 1506–1514. doi: 10.1128/JCM.03611-14

Collaço Verás, N. M., Gray, R. R., de Macedo Brígido, L. F., Rodrigues, R., and Salemi, M. (2012). High-resolution phylogenetics and phylogeography of human immunodeficiency virus type 1 subtype C epidemic in South America. *J. Gen. Virol.* 92, 1698–1709. doi: 10.1099/vir.0.028951-0

de Oliveira, T., Pillay, D., and Gifford, R. J. (2010). The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. *PLoS One* 5:e9311. doi: 10.1371/journal.pone.0009311

De Sá Filho, D. J., Sucupira, M. C., Caseiro, M. M., Sabino, E. C., Diaz, R. S., and Janini, L. M. (2006). Identification of two HIV type 1 circulating recombinant forms in Brazil. *AIDS Res. Hum. Retroviruses* 22, 1–13. doi: 10.1089/aid.2006.22.1

Delgado, E., Benito, S., Montero, V., Cuevas, M. T., Fernández-García, A., Sánchez-Martínez, M., et al. (2019). Diverse large HIV-1 non-subtype B clusters are spreading among men who have sex with men in Spain. *Front. Microbiol.* 10:655. doi: 10.3389/fmicb.2019.00655

Delgado, E., Cuevas, M. T., Domínguez, F., Vega, Y., Cabello, M., Fernández-García, A., et al. (2015). Phylogeny and phylogeography of a recent HIV-1 subtype F outbreak among men who have sex with men in Spain deriving from a cluster with a wide geographic circulation in Western Europe. *PLoS One* 10:e0143325. doi: 10.1371/journal.pone.0143325

Delgado, E., Fernández-García, A., Pérez-Losada, M., Moreno-Lorenzo, M., Fernández-Miranda, I., Benito, S., et al. (2021). Identification of CRF89_BF, a new member of an HIV-1 circulating BF intersubtype recombinant form family widely spread in South America. *Sci. Rep.* 11:11442. doi: 10.1038/s41598-021-90023-x

Delgado, E., Ríos, M., Fernández, J., Pérez-Alvarez, L., Nájera, R., and Thomson, M. M. (2010). Identification of a new HIV type 1 BF intersubtype circulating recombinant form (CRF44_BF) in Chile. *AIDS Res. Hum. Retroviruses* 26, 821–826. doi: 10.1089/aid.2010.0006

Delgado, E., Thomson, M. M., Villahermosa, M. L., Sierra, M., Ocampo, A., Miralles, C., et al. (2002). Identification of a newly characterized HIV-1 BG intersubtype circulating recombinant form in Galicia, Spain, which exhibits a pseudotype-like virion structure. *J. Acquir. Immune Defic. Syndr.* 29, 536–543. doi: 10.1097/00126334-200204150-00016

Dilernia, D. A., Jones, L. R., Pando, M. A., Rabinovich, R. D., Damilano, G. D., Turk, G., et al. (2011). Analysis of HIV type 1 BF recombinant sequences from South America dates the origin of CRF12_BF to a recombination event in the 1970s. *AIDS Res. Hum. Retroviruses* 27, 569–578. doi: 10.1089/AID.2010.0118

Fabeni, L., Alteri, C., Orchi, N., Gori, C., Bertoli, A., Forbici, F., et al. (2015). Recent transmission clustering of HIV-1 C and CRF17_BF strains characterized by NNRTI-related mutations among newly diagnosed men in Central Italy. *PLoS One* 10:e0135325. doi: 10.1371/journal.pone.0135325

Fabeni, L., Santoro, M. M., Lorenzini, P., Rusconi, S., Gianotti, N., Costantini, A., et al. (2020). Evaluation of HIV transmission clusters among natives and foreigners living in Italy. *Viruses* 12:791. doi: 10.3390/v12080791

Fernández-García, A., Pérez-Alvarez, L., Cuevas, M. T., Delgado, E., Muñoz-Nieto, M., Cilla, G., et al. (2010). Identification of a new HIV type 1 circulating BF intersubtype recombinant form (CRF47_BF) in Spain. *AIDS Res. Hum. Retroviruses* 26, 827–832. doi: 10.1089/aid.2009.0311

Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B., and Suchard, M. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* 30, 713–724. doi: 10.1093/molbev/mss265

Guimarães, M. L., Eyer-Silva, W. A., Couto-Fernandez, J. C., and Morgado, M. G. (2008). Identification of two new CRF_BF in Rio de Janeiro State, Brazil. *AIDS* 22, 433–435. doi: 10.1097/QAD.0b013e3282f47ad0

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020). Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781. doi: 10.1016/S1473-3099(18)30647-9

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Instituto Nacional de Estadística (2021a). *Principales Series De Población Desde 1998*. Available online at: https://www.ine.es/jaxi/Datos.htm?path=/t20/e245/p08/l0/, andfile=01006.px#!tabs-tabla (Accessed August 17, 2021).

Instituto Nacional de Estadística (2021b). *Migraciones Exteriores*. Available online at: https://www.ine.es/jaxiT3/Tabla.htm?t=24295 (Accessed August 17, 2021).

Instituto Nacional de Estadísticas y Censos, República Argentina (2021). *Censo 2010*. Available online at: https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-135 (Accessed August 17, 2021).

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lai, A., Bozzi, G., Franzetti, M., Binda, F., Simonetti, F. R., Micheli, V., et al. (2014). Phylogenetic analysis provides evidence of interactions between Italian heterosexual and South American homosexual males as the main source of national HIV-1 subtype C epidemics. *J. Med. Virol.* 86, 729–736. doi: 10.1002/jmv.23891

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/JVI.73.1.152-160.1999

Los Alamos National Laboratory (2021). *HIV Sequence Database*. Available online at: https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html (Accessed August 17, 2021).

Louwagie, J., Delwart, E. L., Mullins, J. I., McCutchan, F. E., Eddy, G., and Burke, D. S. (1994). Genetic analysis of HIV-1 isolates from Brazil reveals presence of two distinct genetic subtypes. *AIDS Res. Hum. Retroviruses* 10, 561–567. doi: 10.1089/aid.1994.10.561

Nájera, R., Delgado, E., Pérez-Álvarez, L., and Thomson, M. M. (2002). Genetic recombination and its role in the development of the HIV-1 pandemic. *AIDS* 16, S3–S16. doi: 10.1097/00002030-200216004-00002

Pessôa, R., Watanabe, J. T., Calabria, P., Felix, A. C., Loureiro, P., Sabino, E. C., et al. (2014). Deep sequencing of HIV-1 near full-length proviral genomes identifies high rates of BF1 recombinants including two novel circulating recombinant forms (CRF) 70_BF1 and a disseminating 71_BF1 among blood donors in Pernambuco, Brazil. *PLoS One* 9:e112674. doi: 10.1371/journal.pone.0112674

Pineda-Peña, A. C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., et al. (2013). Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* 19, 337–348. doi: 10.1016/j.meegid.2013.04.032

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rambaut, A., Lam, T. T., Max, C. L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

Reis, M. N. G., Bello, G., Guimarães, M. L., and Stefani, M. M. A. (2017). Characterization of HIV-1 CRF90_BF1 and putative novel CRFs_BF1 in Central West, North and Northeast Brazilian regions. *PLoS One* 12:e0178578. doi: 10.1371/journal.pone.0178578

Reis, M. N. G., Guimarães, M. L., Bello, G., and Stefani, M. M. A. (2019). Identification of new HIV-1 circulating recombinant forms CRF81_cpx and CRF99_BF1 in Central Western Brazil and of Unique BF1 recombinant forms. *Front. Microbiol.* 10:97. doi: 10.3389/fmicb.2019.00097

Ristic, N., Zukurov, J., Alkmim, W., Diaz, R. S., Janini, L. M., and Chin, M. P. (2011). Analysis of the origin and evolutionary history of HIV-1 CRF28_BF and CRF29_BF reveals a decreasing prevalence in the AIDS epidemic of Brazil. *PLoS One* 6:e17485. doi: 10.1371/journal.pone.0017485

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Ruchansky, D., Casado, C., Russi, J. C., Arbiza, J. R., and López-Galíndez, C. (2009). Identification of a new HIV type 1 circulating recombinant form (CRF38_BF1) in Uruguay. *AIDS Res. Hum. Retroviruses* 25, 351–356. doi: 10.1089/aid.2008.0248

Salminen, M. O., Carr, J. K., Burke, D. S., and McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res. Hum. Retroviruses* 11, 1423–1425. doi: 10.1089/aid.1995.11.1423

Sanabani, S. S., Pastena, E. R., Neto, W. K., Martinez, V. P., and Sabino, E. C. (2010). Characterization and frequency of a newly identified HIV-1 BF1 intersubtype circulating recombinant form in São Paulo Brazil. *Virol. J.* 7:74. doi: 10.1186/1743-422X-7-74

Schultz, A. K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., et al. (2009). jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res.* 37, W647–W651. doi: 10.1093/nar/gkp371

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* 23, 7–9. doi: 10.1093/molbev/msj021

Sierra, M., Thomson, M. M., Ríos, M., Casado, G., Ojea de Castro, R., Delgado, E., et al. (2005). The analysis of near full-length genome sequences of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Chile, Venezuela and Spain reveals their relationship to diverse lineages of recombinant viruses related to CRF12_BF. *Infect. Genet. Evol.* 5, 209–217. doi: 10.1016/j.meegid.2004.07.010

Simonetti, F. R., Lai, A., Monno, L., Binda, F., Brindicci, G., Punzi, G., et al. (2014). Identification of a new HIV-1 BC circulating recombinant form (CRF60_BC) in Italian young men having sex with men. *Infect. Genet. Evol.* 23, 176–181. doi: 10.1016/j.meegid.2014.02.007

Struck, D., Roman, F., De Landtsheer, S., Servais, J. Y., Lambert, C., Masquelier, C., et al. (2015). Near full-length characterization and population dynamics of the human immunodeficiency virus type 1 circulating recombinant form 42 (CRF42_BF) in Luxembourg. *AIDS Res. Hum. Retroviruses* 31, 554–558. doi: 10.1089/AID.2014.0364

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

Thomson, M. M., and Nájera, R. (2005). Molecular epidemiology of HIV-1 variants in the global AIDS pandemic: an update. *AIDS Rev.* 7, 210–224.

Thomson, M. M., Fernández-García, A., Delgado, E., Vega, Y., Díez-Fuertes, F., Sánchez-Martínez, M., et al. (2012). Rapid expansion of a HIV-1 subtype F cluster of recent origin among men who have sex with men in Galicia, Spain. *J. Acquir. Immune Defic. Syndr.* 59, e49–e51. doi: 10.1097/QAI.0b013e3182400fc4

Thomson, M. M., Herrero, I., Villahermosa, M. L., Vázquez de Parga, E., Cuevas, M. T., Carmona, R., et al. (2002). Diversity of mosaic structures and common ancestry of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Argentina revealed by analysis of near full-length genome sequences. *J. Gen. Virol.* 83, 107–119. doi: 10.1099/0022-1317-83-1-107

Thomson, M. M., Villahermosa, M. L., Vázquez de Parga, E., Cuevas, M. T., Delgado, E., Manjón, N., et al. (2000). Widespread circulation of a B/F intersubtype recombinant form among HIV-1-infected individuals in Buenos Aires, Argentina. *AIDS* 14, 897–899. doi: 10.1097/00002030-200005050-00020

Trifinopoulos, J., Nguyen, L. T., von Haeseler, A., and Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44, W232–W235. doi: 10.1093/nar/gkw256

Vinken, L., Fransen, K., Cuypers, L., Alexiev, I., Balotta, C., Debaisieux, L., et al. (2019). Earlier initiation of antiretroviral treatment coincides with an initial control of the HIV-1 sub-subtype F1 outbreak among men-having-sex-with-men in Flanders, Belgium. *Front. Microbiol.* 10:613. doi: 10.3389/fmicb.2019.00613

Zhang, M., Foley, B., Schultz, A. K., Macke, J. P., Bulla, I., Stanke, M., et al. (2010). The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 November 2021
doi: 10.3389/fmicb.2021.779472

# The Evolution of Regulatory Elements in the Emerging Promoter-Variant Strains of HIV-1 Subtype C

*Disha Bhange[1], Nityanand Prasad[1], Swati Singh[1], Harshit Kumar Prajapati[1], Shesh Prakash Maurya[2], Bindu Parachalil Gopalan[3], Sowmya Nadig[3], Devidas Chaturbhuj[4], Boobalan Jayaseelan[5], Thongadi Ramesh Dinesha[5], Syed Fazil Ahamed[3], Navneet Singh[1], Anangi Brahmaiah[1], Kavita Mehta[1], Yuvrajsinh Gohil[1], Pachamuthu Balakrishnan[6], Bimal Kumar Das[2], Mary Dias[3], Raman Gangakhedkar[7], Sanjay Mehendale[8], Ramesh S Paranjape[7], Shanmugam Saravanan[5], Anita Shet[3], Sunil Suhas Solomon[9,10], Madhuri Thakar[4] and Udaykumar Ranga[1]**

*[1] HIV-AIDS Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Bengaluru, India, [2] HIV Immunology Laboratory, Department of Microbiology, All India Institute of Medical Sciences (AIIMS), New Delhi, India, [3] Division of Microbiology/Infectious Diseases Unit, St. John's National Academy of Health Sciences, Bengaluru, India, [4] Department of Serology and Immunology, National AIDS Research Institute (NARI), Pune, India, [5] Department of Molecular Biology and Genotyping, Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), Chennai, India, [6] Infectious Diseases Laboratory, Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), Chennai, India, [7] Department of Clinical Sciences, National AIDS Research Institute (NARI), Pune, India, [8] Department of Research, P. G. Hinduja National Hospital and Medical Research Centre, Mumbai, India, [9] YRGCARE Suniti Solomon Outpatient Clinic, Y. R. Gaitonde Center for AIDS Research and Education (YRG CARE), Chennai, India, [10] Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, United States*

OPEN ACCESS

**Edited by:**
*Kok Keng Tee,*
*University of Malaya, Malaysia*

**Reviewed by:**
*Hirotaka Ode,*
*Nagoya Medical Center, Japan*
*Antoinette Van Der Kuyl,*
*University of Amsterdam, Netherlands*

***Correspondence:**
*Udaykumar Ranga*
*udaykumar@jncasr.ac.in*

**Specialty section:**
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

**Received:** *18 September 2021*
**Accepted:** *19 October 2021*
**Published:** *16 November 2021*

**Citation:**
*Bhange D, Prasad N, Singh S, Prajapati HK, Maurya SP, Gopalan BP, Nadig S, Chaturbhuj D, Jayaseelan B, Dinesha TR, Ahamed SF, Singh N, Brahmaiah A, Mehta K, Gohil Y, Balakrishnan P, Das BK, Dias M, Gangakhedkar R, Mehendale S, Paranjape RS, Saravanan S, Shet A, Solomon SS, Thakar M and Ranga U (2021) The Evolution of Regulatory Elements in the Emerging Promoter-Variant Strains of HIV-1 Subtype C. Front. Microbiol. 12:779472. doi: 10.3389/fmicb.2021.779472*

In a multicentric, observational, investigator-blinded, and longitudinal clinical study of 764 ART-naïve subjects, we identified nine different promoter variant strains of HIV-1 subtype C (HIV-1C) emerging in the Indian population, with some of these variants being reported for the first time. Unlike several previous studies, our work here focuses on the evolving viral regulatory elements, not the coding sequences. The emerging viral strains contain additional copies of the existing transcription factor binding sites (TFBS), including TCF-1α/LEF-1, RBEIII, AP-1, and NF-κB, created by sequence duplication. The additional TFBS are genetically diverse and may blur the distinction between the modulatory region of the promoter and the viral enhancer. In a follow-up analysis, we found trends, but no significant associations between any specific variant promoter and prognostic markers, probably because the emerging viral strains might not have established mono infections yet. Illumina sequencing of four clinical samples containing a coinfection indicated the domination of one strain over the other and establishing a stable ratio with the second strain at the follow-up time points. Since a single promoter regulates viral gene expression and constitutes the master regulatory circuit with Tat, the acquisition of additional and variant copies of the TFBS may significantly impact viral latency and latent reservoir characteristics. Further studies are urgently warranted to understand how the diverse TFBS profiles of the viral promoter may modulate the characteristics of the latent reservoir, especially following the initiation of antiretroviral therapy.

**Keywords: HIV-1, subtype C, evolution, sequence duplication, latency**

## INTRODUCTION

Based on phylogenetic association, the viral strains of HIV-1 are classified into four groups (M, N, O, and P), and within group M, into 10 different genetic subtypes, A, B, C, D, F, G, H, J, K, L (Yamaguchi et al., 2020), and numerous recombinant forms. Of the various genetic subtypes of HIV-1 unevenly distributed globally, HIV-1C and its recombinant forms are responsible for nearly half of the global infections (Locateli et al., 2007; Tebit and Arts, 2011; Hemelaar et al., 2019). Despite the high prevalence of HIV-1C, only a limited number of studies are available examining the causes underlying the expansion of these viral strains and their impact on disease manifestation.

Although the basic architecture of HIV-1 LTR is broadly conserved among the diverse HIV-1 genetic families, subtype-associated differences are manifested (Ramírez de Arellano et al., 2006). The configuration of transcription factor binding sites (TFBS), including those of NF-κB, NF-AT, AP-1, and other regulatory elements such as the TATA box, and the TAR region, in HIV-1C LTR (C-LTR) differs from that of the other viral subtypes (Ramírez de Arellano et al., 2006). Of the TFBS variations, differences in the copy number and sequence of the NF-κB motif are unique to HIV-1C. C-LTR typically contains three or four NF-κB motifs in the enhancer region compared to only one motif present in HIV-1A/E or two motifs in all the other HIV-1 subtypes (Bachu et al., 2012a). Further, the additional copies of the NF-κB motif in C-LTR are genetically variable, alluding to a possibility that the viral promoters are receptive to a diverse and broader range of cellular signals. For instance, the four copies of the NF-κB motif in the enhancer of 4-κB viral strains represent three genetically distinct NF-κB binding sites (Bachu et al., 2012a). Apart from the NF-κB motif, other regulatory elements, including AP-1, RBEIII, and TCF-1α/LEF-1, also show subtype-associated variations, although the impact of such variations has not been examined.

Several publications reported the insertion or deletion of TFBS in HIV-1 LTR. One example is the sequence duplication of the TCF-1α/LEF-1, RBF-2, AP-1, and c-EBPα binding motif in the modulatory region of the LTR, technically called the most frequent naturally occurring length polymorphism (MFNLP; Koken et al., 1992; Ait-Khaled et al., 1995; Estable et al., 1996; Zhang et al., 1997). Approximately 38% of the HIV-1B viral isolates contain MFNLP, a phenomenon believed to be a compensatory mechanism to ensure the presence of at least one functional RBEIII site in the LTR (Estable et al., 1998). Although RBEIII duplication has been found in several subtypes, the significance of this phenomenon has been examined predominantly in HIV-1B. It, however, remains inconclusive whether the presence of RBEIII duplication is directly associated with reduced viral replication or slower disease progression (Estable, 2007). A small number of reports reported RBEIII duplication in HIV-1C infection, however, without evaluating its effect on the replication fitness of the viral strains and disease progression (Rodriguez et al., 2007).

Over the past several years, our laboratory has documented the emergence of LTR variant strains of HIV-1C in India and elsewhere (Siddappa et al., 2004; Bachu et al., 2012b; Boullosa et al., 2014). While the appearance of genetic diversity and such diversity impacting viral evolution are common to the various genetic subtypes of HIV-1, the genetic variation we describe in HIV-1C is non-sporadic and radically different in an important aspect. Viral evolution in HIV-1C appears to be directional toward modulating transcriptional strength of the promoter by creating additional copies of the existing TFBS, such as NF-κB, AP-1, RBEIII, and TCF-1α/LEF-1 motifs, by sequence duplication and co-duplication. A single viral promoter in HIV-1 regulates two diametrically opposite functions critical for viral survival – transcriptional activation and silencing. Hence, any variation in the constitution of the TFBS (copy number difference and/or genetic variation) may have a profound impact on viral replication fitness.

Here, in a multicentric, observational, non-interventional, investigator-blinded, and longitudinal clinical study, we examined the promoter sequences of 455 primary viral isolates derived from ART-naïve subjects. We show that the magnitude of TFBS variation is much larger than we reported previously. At least nine different TFBS variant viral strains have emerged in recent years. Using the Illumina MiSeq platform, we attempted to characterize the proviral DNA of a selected subset of viral variants containing the RBEIII motif duplication. The data allude to the possibility that some of the emerging strains could achieve greater replication fitness levels and may establish expanding epidemics in the future, which requires monitoring. This work provides important insights into the HIV-1 evolution taking place at the level of population in India.

## MATERIALS AND METHODS

### Study Participants and Samples

Participants were recruited at four different sites in India for primary screening (PS) and longitudinal study (LS) – All India Institute of Medical Sciences (AIIMS), New Delhi (PS = 107, LS = 73); National AIDS Research Institute (NARI), Pune (PS = 61, LS = 38); St. John's National Academy of Health Sciences, Bangalore (PS = 116, LS = 60); and Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), Chennai (PS = 171, LS = 37). Subjects above 18 years of age with documented evidence of serological positive test for HIV-1 were recruited to the study. From the hospital records, all the study participants were reportedly ART-naive at baseline. Further, the approximate date of initial infection and the duration of the infection are not available for most of the participants. The study subjects were likely exposed to the viral infection for several years before they reported to the clinics.

**Abbreviations:** TFBS, Transcription factor binding site; HIV-1C, HIV-1 subtype C; LTR, Long terminal repeats; C-LTR, HIV-1C LTR; NGS, Next-generation sequencing; RT-PCR, Reverse transcription polymerase chain reaction; ART, Antiretroviral therapy.

## Ethics Statement

Written informed consent was obtained from all study participants, following specific institutional review board-approved protocols. Ethical approval for the study was granted by the Institutional Review Board of each clinical site. All the clinical sites screened the potential subjects, counselled, recruited the study participants, and maintained the clinical cohorts for the present study. The Human Ethics and Biosafety Committee of Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Bangalore, reviewed the proposal and approved the study.

## Primary Screening: LTR Amplification and Molecular Typing of the Viral Promoter

For the molecular typing of the viral promoter, 15 ml of peripheral blood was collected from every participant at one time. Of this sample, 3–5 ml of blood was allocated to determine the CD4 cell count by flow cytometry and the extraction of genomic DNA from whole blood. From the rest of the blood sample, PBMC were isolated by density-gradient centrifugation. The PBMC and plasma samples were stored in 1 ml aliquots in a liquid nitrogen container or a deep freezer, respectively, for further clinical analysis.

Genomic DNA was extracted from 200 µl of the whole blood using a commercial DNA extraction kit (GenElute™ Blood Genomic DNA kit, Cat. No. NA2020, Sigma-Aldrich, United States) and was eluted in 200 µl volume. The extracted DNA samples from the clinical sites were shipped to JNCASR on cold packages. The LTR sequences were amplified using 200–300 ng of genomic DNA as a template from each of the clinical samples using Taq DNA Polymerase (Cat. No. M073L, New England Biolabs, MA, United States) in a Peqstar 2x thermal cycler (Peqlab, VWR). The U3 region of LTR was amplified using a nested-PCR strategy with the primers listed in **Supplementary Table S1**. Strict procedural and physical safeguards were implemented to minimize carryover contamination that included reagent preparation and PCR setup, amplification, and post-PCR processing of samples in separate rooms. The PCR products were purified using a commercial DNA purification kit (FavorPrep Gel/PCR Purification Kit, Cat. No. FAGCK001, Favorgen Biotech Corp. Ping-Tung 908, Taiwan). The amplified LTR sequences were analyzed using Sanger Dideoxy sequencing (Applied Biosystems, CA, United States). All the sequences were subjected to further quality control by multiple sequence alignment and phylogenetic analysis with the in-house laboratory sequence database using the ClustalW algorithm of BioEdit sequence alignment editor and MEGA6.0 software, respectively. Different viral strains were categorized by analyzing the LTR spanning modulatory and enhancer region from the TCF-1α/LEF-1 motif up to the Sp1III motif.

## Phylogenetic Analysis

The phylogenetic analysis of the HIV-1C LTR variants derived from 455 patient samples was performed with 31 reference sequences representing different HIV-1 subtypes. The analysis was performed with 1,000 bootstrap values. The evolutionary history was inferred by using the maximum likelihood method based on the Tamura-Nei model (Tamura and Nei, 1993). The tree with the highest log likelihood (−36587.28) is shown. Initial tree(s) for the heuristic search were obtained by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 486 nucleotide sequences. A total of 292 positions were included in the final data set. Evolutionary analysis was conducted in MEGA7.0 (Kumar et al., 2016).

## The Follow-Up Clinical Procedures

Following successful characterization of the viral promoter at JNCASR, the clinical sites were advised to recruit specific study subjects without disclosing the nature of the viral LTR. The clinical sites were, thus, blinded to the identity of the viral LTR. All the clinical procedures were performed at the clinical sites using the same protocols and kits as described below.

A total of 15 ml of peripheral blood were collected in a BD vacutainer (Cat. no. 367525, Becton Dickinson, CA, United States) from each participant at 0, 6, 12, 18, 24, and 36 months during 2015–19. The PBMC and plasma samples were stored in 1 ml aliquots in a liquid nitrogen container or a deep freezer, respectively. The CD4 T-cell count was determined using the BD Multitest commercial kit-CD3/CD8/CD45/CD4 (Cat. No. 340491, Becton Dickinson, CA, United States) following the manufacturer's instructions. The samples were analyzed using a BD FACSCalibur flow cytometer or any other suitable machine. Calibration of the flow cytometer was performed using BD CaliBRITE 3 and APC beads (Cat. no. 340486 and 340487, respectively, Becton Dickinson, CA, United States). The plasma viral RNA load was determined at 0 and 12-month time points using the Abbott m2000rt Real-Time PCR machine (Abbott Molecular Inc. Des Plaines, IL, United States). Levels of soluble CD14 (sCD14) in the plasma were determined using Human sCD14 Quantikine ELISA Kit (Cat No. DC140, R & D Systems, MN, United States). Analysis of sCD14 was performed at months 0 and 12.

## RNA Isolation and RT-PCR for the Next-Generation Sequencing

RNA was extracted from 1 ml of the stored plasma samples using a commercial viral RNA isolation kit (NucliSENS miniMAG nucleic acid extraction kit, Ref. No. 200293, BioMerieux, France). The complementary DNA (cDNA) was synthesized using HIV-specific primers (**Supplementary Table S1**) and a commercial kit (SuperScript™ IV First-Strand Synthesis System with ezDNase™ Enzyme (Cat. No. 18091150, Invitrogen, Carlsbad, CA, United States). The reaction vials were incubated at 65°C for 5 min, following 2-min incubation on ice and 50°C for 50 min. The reactions were terminated by incubating the samples at 85°C for 5 min, followed by RNaseH treatment. The cDNA was used for the amplification of LTR.

## The Next-Generation Sequencing

The PCR products containing the RBEIII motif duplication were subjected to the NGS analysis using the Miseq Illumina platform. Each sample was amplified in duplicates using primers containing a unique 8 bp barcode sequence specific for each sample. The amplification of the U3 region (~300–350 bp) using genomic DNA or cDNA prepared from plasma RNA was performed the same way as described above for the primary screening except that the primers contained a unique sequence barcode at the 5′-end as listed (**Supplementary Tables S1, S2**). The concentration of the purified PCR product was determined using the Qubit™ dsDNA BR assay kit (Cat. No. Q32850, Invitrogen, CA, United States). All the samples were pooled at an equal concentration and were processed further. We pulsed the LTR amplicon of Indie.C1, a reference HIV-1C molecular clone, as internal quality control for sequencing.

DNA was quantified using the QUBIT 3 Fluorometer and a dsDNA HS Dye. After adding an "A" nucleotide to the 3′ ends, the adenylated fragments were ligated with loop adapters and cleaved with the uracil-specific excision reagent (USER) enzyme. The DNA was further purified using AMPure beads and then enriched by PCR in 6 cycles using NEBNext Ultra II Q5 master mix (Cat. No. E7645L, New England Biolabs, Inc., MA, United States), Illumina universal primer, and sample-specific octamer primers. The amplified products were cleaned by using AMPure beads, and the final DNA library was eluted in 15 μl of 0.1X TE buffer. The volume of 1 μl of the library was used to quantify by QUBIT 3 Fluorometer using dS DNA HS reagent. The analysis of the fragment size was performed on Agilent 4150 Tape Station by loading 1 μl of the library to Agilent D1000 Screen Tape. The library was sequenced using the Illumina MiSeq system and MiSeq Reagent kit v3 (Cat. No. MS-102-3003, Illumina, San Diego, CA, United States) following the manufacturer's instructions.

Data analysis was performed using a custom pipeline as depicted (**Supplementary Figure S2**). First, the quality assessment was performed using FastQC (version 0.11.5), and the sequencing adapters were removed using Trimmomatic-master (version 0.33) from the raw paired-end data. Second, the paired reads from both the forward and reverse files were merged using the PEAR algorithm of PANDAseq (version 3.9.1) with the minimum and maximum read length set to be 200 and 500, respectively, with an overlap of at least 8 base pairs. Next, the merged reads were mapped to the LTR region of the two reference sequences: Indie.C1 (AB023804.1) and D24 (EF469243.2), using local alignment in Bowtie2 software (version 2.3.5.1). Using custom C++ and shell scripts, the mapped reads were demultiplexed in individual samples based on the combination of forward and reverse barcodes (**Supplementary Table S2**). The reads containing the C-κB motif sequence (HIV-1C) were considered for further analysis using a custom C++ script. All the HIV-1C reads were then grouped in multiple categories based on the number of motifs and sequence of NF-κB, RBEIII, and TCF-1α/LEF-1 motifs using a custom C++ script. The percentage of every category was then calculated based on the total number of reads within each sample, using the custom shell, C++, and R scripts. Next, DeconSeq (version 0.4.3) was used to analyze and filter out inter-sample sporadic contamination. A reference database was prepared manually for every major-variant category (category with >10% of total reads at all time points), by taking only the most abundant variant of that category from each time point of the sample. The reference database of a variant category was used to cross-check for contamination in the same variant category of the other samples, where the variant category is <10% in one or more time points. The percent coverage and identity thresholds were set to 90% each to allow a maximum of 5 mismatches or indels in a sequence stretch of 50 bases. The sum of percentages of all the variants classified as clean for a category, in all the time points, was done using custom shell and R scripts. From the total number of reads for each time point of a sample, the number of contaminated reads was eliminated, and the clean reads were taken ahead for calculating the percentage prevalence of single and double-RBEIII variants.

## Statistical Analysis

The data were analyzed using GraphPad prism 9, except the sequences used for phylogeny determination. *p* values of 0.05 or less were considered statistically significant. A non-parametric, Kruskal-Wallis (for multiple comparisons) test was applied to evaluate statistical significance in the case of a cross-sectional analysis of plasma viral load. One-way ANOVA was used to evaluate statistical significance for cross-sectional analysis of CD4 cell count and sCD14 levels. Two-way ANOVA was used to evaluate statistical significance in the case of longitudinal analysis of all the three parameters, plasma viral load, CD4 cell count, and sCD14.

## Data Availability Statement

All the sequences reported in this paper are available from the GenBank database under accession nos. MN840242.1–MN840356.1, MT847032–MT847207, and MT593868–MT594037. The raw data files for Illumina MiSeq are available under accession no. PRJNA720640.

# RESULTS

## The Magnitude of TFBS Variation in HIV-1C LTR

We collected 764 primary clinical samples from four different clinical sites in India, All India Institute of Medical Sciences (AIIMS), New Delhi; National AIDS Research Institute (NARI), Pune; St. John's National Academy of Health Sciences, Bangalore; and Y. R. Gaitonde Centre for AIDS Research and Education (YRG CARE), Chennai, between 2016 and 2019. Using the genomic DNA, we could successfully amplify the U3 region in the LTR of 518 of 764 viral samples, whereas the amplification of the rest of the samples failed. The sequences of the 518 amplified PCR products were determined by the Sanger sequencing method. Of the 518 sequences, the genetic typing of only 455 could be accomplished. The sequence analysis indicated a rapidly changing TFBS profile in HIV-1C (**Figure 1A**).

The proportion of viral variants across the four clinical sites was comparable without geographic skewing (**Table 1**). A large majority of the variant LTRs contain additional copies of the existing TFBS, including that of NF-κB, RBF-2, AP-1, and TCF-1α/LEF-1 binding sites (**Figures 1B, 2A**). Based on the TFBS profile, the sequences of the copied TFBS, and their temporal location, the various viral promoters may be classified into three categories.

Category-1 is represented by the viral promoters containing three NF-κB motifs without the duplication of any other TFBS. This group consists of the canonical LR-HHC-LTR and a new variant LR-FHC-LTR. The canonical LR-HHC-LTR represents the largest group among all HIV-1C promoters, comprising 302 of 455 sequences (66.4%). This LTR contains three tandemly arranged NF-κB sites in the enhancer, representing two distinct κB-motifs, two H-κB motifs (5′-GGGACTTTCC-3′), and one C-κB motif (5′-GGGGCGTTCC-3′, differences underlined). Immediately upstream of the viral enhancer, an RBF-2 binding site (R, RBEIII motif, 5′-ACTGCTGA-3′) and further upstream a TCF-1α/LEF-1 site (L, 5′-TACAAA/GG/A-3′) are located. The canonical HIV-1C promoter is identified here as LR-HHC-LTR to denote the 5′ to 3′ arrangement of the three kinds of TFBS, TCF-1α/LEF-1, RBEIII, and NF-κB. The second member of category-1 contains a variant LR-FHC-LTR comprising 40 of 455 (8.8%) sequences. The characteristic feature of these viral strains is the presence of three genetically distinct NF-κB motifs in the viral enhancer. Thus, the two viral strains of category-1 formed the major proportion of all the LTRs, 342/455 (75.2%; **Figure 1**). Multisequence alignments of several representative viral strains of the canonical LR-HHC (**Supplementary Figure S1A**) and the variant LR-FHC (**Supplementary Figure S1B**) are presented. The LR-FHC-LTR, being reported here for the first time, may have originated from the 4-κB viral strain FHHC of category 2, as described below.

Category-2 viral LTRs contain an additional (fourth) NF-κB binding site (F-κB site, 5′-GGGACTTTCT-3′) located downstream of the RBEIII site. The 4-κB LTRs, thus, contain one F-, two H-, and one C-κB motifs, in that order, hence are labeled LR-FHHC. An alignment of several representative 4-κB viral sequences shows a high degree of sequence conservation at all the TFBS (**Supplementary Figure S1C**).

## Double-RBEIII LTRs Show Profound Variation in Number, Genetic Sequence, and Position of TFBS

Category-3 viral strains, representing 18.7% (85/455) sequences analyzed here, are characterized by the duplication of the RBEIII motif (**Figure 1B**), a duplication analogous of MFNLP described previously in HIV-1B (Estable et al., 1996). This group is denoted as "RR" to signify the duplication of the RBEIII motif (R; **Figure 2A**). Of note, although the RBEIII motif duplication is

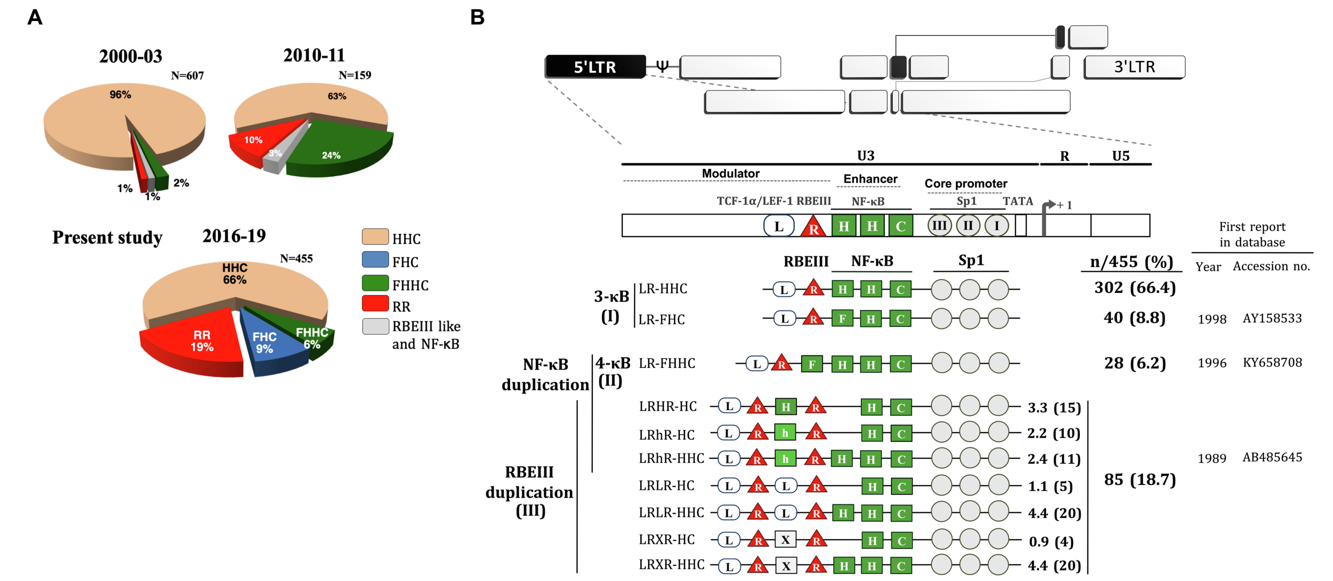


**FIGURE 1 |** Profile of HIV-1C promoter variants in the Indian population. **(A)** The time of sample collection, sample number, and the nature of the promoter configuration are illustrated. The pie charts represent the percentage prevalence of the variant viral strains as color-coded – canonical HHC (beige), FHC (blue), FHHC (green), RR (duplication of RBEIII; red), and duplication of RBEIII like and NF-κB (grey). The data for the periods 2000–2003 and 2010–2011 are adapted from Bachu et al. (2012a) and replotted. The data of the present study are presented in the pie chart 2016–19. **(B)** The magnitude of TFBS variation in HIV-1C LTR. The upper panel represents the genome organization of HIV-1 followed by the TFBS arrangement in the canonical HIV-1C LTR. Sp1 motifs are depicted as grey circles, RBEIII motifs (R) as red triangles, and TCF-1α/LEF-1 sites (L) as open rectangular boxes. The various types of NF-κB binding sites (H, C, F, and h) are depicted as green square boxes. The various HIV-1C viral strains are classified into three main categories based on the NF-κB and/or RBEIII motif duplication. (I) The 3-κB LTR viral strains. (II) The canonical 4-κB LTR viral strains. (III) The viral strains containing the RBEIII site duplication. The two RBEIII sites are separated by an interceding sequence that constitutes an additional copy of a κB-motif (H), κB-like motif (h), TCF-1α/LEF-1 motif (L), or sequence without a distinct pattern (X). The analysis represents 455 of the 518 LTR sequences, and we could not type 63 other sequences.

**TABLE 1 |** Proportion of promoter variants at each clinical site.

| Category | Variant | AIIMS | NARI | St. John's Hospital | YRG CARE | All clinics |
|---|---|---|---|---|---|---|
| I | HHC | 66 (61.6%) | 38 (62.3%) | 79 (68.1%) | 119 (69.6%) | 302 (66.4%) |
|  | FHC | 13 (12.2%) | 5 (8.2%) | 8 (6.9%) | 14 (8.2%) | 40 (8.8%) |
| II | FHHC | 8 (7.5%) | 6 (9.8%) | 9 (7.8%) | 5 (2.9%) | 28 (6.2%) |
| III | RR | 20 (18.7%) | 12 (19.7%) | 20 (17.2%) | 33 (19.3%) | 85 (18.7%) |
| Total |  | 107 | 61 | 116 | 171 | 455 |

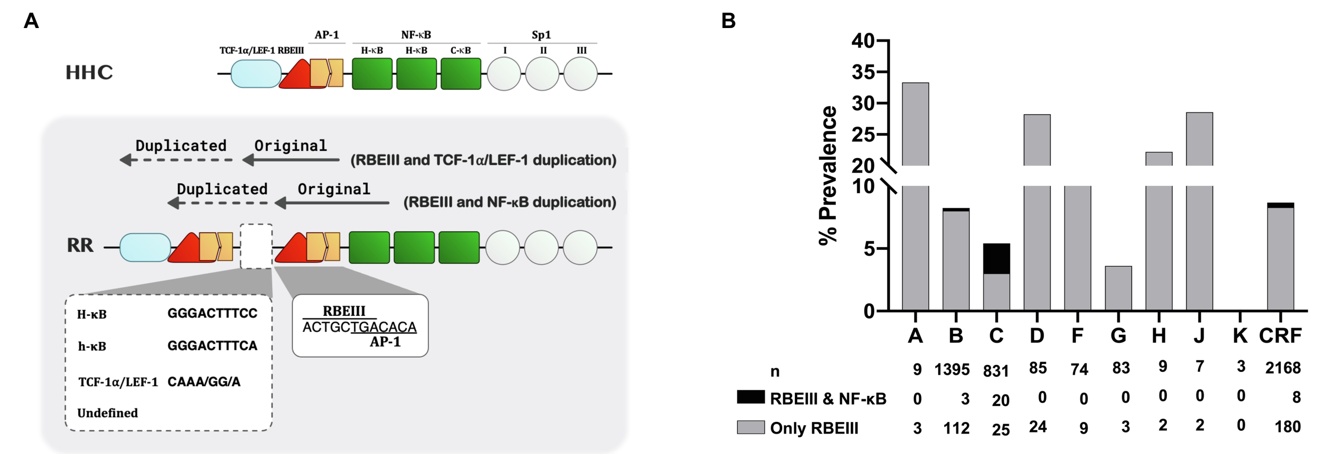


**FIGURE 2 |** The nature of the RBEIII cluster duplication in HIV-1C. **(A)** A schematic representation of the RBEIII cluster duplication in the RR group of HIV-1C variants. The top panel depicts the arrangement of TFBS in the canonical HHC-LTR. The bottom panel portrays the RBEIII cluster duplication (RR, two RBEIII clusters) in two variant LTRs – the co-duplication of RBEIII and TCF-1α/LEF motifs or RBEIII and NF-κB motifs. Of note, the RBEIII cluster duplication comprises the copying of a sequence that recruits both RBF-2 and AP-1 (c-Jun and ATF) collectively. The original and duplicated sequences are marked with solid and dotted arrows, respectively. The two RBEIII clusters are typically separated from each other by an intervening sequence that constitutes a binding site for NF-κB (a canonical H-κB site or a non-canonical h-κB site), TCF-1α/LEF-1 motif, or a sequence of undefined character. **(B)** The relative prevalence of the RBEIII and/or NF-κB motif duplication in HIV-1 subtypes. All the available LTR sequences containing the duplication of one or both TFBS were downloaded from the LANL HIV database and categorized under the major HIV-1 subtypes as shown. A single sequence per patient was included in the analysis. The grey and black bars represent the prevalence of duplication of RBEIII and NF-κB motifs, respectively. n: the number of sequences, CRF: Circulating recombinant forms.

common to all the HIV-1 subtypes (Estable et al., 1998; Gómez-Román et al., 2000), a co-duplication of RBEIII and NF-κB motifs is unique to HIV-1C, not seen in other HIV-1 subtypes (**Figure 2B**). Several unique molecular properties qualify the duplication of the RBEIII site, as described below.

Firstly, the duplicated sequence consists of three different elements – two of the elements invariably coduplicated, along with a third variable element. The two elements coduplicated are the eight base RBEIII core motif (5′-ACTGCTGA-3′) which binds the RBF-2 factor, and a downstream seven base motif (5′-TGACACA-3′) that forms a binding site for AP-1 factors (Jun, Fos, or ATF; **Figure 2A**). Notably, the RBEIII core sequence overlaps with the binding site of the AP-1 factor. Thus, a part of the duplicated sequence (5′-ACTGCTGACACA-3′,) appears to mediate the binding of a TF complex consisting of RBF-2, AP-1, and this cluster is highly conserved in the double-RBEIII LTRs or RR group. Secondly, in addition to the duplication of the 5′-ACTGCTGACACA-3′ sequence, additional sequences are also coduplicated, forming the basis for further classification of the viral strains. In a canonical HIV-1C LTR, the RBEIII motif is flanked by an upstream TCF-1α/LEF-1 site (5′-TACAAA/GG/A-3′) and a downstream NF-κB element (5′-GGGACTTTCC-3′). When the RBEIII motif is duplicated, one of these two TFBS is also coduplicated serving as an intervening sequence, forming at least two subgroups – viral strains containing the co-duplication of a TCF-1α/LEF-1 or NF-κB site. Thirdly, the variant LTRs contain three or two NF-κB motifs in the enhancer region.

Based on the nature of the intervening sequence, the viral strains may be classified into two major subgroups, one containing the presence of a canonical NF-κB motif (5′-GGGACTTTCC-3′, LRHR-HC), or κB-like motif (5′-GGGACTTTCA-3′, which is denoted as "h" κB-motif in the present work, strains LRhR-HC and LRhR-HHC) and the second subgroup with a TCF-1α/LEF-1 (5′-TACAAA/GG/A-3′, LRLR-HC, and LRLR-HHC) or an incomplete TCF-1α/LEF-1 binding sequence. A third subgroup may also be identified where the intervening sequence does not seem to form a binding site for a defined host factor (LRXR-HC and LRXR-HHC). Multisequence alignments of the variant viral promoters are presented (**Supplementary Figures S1D–J**). Of note, the three NF-κB motifs in LRHR-HC, LRhR-HC, and LRhR-HHC strains are not arranged in tandem, thus, blurring the distinction between the viral enhancer (consisting of the only NF-κB motifs) and the upstream modulatory region.

In a phylogenetic analysis of 455 LTR sequences, compared with 31 reference sequences, all the viral strains of the cohort grouped with HIV-1C ascertaining their genetic identity (**Figure 3**). Notably, the various promoter variant viral sequences combined homogeneously, without forming separate clusters based on the TFBS variation.

## Longitudinal Analysis of Prognostic Markers

Increased transcriptional strength of the LTR may augment plasma viral load modulating various prognostic markers and immune activation markers. To this end, we monitored a few prognostic markers, such as the plasma viral load, CD4 cell count, and soluble CD14 levels, in the blood samples at the baseline and at two or three follow-up time points spaced 6 months apart. Unfortunately, this objective could be fulfilled only partially given practical constraints. Following the primary screening and typing of the viral promoters, we could recruit only 208 of the 455 study participants for the follow-up analysis. Subsequently, with the implementation of the "Test and Treat" policy in 2017 in India, several participants enrolled for the follow-up analysis were excluded from the study. Consequently, the longitudinal analysis could be accomplished only with a small number of study participants, who did not prefer to switch to ART.

The demographic features of the 208 study participants at the baseline are summarized (**Table 2**). Of the 208 study participants, the percentage of female, male, and transgender are 55.3% (115/208), 43.8% (91/208), and 1.0% (2/208), respectively. The average (mean) age of the study participants was 34.4 ± 8.45 years (median = 33 years). For the subsequent analyses, all the viral strains were classified into four categories based on the nature of the TFBS variations identified in the viral promoter – HHC (the conventional LTRs containing the HHC κB-binding sites), FHC (all the three κB-binding sites are genetically distinct in these LTRs), FHHC (the LTRs contain four κB-binding sites), and RR (double-RBEIII strains; the seven viral variant strains are pooled into a single category, given the limited number of samples in individual groups).

We compared the levels of plasma viral load, CD4 cell count, and sCD14 among these four categories at the baseline in a cross-sectional analysis. This analysis did not show a significant difference in any of the three parameters among the four groups (**Figure 4**, left panels). At M0, the median PVL values for the HHC, FHC, FHHC, and RR groups were 12,609.0, 13,553.0, 10,440.0, and 6,321.0 copies/ml, respectively (**Figure 4A**, left panel). The median values of CD4 cell count and sCD14 were also similarly comparable among the four groups. We also compared the three clinical parameters between the baseline and 12 M time point for PVL and sCD14 and baseline, 6 M and 12 M for CD4 cell count (**Figure 4**, right panels). All the three parameters appeared to remain stable without a significant change between the baseline and 12 M. The promoter configuration did not appear to make a significant difference for any of the three parameters examined. For instance, the median PVL values at 12 M were 18,450.0, 52,539.0, 13,266.0, and 6,090.0 copies/ml, for HHC, FHC, FHHC, and RR groups, respectively. The CD4 cell count remained stable over the 12-month observation period among all four groups. The sCD14 levels appeared to show an increasing trend among all four groups at the follow-up; these differences, however, were not statistically significant. Of note, in the complete-case analysis, a trend of low-level plasma viral load was manifested for the RR variants compared to the three other groups (**Figure 4A**, right panel). However, the viral load increased between the time points among all the groups.

## The Coexistence of Viral Variants in Natural Infection

Of the various promoter variant viral strains described in the present work, the emergence of LTRs containing RBEIII duplication is relevant to HIV-1 latent reservoirs. In the context of HIV-1B, the RBEIII site and the AP-1 motif are known to play a predominantly suppressive role, especially in the absence of cellular activation (Naghavi et al., 1999; Bernhard et al., 2013; Kropp et al., 2014). The relative proportion of reads representing single vs. double-RBEIII motif-containing viral sequences in a coinfection may offer leads as per the biological significance of RBEIII duplication in natural infection.

To this end, we identified a subset of four of 85 subjects of our cohort who showed the presence of a coinfection of single- and double-RBEIII viral strains in Sanger sequencing. The clinical profile of the four subjects (2079, 3767, 4084, and VFSJ020) is summarized (**Table 3**). We performed an NGS analysis, using the MiSeq Illumina platform (**Supplementary Figure S2**), of the whole blood genomic DNA and plasma viral RNA of the four subjects at the baseline and two or three follow-up time points.

The NGS data confirmed the coexistence of single- and double-RBEIII viral strains in all four subjects in both the DNA and RNA compartments. Importantly, in two subjects (2079 and 4084), the single-RBEIII strains represented a significantly larger proportion of reads in the plasma RNA compared to the double-RBEIII strains (**Figure 5**). Only in subject VFSJ020, the double-RBEIII reads dominated the single-RBEIII reads at all the time points, whereas in subject 3767, a mixed profile was observed. The data between the replicate samples are consistent with each other ascertaining the reproducibility of the analysis. A broad-level concurrence between the plasma RNA and genomic DNA was also noted.

# DISCUSSION

The key finding of the present work is the continuing evolution of the HIV-1C viral promoter and as a consequence, the emergence of new variants at the population level. In 2004, we reported the emergence of HIV-1C strains containing four copies of the NF-κB binding motif in the viral enhancer for the first time in India (Siddappa et al., 2004; Bachu et al., 2012b). The 4-κB viral strains dominated the canonical viral strains containing three copies of NF-κB motifs in natural

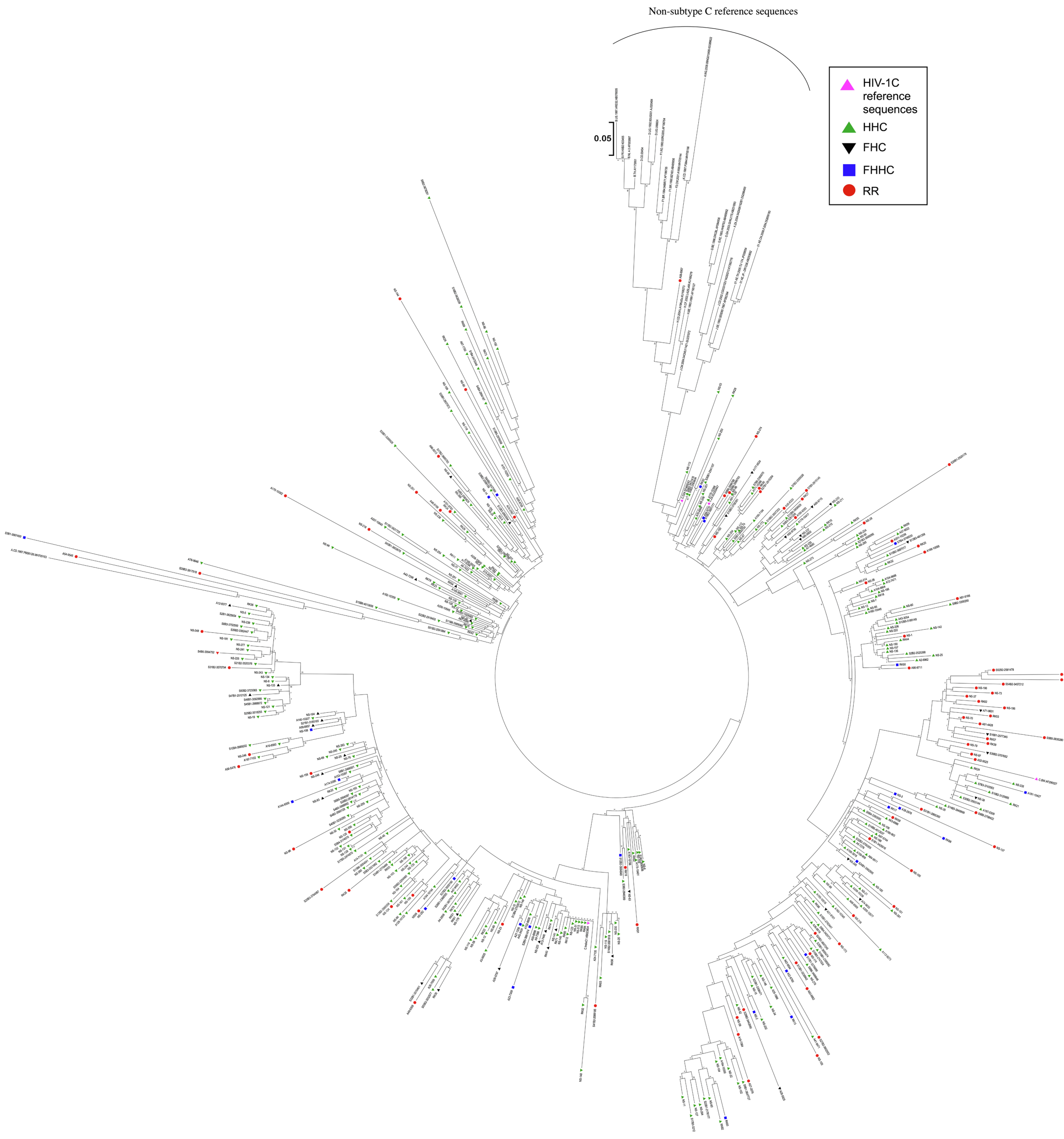


**FIGURE 3 |** Phylogenetic analysis of HIV-1C LTR variant sequences. A total of 455 viral sequences isolated from study participants are included in the analysis. The analysis also includes four HIV-1C reference sequences and three sequences representing each primary genetic subtype of HIV-1, as described in materials and methods. Different LTR variants are represented using different symbols and colors, as depicted. The evolutionary history was inferred by using the maximum likelihood method based on the Tamura-Nei model. The analysis was performed with 1,000 bootstrap values. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. There were a total of 292 positions in the final data set. Evolutionary analyses were performed using MEGA 7.0 software.

infection and under all experimental conditions, alluding to the additional copy of NF-κB motif conferring replication advantage (Bachu et al., 2012a). Subsequently, the 4-κB viral strains were also detected in Brazil and several African countries, suggesting that the phenomenon is not specific to a single country (Boullosa et al., 2014). Although our initial focus was limited to the NF-κB motif duplication and its impact on viral gene expression, we also observed the duplication of other TF binding motifs, including RBEIII and AP-1, though at a lower frequency (Bachu et al., 2012a,b).

## Sequence Motif Duplication in HIV-1C Differs From That of Other HIV-1 Families

Gene duplication accompanied by sequence variation played a crucial role in the acquisition of novel properties leading to the evolutionary success of organisms (Carroll, 2005).

**TABLE 2 |** Characteristics of study participants at the time of recruitment.

| Parameters | |
|---|---|
| Total number of participants | 208 |
| **Gender** | |
| Female | 55.3% (*n* = 115) |
| Male | 43.8% (*n* = 91) |
| Transgender | 1.0% (*n* = 2) |
| **Age (years)** | |
| Mean and SD | 34.4 ± 8.45 |
| Median | 33 |
| ART status | Naïve |

In viruses, the duplication of biologically important sequence motifs may confer the same survival advantage as gene duplication has done in higher organisms (Kropp et al., 2014). The significance of sequence motif duplication in the coding sequences of HIV-1 has attracted more attention conventionally compared to that of regulatory sequences (Marlowe et al., 2004; Guglietta et al., 2010; Sharma et al., 2018). Further, numerous publications reported the deletion or duplication of different regions in the enhancer, core promoter, and modulatory regions that removed or added copies of TFBS (Leonard et al., 1989; Michael et al., 1994; Montano et al., 1997; Naghavi et al., 1999). However, such sequence modifications have been sporadic, typically limited to the individual or a small number of viral strains, and cannot be generalized.

One notable exception to this observation is the occurrence of the MFNLP, which primarily represents the duplication of the RBEIII motif in the viral modulatory region with the concomitant co-duplication of the flanking sequences for the binding of other host factors. The RBEIII motif duplication in HIV-1B was found in approximately 38% of primary viral isolates (Estable et al., 1996). Importantly, RBEIII motif duplication in HIV-1B is believed to ensure the presence of a binding site for RBF-2 when the original copy becomes non-functional due to mutations (Chen et al., 2005; Sadowski and Mitchell, 2005).

RBEIII motif duplication in HIV-1C differs from that of HIV-1B in two crucial qualities. First, the creation of an additional RBEIII motif is not associated with the inactivation of the original motif. In other words, nearly all the double-RBEIII viral strains in our cohort of HIV-1C contained two copies of the intact motif without any mutations in the core sequence. Preliminary leads from our laboratory confirm the functional activity of both the motifs in such LTRs. Of note, the participants of the present study are all reportedly ART-naïve by self-declaration. Our data, however, do not rule out the possibility of ART exposure in HIV-1C leading to the inactivation of the original RBEIII motif, necessitating the need for the creation of a second and functional RBEIII motif in the promoter. Second, the co-duplication of the RBEIII and NF-κB motifs is unique to HIV-1C, a property not seen in any other HIV-1 genetic subtype. Thus, HIV-1C appears to exploit the strategy of sequence motif duplication differently compared to other viral subtypes.

Importantly, the addition of more copies of NF-κB to the viral promoter may be beneficial by enhancing the transcriptional strength of the LTR. However, a stronger LTR can be detrimental to maintaining stable latency. HIV-1C appears to have found two different solutions to the paradox of gene expression regulation – limiting the copy number of the NF-κB motifs to three and duplicating the RBEIII motif.

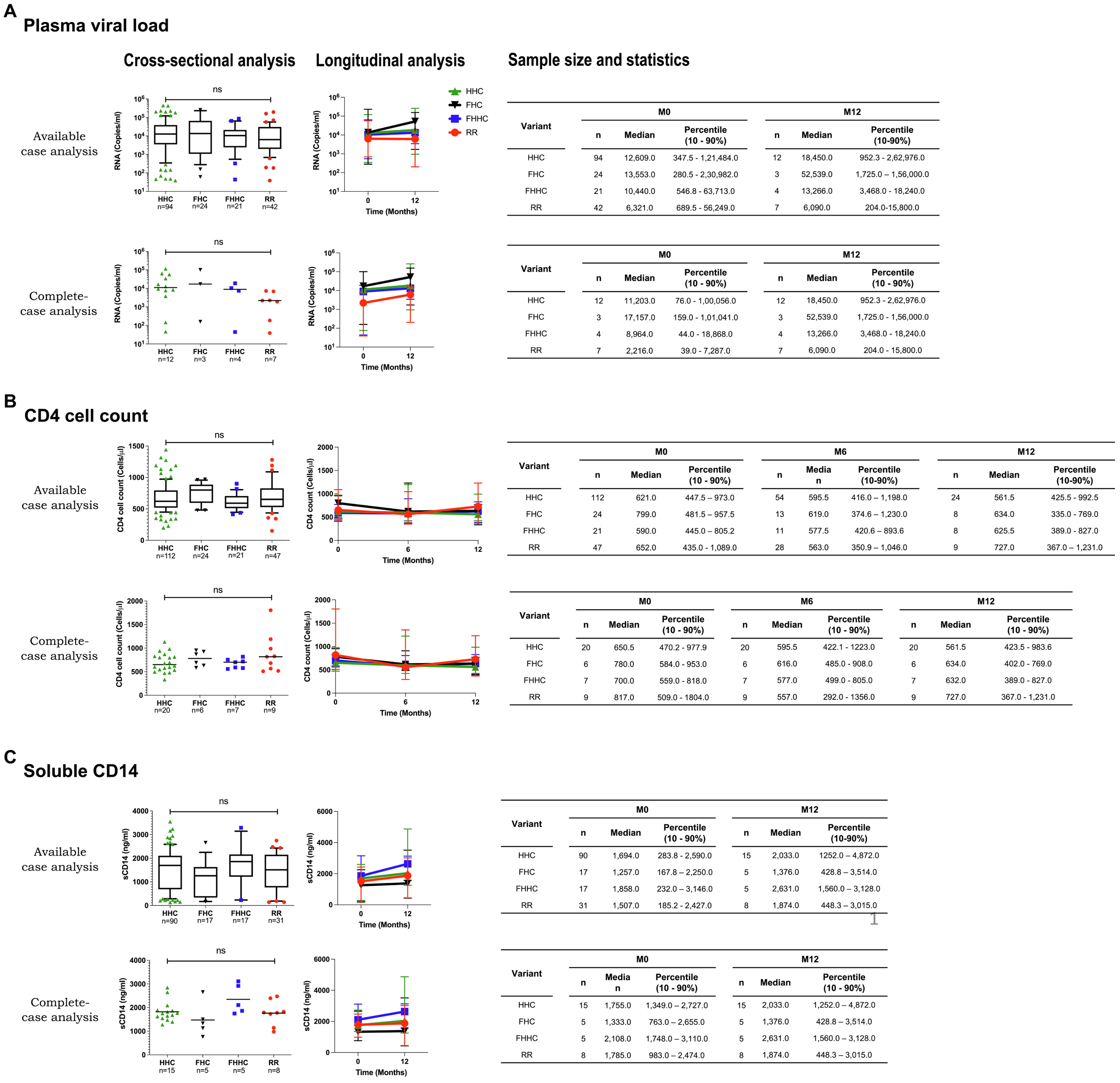
## Limiting the Number of NF-κB Motifs in the Viral Enhancer

Three viral strains, LR-HHC, LR-FHHC, and LR-FHC, lack RBEIII duplication. The prevalence of the LR-FHHC viral strains was only 2% (13 of 607 primary viral isolates) in a southern Indian cohort when discovered during 2000–2003 for the first time (Bachu et al., 2012a). The prevalence of these strains increased to approximately 25% (39/159) during 2010–2011, evaluated at four different clinical sites of India, suggesting transmission success of 4-κB viral strains at the population level (Bachu et al., 2012a). However, in the present study, the prevalence of the LR-FHHC viral strains dropped to 6.2% (28 of 455) during 2017–2019. Notably, a new variant viral group, LR-FHC representing the second-largest proportion among the emerging variants with 8.8% (40 of 455), was identified here for the first time. Given the reduction in the prevalence of LR-FHHC strains and the concomitant appearance of the LR-FHC strains, the former may have originated from the latter.

This observation leads to three logical conclusions. First, LR-FHHC strains, given the stronger transactivation properties of the promoter, may lack replication competence over a sustained period explaining the transient nature of their prevalence in the population. Second, the 4-κB viral strains must relinquish one κB-motif to regain the 3-κB formulation of the enhancer to down modulate the transcriptional strength of the viral promoter. The LR-FHHC strains relinquished one of the two H-κB sites to this end to become LR-FHC. Three, both the canonical LR-HHC and the variant LR-FHC strains contain the same number of NF-κB motifs in the enhancer. However, all the three κB motifs of the FHC-LTR are genetically variable. We propose that the LR-FHC-LTR is likely to be responsive to a broader range of cellular activation signals compared to the LR-HHC-LTR, given the NF-κB motif variation. Thus, by deleting one H-κB site from the LR-FHHC-LTR, HIV-1C appears to have down-modulated transcriptional strength of the viral promoter on the one hand but retained the broader reception potential to cellular signals on the other hand. If the LR-FHC viral strains enjoy a replication advantage at the population level, they are expected to replace the canonical HHC strains in the coming years.

## Is the RBEIII Motif Duplication Essential to Impose Avid Latency of a Stronger Viral Promoter?

Seven different variant LTRs identified in this work contain a second copy of the RBEIII motif added by sequence motif duplication (**Figure 1; Supplementary Figure S1D–J**). Unlike



| Variant | M0 |  |  | M12 |  |  |
|---|---|---|---|---|---|---|
|  | n | Median | Percentile (10 - 90%) | n | Median | Percentile (10-90%) |
| HHC | 94 | 12,609.0 | 347.5 - 1,21,484.0 | 12 | 18,450.0 | 952.3 - 2,62,976.0 |
| FHC | 24 | 13,553.0 | 280.5 - 2,30,982.0 | 3 | 52,539.0 | 1,725.0 – 1,56,000.0 |
| FHHC | 21 | 10,440.0 | 546.8 - 63,713.0 | 4 | 13,266.0 | 3,468.0 - 18,240.0 |
| RR | 42 | 6,321.0 | 689.5 - 56,249.0 | 7 | 6,090.0 | 204.0-15,800.0 |

| Variant | M0 |  |  | M12 |  |  |
|---|---|---|---|---|---|---|
|  | n | Median | Percentile (10 - 90%) | n | Median | Percentile (10 - 90%) |
| HHC | 12 | 11,203.0 | 76.0 - 1,00,056.0 | 12 | 18,450.0 | 952.3 - 2,62,976.0 |
| FHC | 3 | 17,157.0 | 159.0 - 1,01,041.0 | 3 | 52,539.0 | 1,725.0 - 1,56,000.0 |
| FHHC | 4 | 8,964.0 | 44.0 - 18,868.0 | 4 | 13,266.0 | 3,468.0 - 18,240.0 |
| RR | 7 | 2,216.0 | 39.0 - 7,287.0 | 7 | 6,090.0 | 204.0 - 15,800.0 |

| Variant | M0 |  |  | M6 |  |  | M12 |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | n | Median | Percentile (10 - 90%) | n | Media n | Percentile (10-90%) | n | Median | Percentile (10-90%) |
| HHC | 112 | 621.0 | 447.5 – 973.0 | 54 | 595.5 | 416.0 – 1,198.0 | 24 | 561.5 | 425.5 - 992.5 |
| FHC | 24 | 799.0 | 481.5 – 957.5 | 13 | 619.0 | 374.6 – 1,230.0 | 8 | 634.0 | 335.0 - 769.0 |
| FHHC | 21 | 590.0 | 445.0 – 805.2 | 11 | 577.5 | 420.6 – 893.6 | 8 | 625.5 | 389.0 - 827.0 |
| RR | 47 | 652.0 | 435.0 - 1,089.0 | 28 | 563.0 | 350.9 – 1,046.0 | 9 | 727.0 | 367.0 – 1,231.0 |

| Variant | M0 |  |  | M6 |  |  | M12 |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | n | Median | Percentile (10 - 90%) | n | Median | Percentile (10 - 90%) | n | Median | Percentile (10 - 90%) |
| HHC | 20 | 650.5 | 470.2 - 977.9 | 20 | 595.5 | 422.1 - 1223.0 | 20 | 561.5 | 423.5 - 983.6 |
| FHC | 6 | 780.0 | 584.0 - 953.0 | 6 | 616.0 | 485.0 - 908.0 | 6 | 634.0 | 402.0 - 769.0 |
| FHHC | 7 | 700.0 | 559.0 - 818.0 | 7 | 577.0 | 499.0 - 805.0 | 7 | 632.0 | 389.0 - 827.0 |
| RR | 9 | 817.0 | 509.0 - 1804.0 | 9 | 557.0 | 292.0 - 1356.0 | 9 | 727.0 | 367.0 - 1,231.0 |

| Variant | M0 |  |  | M12 |  |  |
|---|---|---|---|---|---|---|
|  | n | Median | Percentile (10 - 90%) | n | Median | Percentile (10-90%) |
| HHC | 90 | 1,694.0 | 283.8 - 2,590.0 | 15 | 2,033.0 | 1252.0 – 4,872.0 |
| FHC | 17 | 1,257.0 | 167.8 – 2,250.0 | 5 | 1,376.0 | 428.8 – 3,514.0 |
| FHHC | 17 | 1,858.0 | 232.0 – 3,146.0 | 5 | 2,631.0 | 1,560.0 – 3,128.0 |
| RR | 31 | 1,507.0 | 185.2 - 2,427.0 | 8 | 1,874.0 | 448.3 – 3,015.0 |

1

| Variant | M0 |  |  | M12 |  |  |
|---|---|---|---|---|---|---|
|  | n | Media n | Percentile (10 - 90%) | n | Median | Percentile (10 - 90%) |
| HHC | 15 | 1,755.0 | 1,349.0 – 2,727.0 | 15 | 2,033.0 | 1,252.0 – 4,872.0 |
| FHC | 5 | 1,333.0 | 763.0 – 2,655.0 | 5 | 1,376.0 | 428.8 – 3,514.0 |
| FHHC | 5 | 2,108.0 | 1,748.0 – 3,110.0 | 5 | 2,631.0 | 1,560.0 – 3,128.0 |
| RR | 8 | 1,785.0 | 983.0 – 2,474.0 | 8 | 1,874.0 | 448.3 – 3,015.0 |

**FIGURE 4 |** Cross-sectional and longitudinal analyses of prognostic markers. **(A)** Plasma viral load, **(B)** CD4 cell count, and **(C)** soluble CD14 levels of the four groups are presented at the baseline (left panels) and follow-up points (right panels). An available-case and complete-case analysis is presented for all the prognostic markers. The sample size used in each evaluation and the corresponding statistics are presented, in the tables. Given the limited sample numbers, the seven RR groups were pooled into a single double-RBEIII arm. Different LTR variant types are represented using different symbols and colors, as depicted. A non-parametric test, that is, the Kruskal-Wallis test, was applied for the statistical analysis of the plasma viral load. One-way ANOVA with Dunnett multiple comparison test was applied to CD4 cell count and sCD14. Two-way ANOVA was used for the comparison of the longitudinal analysis.

in HIV-1B where a new RBEIII site is created as a compensatory mechanism when the original copy is mutated (Estable et al., 1996), in HIV-1C, both the RBEIII motifs are, in contrast, intact without a mutation in the core motif (5′-ACTGCTGA-3′). Thus, RBEIII duplication in HIV-1C appears to confer a novel function or an enhanced phenotype of the existing function but not compensating for a loss of function.

A second quality of the RBEIII motif duplication in HIV-1C is also relevant, especially for HIV latency. While RBEIII motif duplication is common to all the HIV-1 genetic subtypes (**Figure 2B**; Ait-Khaled et al., 1995; Zhang et al., 1997; Estable et al., 1998; Chen et al., 2005), one significant distinction unique to HIV-1C is the co-duplication of the NF-κB motif, not seen in other subtypes. One variant LTR, LRhR-HHC, contains a total of four NF-κB motifs like the

**TABLE 3** | The clinical profile of the four study subjects containing a duplication of the RBEIII motif.

| Subject ID | Age/ Gender | Enrollment date | Promoter variant | PVL (number of RNA copies/ml) | | CD4 (cells/μl) | | | sCDS14 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | M0 | M12 | M0 | M6 | M12 | M0 | M12 |
| 2079 | 32/F | 01/06/2016 | LR-HHC and LRLR-HHC | 185 | 9,989 | 989 | 332 | 558 | 1,146.0 | 729.6 |
| 3767 | 40/F | 29/06/2016 | LR-HHC and LRhR-HHC | 3,742 | - | 566 | 292 | 661 | 139.0 | - |
| 4084 | 38/F | 07/05/2016 | LR-HHC, LR-HC, and LRhR-HC | 6,924 | 5,179 | 509 | 415 | 442 | 983.0 | 448.3 |
| VFSJ020 | 38/F | 22/09/2016 | LR-HHC and LRhR-HC | 21,200 | - | 461 | - | 579 | 2,204.4 | - |

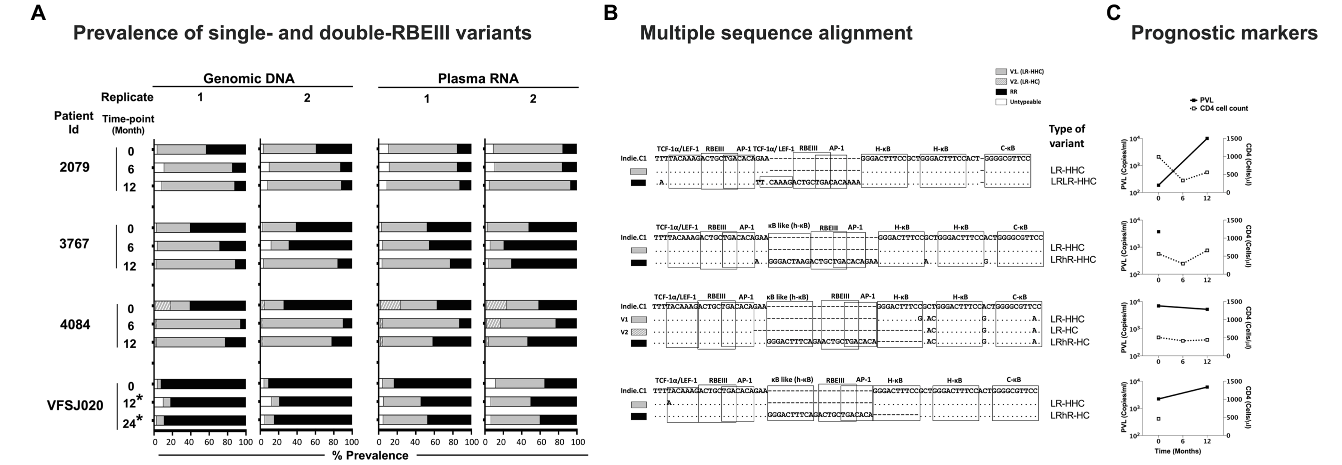


**FIGURE 5** | The frequencies of single- and double-RBEIII variants in a subset of study participants. **(A)** Two independent analyses were performed (replicates 1 and 2) using both the whole blood genomic DNA and plasma viral RNA. The samples were collected at six-month intervals, as shown. An asterisk (*) represents the samples collected post-ART. The dark, grey, and hollow bars represent the percentage prevalence of double-RBEIII, single-RBEIII, and minority/un-typable viral strains, respectively. **(B)** Multiple sequence alignment of single- and double-RBEIII promoter variants in respective subjects, as indicated. TFBS of relevance are marked using open square boxes. The viral variants are aligned with the Indie.C1 reference sequence, which was pulsed in the sequencing sample as an internal control. Dashes and dots represent sequence deletion and sequence homology, respectively. **(C)** Prognostic markers, plasma viral load (PVL), and CD4 cell count are represented by a filled box with a solid line and an open box with a dotted line, respectively.

old FHHC-LTR. However, the variant LTR contains two RBEIII sites, unlike the LR-FHHC-LTR that has only one (**Figure 1**). Additionally, the duplicated κB-motif (h-κB) of LRhR-HHC is genetically distinct (5′-GGGACTTTCA̲-3′) from the other three types (C-, H-, and F-κB) described above. The Single Nucleotide Mutation Model predicted the 5′-GGGACTTTCA̲-3′ motif to bind the p50 homodimer with reduced affinity compared to the consensus NF-κB motif 5′-GGGACTTTCC-3′. This binding prediction was supported by the bimolecular dsDNA microarray analysis, as demonstrated (Du et al., 2014). Lastly, the duplicated κB-motif of LRhR-HHC-LTR is separated from the viral enhancer by one RBEIII motif thus, obliterating the distinction between the viral modulatory and enhancer elements. The biological significance of creating one of each RBEIII and NF-κB motifs in the LRhR-HHC-LTR is of interest.

In the absence of cell activation, the RBEIII motif functions predominantly as a repressive element by recruiting RBF-2 comprising three different cell factors, including TFII-I (Chen et al., 2005). While TFII-I can activate several cellular genes, it can also suppress gene expression from several other cellular promoters, including c-fos (Roy, 2012). Thus, the presence of two copies of the RBEIII motif in the LTR may have a profound impact on viral latency, probably by stabilizing the latency phase. The NF-κB binding motifs, in contrast, play a predominantly positive role in enhancing transcription from the LTR, under the conditions of cell activation. Thus, a higher copy number of the NF-κB motif (4 copies Vs. 3) in the promoter may offset the negative impact of the RBEIII motifs, especially when the provirus is induced out of latency. Of note, unlike the variant LRhR-HHC, a different variant strain LRhR-HC contains one less NF-κB motif (two RBEIII but only three NF-κB motifs). Preliminary results from our laboratory show that the

LRhR-HC-LTR requires a profoundly stronger activation signal, compared to LRhR-HHC-LTR or the canonical LR-HHC-LTR, for latency reversal in Jurkat cells or primary CD4 cells (Bhange et al., unpublished observations).

A different variant promoter pair (LRLR-HHC and LRLR-HC) is also of interest in this respect. This pair also contains an RBEIII motif duplication where the motifs are accompanied by the co-duplication of the TCF-1α/LEF-1 motif, not the NF-κB site. One member of the pair contains three NF-κB motifs (LRLR-HHC), while the other only two (LRLR-HC). This variant promoter pair may have similar gene expression properties as that of the LRhR-HHC- and LRhR-HC-LTRs if the additional copy of TCF-1α/LEF-1 is a functional equivalent of the NF-κB motif.

## The Implication of Promoter Variations for HIV-1 Pathogenesis and Evolution

Since a single promoter regulates the expression of all the HIV-1 proteins and regulates latency, a profound variation in the TFBS composition is expected to have a significant impact on the various properties of the virus, including latency, viral load, disease progression, and viral evolution. The evolution of regulatory elements may play a role as essential or even more important than that of coding sequences (Carroll, 2005). However, little attention has been focused on the evolution of the regulatory elements in HIV-1, than that of the protein-coding regions (Maljkovic Berry et al., 2007).

In our study, the cross-sectional and longitudinal analyses did not find a statistically significant difference in the levels of any of the prognostic markers among the promoter variant strains categorized into four groups, HHC, FHC, FHHC, and RR (**Figure 4**). A significant difference in PVL and CD4 cell count was found in a previous study from our laboratory when a cohort of 80 patients was divided into HHC and FHHC groups (Bachu et al., 2012a). The present study did not find such differences, except for the RR group manifesting a trend in the complete-case analysis, which was not statistically significant. The analytical power of the present study was profoundly compromised given the loss of study participants due to the implementation of the test and treat policy.

Commensurate with our findings, previous studies of the RBEIII motif duplication in HIV-1B also failed to find an association between the LTR profile and clinical or transcriptional phenotypes in a cross-sectional cohort (Estable et al., 1996). A different study failed to identify a correlation between the RBEIII motif duplication and the syncytium-inducing property of envelope and, thus, disease progression (Koken et al., 1994). Likewise, Koken et al. (1994) reported the lack of an association between the copy number of the RBEIII motif and disease progression.

The coexistence of viral strains could be a likely explanation for the absence of association between the LTR variant forms and prognostic markers in our cohort. Deep sequencing of the samples identified the presence of a coinfection in all four subjects in our study. At present, it appears that the double-RBEIII viral infections are found only as coinfections along with the single-RBEIII strains. In three of the four subjects, single-RBEIII viral strains seem to dominate the double-RBEIII variants in both the genomic DNA and RNA compartments and at most of the follow-up time points. If RBEIII cluster duplication indeed manifests a suppressive effect on viral gene expression, a distinct association between the duplication and the prognostic markers may become evident in a mono-infection, but not in a coinfection. The dominant influence of the RBEIII motif duplication on viral gene expression has been confirmed using panels of engineered viral clones (Bhange et al., unpublished observations).

In summary, our work records the emergence of several promoter variant viral strains in HIV-1C of India over recent years. Sequences representing the variant viral forms are also found in the sequence databases derived from different global regions where HIV-1C is predominant. Sequence motif duplication creates additional copies of TFBS that play a crucial role in regulating HIV latency and even blurs the distinction between the viral enhancer and modulatory regions. Given that the RBEIII and AP-1 sites play a crucial role in regulating latency (Duverger et al., 2013), the influence of the RBEIII motif duplication, especially when accompanied by the co-duplication of the NF-κB motifs, needs experimental evaluation. Consistent monitoring will be necessary to understand which variant viral strains will survive to establish spreading epidemics in the coming years. Detailed investigations are warranted to evaluate the impact of the TFBS profile differences on HIV-1 latency and latent reservoir properties. ART administration may have a profound impact on the promoter variations described here by exacerbating the frequency of such sequence duplications. Further, the present study could not examine the influence of the duration of the viral infection and disease state on promoter variations, as such details are not available from clinics. The present study also restricted the scope of sequence duplication evaluation to the LTR and did not examine sequence motif duplication in the other regions of the viral genome.

## DATA AVAILABILITY STATEMENT

The data sets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Ethics Committee of JNCASR, AIIMS, NARI, St. John's Hospital, and YRG CARE. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

DB performed research, analyzed data, and wrote the paper. NP, SPM, BPG, SN, DC, BJ, TRD, SFA, NS, AB, and KM performed research. SSi and YG analyzed data. HP analyzed NGS data. PB, BKD, MD, RG, SM, RSP, SSa, AS, SSo, and MT designed research. UR designed research and wrote the paper. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the Department of Biotechnology, Ministry of Science and Technology, Government of India (Sanction order no. BT/PR7359/MED/29/651/2012).

## ACKNOWLEDGMENTS

We thank Kushgra Bansal for the discussions during the NGS data analysis. We thank the study participants for their participation and cooperation during the complete duration of the clinical study. We thank Ruthu Nagaraj for the help during flow cytometry analysis of the samples from St. John's National Academy of Health Sciences.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.779472/full#supplementary-material

## REFERENCES

Ait-Khaled, M., McLaughlin, J. E., Johnson, M. A., and Emery, V. C. (1995). Distinct HIV-1 long terminal repeat quasispecies present in nervous tissues compared to that in lung, blood and lymphoid tissues of an AIDS patient. *AIDS* 9, 675–684. doi: 10.1097/00002030-199507000-00002

Bachu, M., Mukthey, A. B., Murali, R. V., Cheedarla, N., Mahadevan, A., Shankar, S. K., et al. (2012b). Sequence insertions in the HIV type 1 subtype C viral promoter predominantly generate an additional NF-κB binding site. *AIDS Res. Hum. Retrovir.* 28, 1362–1368. doi: 10.1089/AID.2011.0388

Bachu, M., Yalla, S., Asokan, M., Verma, A., Neogi, U., Sharma, S., et al. (2012a). Multiple NF-κB sites in HIV-1 subtype C long terminal repeat confer superior magnitude of transcription and thereby the enhanced viral predominance. *J. Biol. Chem.* 287, 44714–44735. doi: 10.1074/jbc.M112.397158

Bernhard, W., Barreto, K., Raithatha, S., and Sadowski, I. (2013). An upstream YY1 binding site on the HIV-1 LTR contributes to latent infection. *PLoS One* 8:e77052. doi: 10.1371/journal.pone.0077052

Berry, I. M., Ribeiro, R., Kothari, M., Athreya, G., Daniels, M., Lee, H. Y., et al. (2007). Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. *J. Virol.* 81, 10625–10635. doi: 10.1128/JVI.00985-07

Boullosa, J., Bachu, M., Bila, D., Ranga, U., Süffert, T., Sasazawa, T., et al. (2014). Genetic diversity in HIV-1 subtype C LTR from Brazil and Mozambique generates new transcription factor-binding sites. *Viruses* 6, 2495–2504. doi: 10.3390/v6062495

Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS Biol.* 3:e245. doi: 10.1371/journal.pbio.0030245

Chen, J., Malcolm, T., Estable, M. C., Roeder, R. G., and Sadowski, I. (2005). TFII-I regulates induction of chromosomally integrated human immunodeficiency virus type 1 long terminal repeat in cooperation with USF. *J. Virol.* 79, 4396–4406. doi: 10.1128/JVI.79.7.4396-4406.2005

du, W., Gao, J., Wang, T., and Wang, J. (2014). Single-nucleotide mutation matrix: A new model for predicting the NF-κB DNA binding sites. *PLoS One* 9:e101490. doi: 10.1371/journal.pone.0101490

Duverger, A., Wolschendorf, F., Zhang, M., Wagner, F., Hatcher, B., Jones, J., et al. (2013). An AP-1 binding site in the enhancer/core element of the HIV-1 promoter controls the ability of HIV-1 to establish latent infection. *J. Virol.* 87, 2264–2277. doi: 10.1128/JVI.01594-12

Estable, M. C. (2007). In search of a function for the most frequent naturally-occurring length polymorphism (MFNLP) of the HIV-1 LTR: retaining functional coupling, of Nef and RBF-2, at RBEIII? *Int. J. Biol. Sci.* 3, 318–327. doi: 10.7150/ijbs.3.318

Estable, M. C., Bell, B., Hirst, M., and Sadowski, I. (1998). Naturally occurring human immunodeficiency virus type 1 long terminal repeats have a frequently observed duplication That binds RBF-2 and represses transcription. *J. Virol.* 72, 6465–6474. doi: 10.1128/JVI.72.8.6465-6474.1998

Estable, M. C., Bell, B., Merzouki, A., Montaner, J. S., O'Shaughnessy, M. V., and Sadowski, I. J. (1996). Human immunodeficiency virus type 1 long terminal repeat variants from 42 patients representing all stages of infection display a wide range of sequence polymorphism and transcription activity. *J. Virol.* 70, 4053–4062. doi: 10.1128/jvi.70.6.4053-4062.1996

Gómez-Román, V. R., Vásquez, J. A., Basualdo, M. D. C., Estrada, F. J., Ramos-Kuri, M., and Soler, C. (2000). nef/long terminal repeat quasispecies from HIV type 1-infected mexican patients with different progression patterns and their pathogenesis in hu- PBL-SCID mice. *AIDS Res. Hum. Retrovir.* 16, 441–452. doi: 10.1089/088922200309106

Guglietta, S., Pantaleo, G., and Graziosi, C. (2010). Long sequence duplications, repeats, and palindromes in HIV-1 gp120: length variation in V4 as the product of misalignment mechanism. *Virology* 399, 167–175. doi: 10.1016/j.virol.2009.12.030

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Koken, S. E., van Wamel, J. L. B., Geelen, J. L. M. C., and Berkhout, B. (1994). Functional analysis of the ACTGCTGA sequence motif in the human immunodeficiency virus Type-1 long terminal repeat promoter. *J. Biomed. Sci.* 1, 83–92. doi: 10.1007/BF02257981

Koken, S. E. C., van Wamel, J. L. B., Goudsmit, J., Berkhout, B., and Geelent, J. L. M. C. (1992). Natural variants of the HIV-1 long terminal repeat: analysis of promoters with duplicated DNA regulatory motifs. *Virology* 191, 968–972. doi: 10.1016/0042-6822(92)90274-S

Kropp, K. A., Angulo, A., and Ghazal, P. (2014). Viral enhancer mimicry of host innate-immune promoters. *PLoS Pathog.* 10:e1003804. doi: 10.1371/journal.ppat.1003804

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Leonard, J., Parrott, C., Buckler-White, A. J., Turner, W., Ross, E. K., Martin, M. A., et al. (1989). The NF-kappa B binding sites in the human immunodeficiency virus type 1 long terminal repeat are not required for virus infectivity. *J. Virol.* 63, 4919–4924. doi: 10.1128/jvi.63.11.4919-4924.1989

Locateli, D., Stoco, P. H., de Queiroz, A. T. L., Alcântara, L. C. J., Ferreira, L. G. E., Zanetti, C. R., et al. (2007). Molecular epidemiology of HIV-1 in Santa Catarina state confirms increases of subtype C in southern Brazil. *J. Med. Virol.* 79, 1455–1463. doi: 10.1002/jmv.20955

Marlowe, N., Flys, T., Hackett, J. Jr., Schumaker, M., Jackson, J. B., and Eshleman, S. H. (2004). Analysis of insertions and deletions in the gag p6 region of diverse HIV type 1 strains. *AIDS Res. Hum. Retrovir.* 20, 1119–1125. doi: 10.1089/aid.2004.20.1119

Michael, N. L., D'Arcy, L., Ehrenberg, P. K., and Redfield, R. R. (1994). Naturally occurring genotypes of the human immunodeficiency virus type 1 long terminal repeat display a wide range of basal and tat-induced transcriptional activities. *J. Virol.* 68, 3163–3174. doi: 10.1128/jvi.68.5.3163-3174.1994

Montano, M. A., Novitsky, V. A., Blackard, J. T., Cho, N. L., Katzenstein, D. A., and Essex, M. (1997). Divergent transcriptional regulation among expanding human immunodeficiency virus type 1 subtypes. *J. Virol.* 71, 8657–8665. doi: 10.1128/jvi.71.11.8657-8665.1997

Naghavi, M. H., Schwartz, S., Sonnerborg, A., and Vahlne, A. (1999). Long terminal repeat promoter/enhancer activity of different subtypes of HIV type 1. *AIDS Res. Hum. Retrovir.* 15, 1293–1303. doi: 10.1089/088922299310197

Ramírez de Arellano, E., Soriano, V., Alcamil, J., and Holguín, A. (2006). New findings on transcription regulation across different HIV-1 subtypes. *AIDS Rev.* 8, 9–16.

Rodriguez, M. A., Shen, C., Ratner, D., Paranjape, R. S., Kulkarni, S. S., Chatterjee, R., et al. (2007). Genetic and functional characterization of the LTR of HIV-1 subtypes A and C circulating in India. *AIDS Res. Hum. Retrovir.* 23, 1428–1433. doi: 10.1089/aid.2007.0152

Roy, A. L. (2012). Biochemistry and biology of the inducible multifunctional transcription factor TFII-I: 10 years later. *Gene* 492, 32–41. doi: 10.1016/j.gene.2011.10.030

Sadowski, I., and Mitchell, D. A. (2005). TFII-I and USF (RBF-2) regulate Ras/MAPK-responsive HIV-1 transcription in T cells. *Eur. J. Cancer* 41, 2528–2536. doi: 10.1016/j.ejca.2005.08.011

Sharma, S., Arunachalam, P. S., Menon, M., Ragupathy, V., Satya, R. V., Jebaraj, J., et al. (2018). PTAP motif duplication in the p6 gag protein confers a replication advantage on HIV-1 subtype C. *J. Biol. Chem.* 293, 11687–11708. doi: 10.1074/jbc.M117.815829

Siddappa, N. B., Dash, P. K., Mahadevan, A., Jayasuryan, N., Hu, F., Dice, B., et al. (2004). Identification of subtype C human immunodeficiency virus type 1 by subtype-specific PCR and its use in the characterization of viruses circulating in the southern parts of India. *J. Clin. Microbiol.* 42, 2742–2751. doi: 10.1128/JCM.42.6.2742-2751.2004

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. doi: 10.1093/oxfordjournals.molbev.a040023

Tebit, D. M., and Arts, E. J. (2011). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infect. Dis.* 11, 45–56. doi: 10.1016/S1473-3099(10)70186-9

Yamaguchi, J., Vallari, A., McArthur, C., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2020). Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *J. Acquir. Immune Defic. Syndr.* 83, 319–322. doi: 10.1097/QAI.0000000000002246

Zhang, L., Huang, Y., Yuan, H., Chen, B. K., Ip, J., and Ho, D. D. (1997). Identification of a replication-competent pathogenic human immunodeficiency virus type 1 with a duplication in the TCF-1alpha region but lacking NF-kappaB binding sites. *J. Virol.* 71, 1651–1656. doi: 10.1128/jvi.71.2.1651-1656.1997

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 31 January 2022
doi: 10.3389/fmicb.2021.828765

# Updated HIV-1 Consensus Sequences Change but Stay Within Similar Distance From Worldwide Samples

*Gregorio V. Linchangco Jr., Brian Foley and Thomas Leitner**

*Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, United States*

OPEN ACCESS

***Edited by:***
*Michael M. Thomson,*
*Instituto de Salud Carlos III (ISCIII),*
*Spain*

***Reviewed by:***
*Vlad Novitsky,*
*HSPH, United States*
*Antoinette Van Der Kuyl,*
*University of Amsterdam, Netherlands*

***Correspondence:**
*Thomas Leitner*
*tkl@lanl.gov*

***Specialty section:***
*This article was submitted to*
*Virology,*
*a section of the journal*
*Frontiers in Microbiology*

***Received:*** *03 December 2021*
***Accepted:*** *20 December 2021*
***Published:*** *31 January 2022*

***Citation:***
*Linchangco GV Jr, Foley B and Leitner T (2022) Updated HIV-1 Consensus Sequences Change but Stay Within Similar Distance From Worldwide Samples.*
*Front. Microbiol. 12:828765.*
*doi: 10.3389/fmicb.2021.828765*

HIV consensus sequences are used in various bioinformatic, evolutionary, and vaccine related research. Since the previous HIV-1 subtype and CRF consensus sequences were constructed in 2002, the number of publicly available HIV-1 sequences have grown exponentially, especially from non-EU and US countries. Here, we reconstruct 90 new HIV-1 subtype and CRF consensus sequences from 3,470 high-quality, representative, full genome sequences in the LANL HIV database. While subtypes and CRFs are unevenly spread across the world, in total 89 countries were represented. For consensus sequences that were based on at least 20 genomes, we found that on average 2.3% (range 0.8–10%) of the consensus genome site states changed from 2002 to 2021, of which about half were nucleotide state differences and the rest insertions and deletions. Interestingly, the 2021 consensus sequences were shorter than in 2002, and compared to 4,674 HIV-1 worldwide genome sequences, the 2021 consensuses were somewhat closer to the worldwide genome sequences, i.e., showing on average fewer nucleotide state differences. Some subtypes/CRFs have had limited geographical spread, and thus sampling of subtypes/CRFs is uneven, at least in part, due to the epidemiological dynamics. Thus, taken as a whole, the 2021 consensus sequences likely are good representations of the typical subtype/CRF genome nucleotide states. The new consensus sequences are available at the LANL HIV database.

**Keywords: HIV, subtypes, consensus sequences, evolution, molecular epidemiology, pandemic**

## INTRODUCTION

In 2020, 37.7 million people worldwide were living with HIV, of which 1.5 million became infected in 2020. Until 2020, 36.3 million people have died from AIDS-related illnesses (UNAIDS, 2021). Most of these infections are by HIV-1. The burden of HIV is uneven across the world, between countries, within and between risk groups, and between ethnic groups in different geographical regions. In large due to founder effects, different genetic variants, i.e., subtypes and circulating recombinant forms (CRFs), have spread unevenly across the world (Hemelaar et al., 2019, 2020).

While analyses of individual HIV sequences provides comprehensive information about worldwide and local epidemics as well as detailed information about within-host evolution, global reference sequences have many uses. One type of reference sequences is consensus sequences, i.e., a sequence that represents the most commonly found nucleotide (or amino acid) at each site.

Such sequences are useful as references for bioinformatic processing in, for instance, alignments and contig assembly, for detection of hypermutants, gene detection and annotation, and for representing simplified views and data from complex populations (Rose and Korber, 2000; Lee, 2003; Seah et al., 2020; Domingo et al., 2021; Frith et al., 2021; Kulikova et al., 2021; Zhang et al., 2021). Consensus sequences have also been used in studies of protein functions, binding, and vaccine designs (Novitsky et al., 2002; Gao et al., 2005; Nickle et al., 2007; Yan et al., 2007; Sternke et al., 2019).

The LANL HIV database (Foley et al., 2018) provides global consensus sequences for HIV-1 subtypes and CRFs. The most recent genome level consensus sequences are from 2002 (and some gene specific consensus sequences from 2004). Since 2002, the number of available sequences in the database has grown exponentially, from 85,926 to 1,073,050 in 2021, a >12-fold increase (**Figure 1**). Similarly, near full length genomes (sequences > 7,000 nt long) have increased from 574 to 21,952, a massive > 38-fold increase. Over this time, sequencing of non-EU and non-US samples has increased the most, and thus the increase mostly reflects HIV-1 sequences from the rest of the world, where most of the infected people live (UNAIDS, 2021). Therefore, it is necessary to re-evaluate the global consensus sequences.

## MATERIALS AND METHODS

### Sequence Data

To generate new HIV-1 consensus sequences, we used the LANL HIV database 2019 filtered web alignments of full genomes, available at https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html. This alignment is a high-quality selection of the complete 2019 web alignment. The sequences in this set have no or only one minor frameshift, <1% nucleotide ambiguities, no nucleotide ambiguities that affect translation, and no unusual indels. This set was considered ideal for global consensus sequence generation. This set contained 4,312 sequences. For comparison to our new consensus sequences, we used the latest previously calculated consensus sequences, from 2002, also available at https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html.

To evaluate how distant actual HIV-1 genomes are from the consensus sequences, we included (1) HIV-1 genome sequences with >7,000 nt, (2) sequences that have a sampling year, (3) sequences that were not labeled as "problematic" in the LANL HIV database (see https://www.hiv.lanl.gov/components/sequence/HIV/search/help.html for an explanation of what "problematic" means), and (4) restricted the data to only include one sequence per patient when >1 sequence was known to come from a patient. This set contained 4,674 sequences, accessed 2021-06-23.

### Consensus Calculation

Consensus sequence calculations were performed with the Advanced Consensus Maker, available at https://www.hiv.lanl.gov/content/sequence/CONSENSUS/AdvCon.html. We used a minimum of three sequences per HIV-1 subtype or circulating recombinant form (CRF) to generate new consensus sequences (reducing the number of useable sequences to 3,470 from the 2019 web alignment of 4,312 sequences), a majority rule that assigns the most common nucleotide state to each site, tie-breaking that follows the typical nucleotide frequency in HIV-1 sequences (i.e., priority in order A, G, T, C), and no gap removal. These settings are the current defaults for these consensus calculations, and have been used for the previous consensus sequence calculations at the LANL HIV database.

### Sequence Comparisons

Pairwise alignments were made with MAFFT V7 (Katoh and Standley, 2013), followed by codon correction using GeneCutter,[1] in all sequence comparisons. Pairwise comparisons were performed between previous and new consensus sequences as well as between individual HIV-1 genome sequences (>7,000 nt) and consensus sequences (**Figure 2**). Each pairwise alignment was then analyzed with a custom python script that counted state changes, insertions, deletions, and sequence length. Flanking gaps in each pairwise alignment were ignored. The R programming environment and ggplot (R Development Core Team, 2003; Wickham, 2016) were used to generate violin plots to display distributions of these categories, and Wilcoxon rank sum tests with Bonferroni multiple-test correction to assess potential differences.

## RESULTS

### Changes in HIV-1 Consensus Sequences

The number of HIV-1 sequences in the LANL HIV database has grown over time (**Figure 1**). Both the total number of sequences and the number of near full genomes (>7,000 nt) has grown roughly exponentially. The substantial growth of the database since 2002, when genome level consensus sequences were last updated, motivated us to assess potential changes in the consensus sequences. In total, 90 new HIV-1 subtype or CRF consensus sequences were generated based on at least three available near full genome sequences in each such set (**Supplemental Results**). Out of those, 18 subtypes/CRFs (and CPZ) allowed for comparison between the 2002 and 2021 consensus sequences (**Table 1**). In 2002, only four of these subtype consensus sequences were based on a substantial number of sequences (A1, B, C, and D used > 30 sequences), while the rest used <10 sequences each. In 2021, nearly all used substantial numbers; subtypes B and C, the two most sequenced subtypes in the database, used 1,294 and 744 sequences, respectively, for the 2021 consensus sequences. Typically, the 2021 consensus sequences were shorter than in 2002, i.e., they had more "deletions" than "insertions" relative to the 2002 consensuses. Typically, there were also many "substitutions" between the 2002 and 2021 consensuses, on average 109 nucleotide state differences across the entire genome (1.1%), excluding HIV-1 group O and CPZ consensuses, which had more. Overall, counting all indel

[1] https://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html

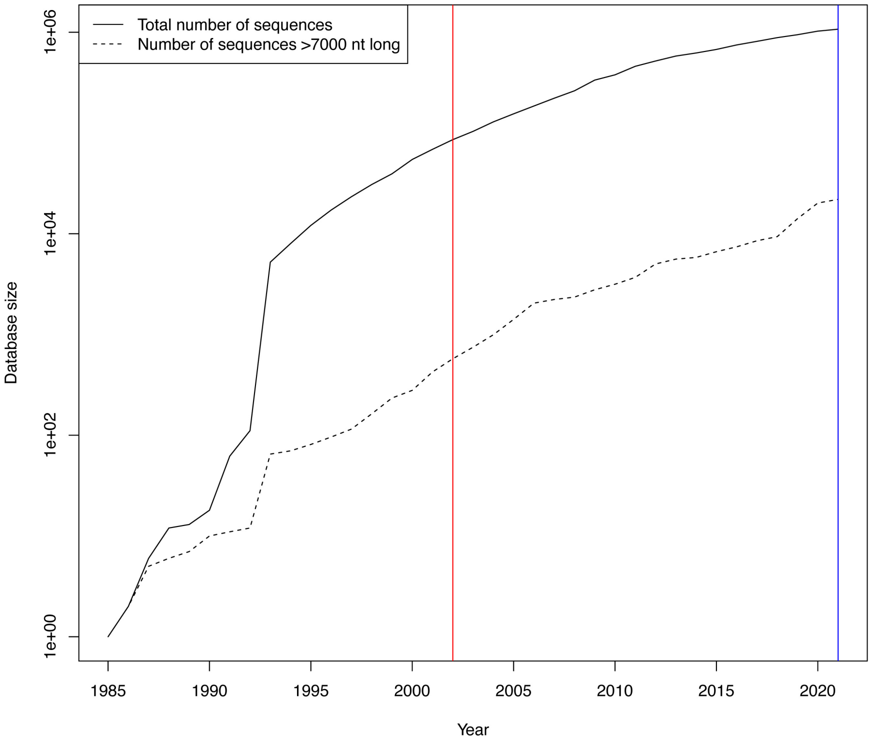


**FIGURE 1 |** Growth of HIV-1 sequences in the LANL HIV database. The growth of the number of publicly available HIV sequences has been roughly exponential since the beginning of the HIV era. The y-axis is logarithmic to make the near full genome (>7,000 nt) sequence count visible. The red vertical line shows when the last previous HIV consensus sequences were calculated in 2002, and the blue line when we calculated the new ones in this publication in 2021.

and nucleotide state differences (including those in group O and CPZ), on average 2.3% (range 0.8–10%) of the consensus genomes changed from 2002 to 2021.

Interestingly, non-synonymous "substitutions" dominated in the 2002 to 2021 consensus comparisons (**Figure 3**). Overall, "substitutions" in codon positions 1 and 2 were about 3.5 times more frequent than in codon position 3. This result should not be surprising because the "substitutions" (as well as "insertions" and "deletions") are simply differences between the 2002 and 2021 consensus sequences, which are manmade constructs not only reflecting evolutionary processes but also sampling effects. On the other hand, most nucleotide state differences ("substitutions") occurred in *env*, and least in *pol* (**Figure 3**), which is expected from the known differences in the evolutionary rate across the HIV-1 genome.

## Consensus Sequences Remain Equally Distant From Worldwide Sequences Over Time

Even though the consensus sequences have changed since 2002 until 2021 (**Table 1**), most subtypes/CRFs have stayed within a similar genetic distance to the consensuses over this time span (**Figure 4**). We compared eight subtypes/CRFs that had at least 20 worldwide genome sequences sampled in 2002 (and 2021). Overall, 2021 consensuses were somewhat closer to the worldwide genome sequences, i.e., showing on average fewer nucleotide state differences, but only subtypes B, G, and group O sequences displayed significant differences (**Figure 4A**).

To assess whether the changes in the 2021 consensus sequences induced significant differences over time, we compared the 2021 consensus sequences to genome sequences sampled until 2002 or 2021, i.e., the 2021 set had additional sequences that became available after 2002 ("N Genome Seq" columns in **Table 1**). Again, on average most subtypes/CRFs showed no significant change in their distance to the worldwide sequences available until 2002 or 2021 (**Figure 4B**). Only group O sequences showed a significant difference. We note that group O consensus sequences had the biggest change from 2002 to 2021 (401 nucleotide state changes) and a 29% growth in available genome sequences (**Table 1**).

While comparing 2002–2021 consensuses to each other showed more deletions than insertions (**Table 1**), comparing consensuses to worldwide genome sequences showed the opposite (**Supplementary Figures 1**, **2**). Thus, Subtypes/CRFs 01_AE, 02_AG, B, C, D, and group O had significant changes in insertions, while only 01_AE, B, and C showed significant changes in deletions.

## DISCUSSION

The LANL HIV database has grown exponentially, adding hundreds of thousands of sequences since the 2002 and

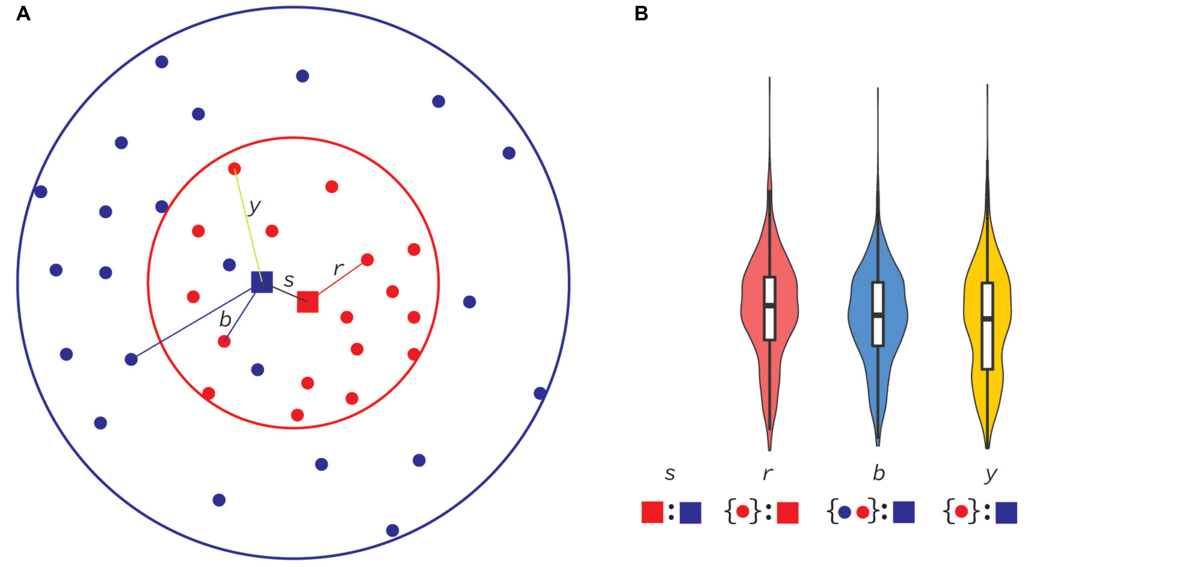


**FIGURE 2 |** Principal distances between consensus and database sequences. In this cartoon, sequences evolve through mutations over time, radiating out from the origin (center of circles) and sampled through time **(A)**. At some time (red circle) all sequences sampled until then (red dots) are used to compute a consensus sequence (red square). Individual distances from that consensus sequence to all sequences available until then form a distance distribution in panel **(B)**, displayed as a red violin plot of all distances *r*. An example of a *r* distance is shown in panel **(A)**. At a later time (blue circle), a new consensus sequence is computed (blue square), and, similarly, all distances (*b*) to sequences available until that time (red and blue dots) form the blue distribution in panel **(B)**. The distance between the first and second consensus sequences is *s*. We can also consider distances from the second consensus sequence to samples collected only available until the first consensus was made (*y*). Note that some samples that originated from a time before the first consensus was made were not publicly available until the second consensus was made (blue dots inside red circle).

**TABLE 1 |** 2002–2021 HIV-1 consensus sequence comparison.

| Subtype/CRF | Insertions | Deletions | Substitutions | N seq used in cons 2002 | N seq used in cons 2021 | N genome seq in 2002 | N genome seq in 2021 |
|---|---|---|---|---|---|---|---|
| A1 | 3 | 10 | 60 | 40 | 173 | 57 | 188 |
| B | 3 | 403 | 96 | 31 | 1,294 | 326 | 2,024 |
| C | 6 | 35 | 56 | 66 | 744 | 189 | 1,214 |
| D | 0 | 25 | 68 | 33 | 71 | 53 | 77 |
| F1 | 17 | 23 | 135 | 4 | 42 | 12 | 73 |
| G | 9 | 22 | 205 | 5 | 80 | 21 | 85 |
| H | 16 | 4 | 221 | 3 | 10 | 8 | 10 |
| O | 24 | 97 | 401 | 4 | 49 | 35 | 45 |
| 01_AE | 4 | 110 | 52 | 9 | 350 | 122 | 636 |
| 02_AG | 4 | 66 | 94 | 7 | 130 | 49 | 160 |
| 04_CPX | 29 | 13 | 109 | 3 | 5 | 5 | 5 |
| 06_CPX | 10 | 21 | 118 | 4 | 11 | 4 | 11 |
| 07_BC | 1 | 46 | 86 | 3 | 22 | 2 | 38 |
| 08_BC | 6 | 12 | 121 | 4 | 21 | 8 | 33 |
| 10_CD | 15 | 16 | 51 | 3 | 3 | 3 | 3 |
| 11_CPX | 8 | 20 | 149 | 6 | 22 | 12 | 23 |
| 12_BF | 20 | 10 | 53 | 6 | 9 | 12 | 15 |
| 14_BG | 27 | 7 | 91 | 6 | 5 | 8 | 12 |
| CPZ | 181 | 62 | 736 | 5 | 21 | 7 | 18 |

*Insertions, deletions, and substitutions are relative differences comparing 2002–2021 consensus sequences.*

thousands of full genome sequences that informed the new HIV-1 subtype/CRF consensus sequences in this study (in 2021). The new consensuses differed overall in about 2.3% of the genome, of which about half were nucleotide state differences. Of that, nearly 3/4 were non-synonymous changes, i.e., changes inducing amino acid differences. Such changes may be important

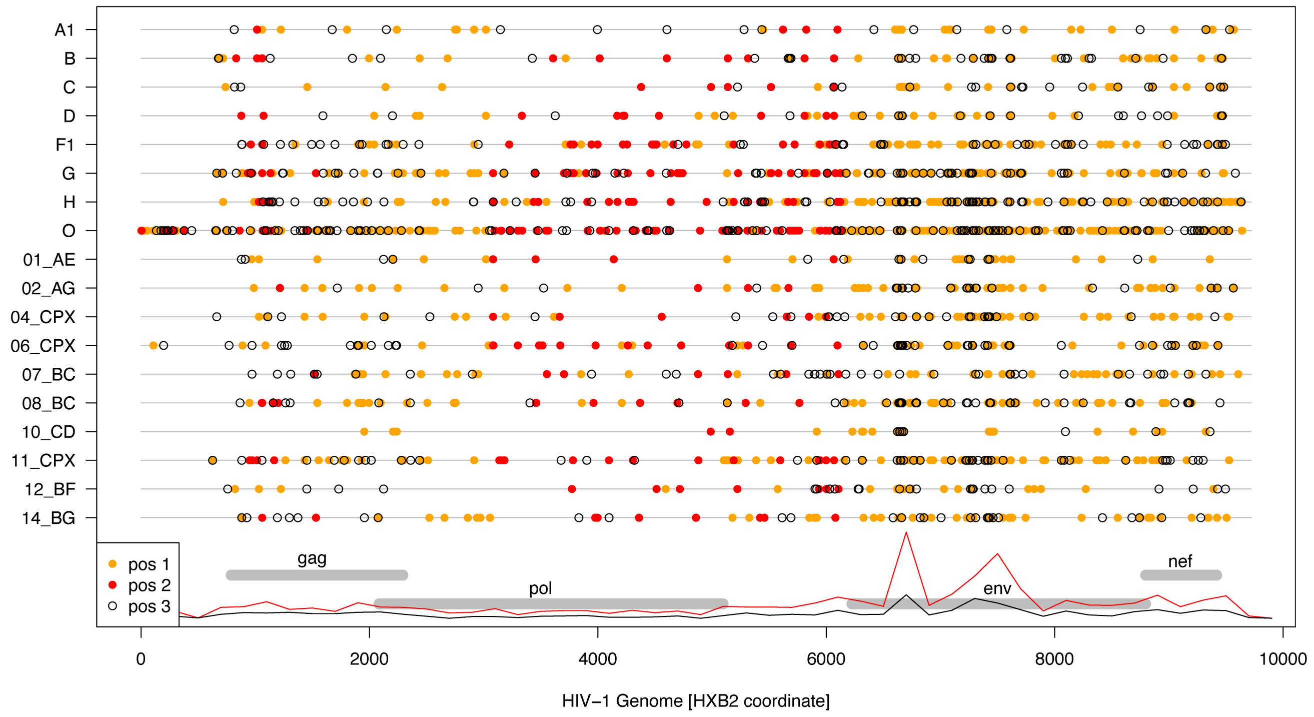


**FIGURE 3 |** Differences between 2002 and 2021 HIV-1 subtype/CRF consensus sequences. Dots show nucleotide state differences for 1st (orange), 2nd (red), and 3rd (open) codon positions along the respective subtype/CRF genome (gray lines). The genome locations for the structural genes are shown for reference, and overlayed with relative densities of 1st + 2nd (red) and 3rd (black) codon position differences across all subtypes/CRFs.

for vaccine design and other scientific purposes where protein sequences are important.

As shown in **Figure 4**, most real-world HIV-1 genome sequences stayed at about the same distance from the 2021 consensuses as they did in 2002. This is explained by the relatively small overall difference between the 2002 and 2021 consensuses as compared to the distances to the real-world genome sequences, i.e., at about 1.1% consensus-to-consensus distance and about 5% consensus-to-real sequence distance. The principle of this is shown in **Figure 2**. The differences were, however, uneven across many aspects of the data. On the genome level, *env* had most differences because it (mostly the variable loop regions) evolves faster than other parts of the genome. Moreover, for certain uses, a 1% overall genome difference is meaningless because a specific amino acid at a certain site may make all the difference. On the subtype/CRF level, some subtype/CRF consensus sequences changed more than others, ranging from 0.8 to 10% (**Table 1**), e.g., while CRF01 only changed nucleotide state at 49 sites when going from building consensus sequences based on nine sequences in 2002 to 350 in 2021, subtype H consensus sequences differed at 222 sites going from 3 to 10 underlying sequences.

Consensus sequences are computational constructs rather than real world biological entities. As such, consensus sequences may not exist in nature, yet it has been shown that they may describe stable and representative protein structures (Sternke et al., 2019) that may be suitable for vaccines (Novitsky et al., 2002; Nickle et al., 2007). Furthermore, consensus sequences are affected by potential sampling biases. In our case, worldwide HIV-1 genome sequences have not been randomly sampled, instead they are simply all sequences ever published in the international literature, for whatever purpose. Nevertheless, the new HIV-1 subtype/CRF consensus sequences in this study were based on up to 1,294 observed genome sequences each, and by now most geographical regions of the world have had subtype/CRF surveys, all which contributed near full genome sequences included in these new consensus sequences. Here, 89 countries were included among these sequences. Some subtypes/CRFs have had limited geographical spread, and thus sampling, which is not the same as unrepresentative sampling, is uneven due to the epidemiological dynamics. Two other potential reasons for change from 2002 until 2021 is more use of antiviral drugs in some parts of the world, and changes in sequencing technologies. Recall, however, that the 2021 consensuses include all high-quality sequences, including those used in 2002. Thus, overall, the 2021 consensus sequences likely are good representations of the typical subtype/CRF genome nucleotide states.

Alternatives to consensus sequences include phylogenetically inferred ancestral sequences (Thornton, 2004), the most frequently observed actual sequence in a population, the most central real sequence in a population, and so-called mosaic sequences (Thurmond et al., 2008). Each one of these alternatives are also computational constructs that depend on assumptions related to sampling and evolutionary processes. They may

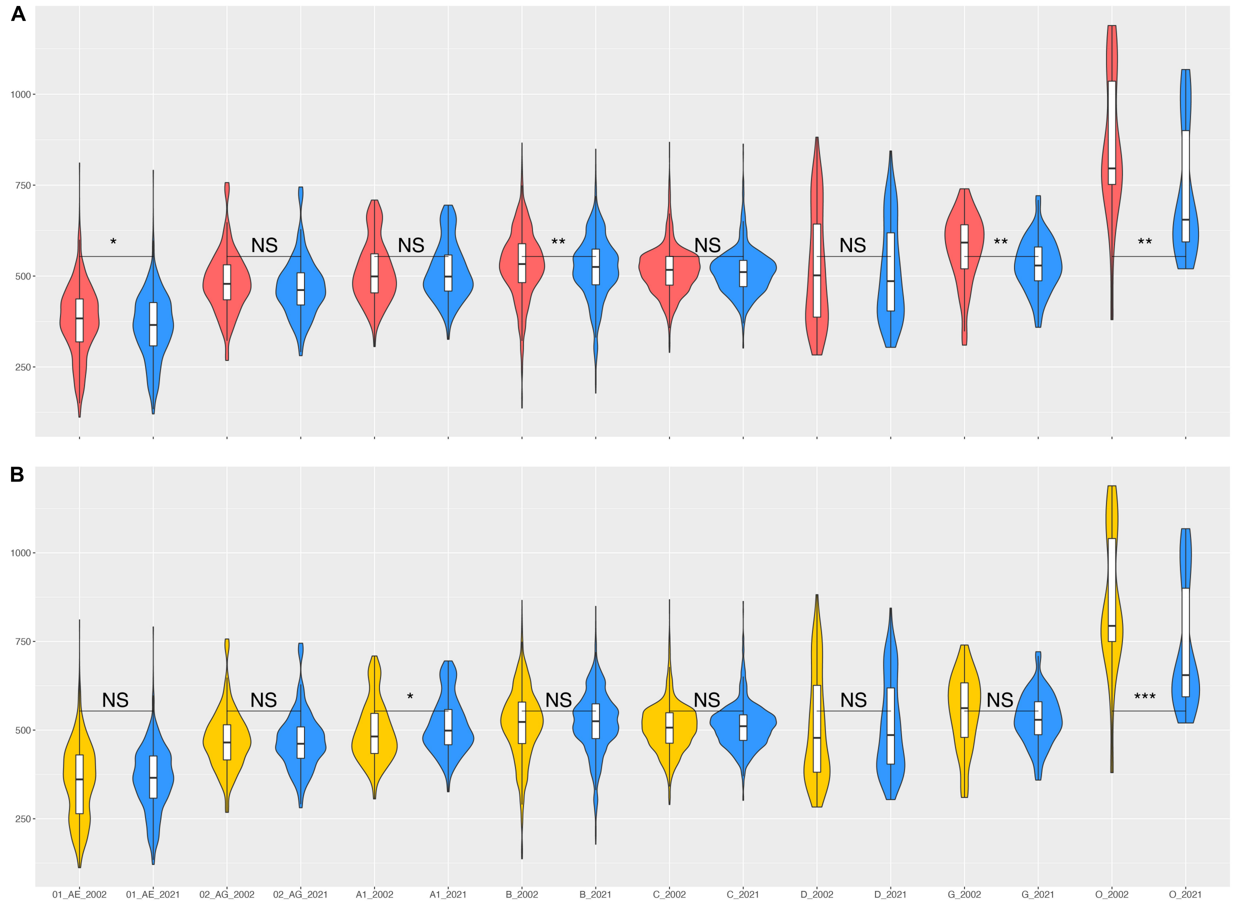


**FIGURE 4 |** Nucleotide state differences between HIV-1 consensus sequences and individual HIV-1 genomes from across the world. **(A)** Violin plots of the distribution of nucleotide state differences between individual HIV-1 sequences sampled up until 2002 and the 2002 consensuses (red) and individual HIV-1 sequences sampled up until 2021 and the 2021 consensuses (blue). **(B)** Violin plots of the distribution of nucleotide state differences between individual HIV-1 sequences sampled up until 2002 and the 2021 consensuses (yellow) and, again, individual HIV-1 sequences sampled up until 2021 and the 2021 consensuses (blue). Violin plot margins show the distribution of possible values, box margins 25% (Q1) and 75% (Q3) quantiles (*IQR*), box whiskers indicate *Q*1 $- 1.5 \times$ *IQR* and Q3 1.5 × *IQR*, the median is depicted by a horizontal line. Pairwise comparisons of the distributions show significance assessed by a two-sided Wilcoxon rank sum test with Bonferroni multiple-test correction ($p = \alpha/m$, with $\alpha = 0.05$ (*), $\alpha = 0.01$ (**), $\alpha = 0.001$(***), and NS = not significant, for $m = 16$ tests).

each have their strengths and limitations in whatever use they are put to.

## CONCLUSION

In conclusion, with the large increase of available full genome sequences from across the world, the 2021 consensus sequences likely are good representations of the typical subtype/CRF genome nucleotide states. The new consensus sequences are available at the LANL HIV database for public use.

## DATA AVAILABILITY STATEMENT

All new HIV-1 consensus sequences calculated in this study are available at https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html under the Alignment type "Consensus/Ancestral" type, Year "2021".

## AUTHOR CONTRIBUTIONS

GL and TL conceived and designed the study. GL, BF, and TL analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by NIH/NIAID grant R01AI087520.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.828765/full#supplementary-material

## REFERENCES

Domingo, E., Garcia-Crespo, C., and Perales, C. (2021). Historical perspective on the discovery of the quasispecies concept. *Annu. Rev. Virol.* 8, 51–72. doi: 10.1146/annurev-virology-091919-105900

Foley, B., Marie, K. B. T., Kenneth, L. T., Beatrice, A. C. H., Ilene, M., James, M., et al. (2018). *HIV Sequence Compendium 2018*. Los Alamos, NM: Los Alamos National Laboratory.

Frith, M. C., Mitsuhashi, S., and Katoh, K. (2021). lamassemble: multiple alignment and consensus sequence of long reads. *Methods Mol. Biol.* 2231, 135–145. doi: 10.1007/978-1-0716-1036-7_9

Gao, F., Weaver, E. A., Lu, Z., Li, Y., Liao, H. X., Ma, B., et al. (2005). Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. *J. Virol.* 79, 1154–1163. doi: 10.1128/JVI.79.2.1154-1163.2005

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect Dis.* 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020). Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781. doi: 10.1016/S2352-3018(20)30252-6

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kulikova, A. V., Diaz, D. J., Loy, J. M., Ellington, A. D., and Wilke, C. O. (2021). Learning the local landscape of protein structures with convolutional neural networks. *J. Biol. Phys.* 47, 435–454. doi: 10.1007/s10867-021-09593-6

Lee, C. (2003). Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics* 19, 999–1008. doi: 10.1093/bioinformatics/btg109

Nickle, D. C., Rolland, M., Jensen, M. A., Pond, S. L., Deng, W., Seligman, M., et al. (2007). Coping with viral diversity in HIV vaccine design. *PLoS Comput. Biol.* 3:e75. doi: 10.1371/journal.pcbi.0030075

Novitsky, V., Smith, U. R., Gilbert, P., McLane, M. F., Chigwedere, P., Williamson, C., et al. (2002). Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J. Virol.* 76, 5435–5451. doi: 10.1128/jvi.76.11.5435-5451.2002

R Development Core Team (2003). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rose, P. P., and Korber, B. T. (2000). Detecting hypermutations in viral sequences with an emphasis on G –> A hypermutation. *Bioinformatics* 16, 400–401. doi: 10.1093/bioinformatics/16.4.400

Seah, A., Lim, M. C. W., McAloose, D., Prost, S., and Seimon, T. A. (2020). MinION-Based DNA barcoding of preserved and non-invasively collected wildlife samples. *Genes (Basel)* 11:445. doi: 10.3390/genes11040445

Sternke, M., Tripp, K. W., and Barrick, D. (2019). Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U.S.A.* 116, 11275–11284. doi: 10.1073/pnas.1816707116

Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* 5, 366–375. doi: 10.1038/nrg1324

Thurmond, J., Yoon, H., Kuiken, C., Yusim, K., Perkins, S., Theiler, J., et al. (2008). Web-based design and evaluation of T-cell vaccine candidates. *Bioinformatics* 24, 1639–1640. doi: 10.1093/bioinformatics/btn251

UNAIDS (2021). *Global HIV Statistics, Fact Sheet*. Available online at: https://www.unaids.org/en/resources/fact-sheet (accessed November 29, 2021).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Yan, J., Yoon, H., Kumar, S., Ramanathan, M. P., Corbitt, N., Kutzler, M., et al. (2007). Enhanced cellular immune responses elicited by an engineered HIV-1 subtype B consensus-based envelope DNA vaccine. *Mol. Ther.* 15, 411–421. doi: 10.1038/sj.mt.6300036

Zhang, D., Zhang, T., Liu, S., Sun, D., Ding, S., Cheng, X., et al. (2021). SARS2020: an integrated platform for identification of novel coronavirus by a consensus sequence-function model. *Bioinformatics* 37, 1182–1183. doi: 10.1093/bioinformatics/btaa767

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 09 February 2022
doi: 10.3389/fmicb.2021.812391

# Antiretroviral Imprints and Genomic Plasticity of HIV-1 *pol* in Non-clade B: Implications for Treatment

***Jude S. Bimela[1,2,3], Aubin J. Nanfack[4,5], Pengpeng Yang[6], Shaoxing Dai[6], Xiang-Peng Kong[7], Judith N. Torimiro[5,8†] and Ralf Duerr[1,9*†]***

*[1] Department of Pathology, New York University School of Medicine, New York, NY, United States, [2] Department of Biochemistry, University of Yaoundé 1, Yaoundé, Cameroon, [3] Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, United States, [4] Medical Diagnostic Center, Yaoundé, Cameroon, [5] Chantal Biya International Reference Centre for Research on HIV/AIDS Prevention and Management (CIRCB), Yaoundé, Cameroon, [6] Yunnan Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, China, [7] Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, United States, [8] Department of Biochemistry, Faculty of Medicine and Biomedical Sciences, University of Yaoundé 1, Yaoundé, Cameroon, [9] Department of Microbiology, New York University School of Medicine, New York, NY, United States*

Combinational antiretroviral therapy (cART) is the most effective tool to prevent and control HIV-1 infection without an effective vaccine. However, HIV-1 drug resistance mutations (DRMs) and naturally occurring polymorphisms (NOPs) can abrogate cART efficacy. Here, we aimed to characterize the HIV-1 *pol* mutation landscape in Cameroon, where highly diverse HIV clades circulate, and identify novel treatment-associated mutations that can potentially affect cART efficacy. More than 8,000 functional Cameroonian HIV-1 *pol* sequences from 1987 to 2020 were studied for DRMs and NOPs. Site-specific amino acid frequencies and quaternary structural features were determined and compared between periods before (≤2003) and after (2004–2020) regional implementation of cART. cART usage in Cameroon induced deep mutation imprints in reverse transcriptase (RT) and to a lower extent in protease (PR) and integrase (IN), according to their relative usage. In the predominant circulating recombinant form (CRF) 02_AG (CRF02_AG), 27 canonical DRMs and 29 NOPs significantly increased or decreased in RT during cART scale-up, whereas in IN, no DRM and only seven NOPs significantly changed. The profound genomic imprints and higher prevalence of DRMs in RT compared to PR and IN mirror the dominant use of reverse transcriptase inhibitors (RTIs) in sub-Saharan Africa and the predominantly integrase strand transfer inhibitor (InSTI)-naïve study population. Our results support the potential of InSTIs for antiretroviral treatment in Cameroon; however, close surveillance of IN mutations will be required to identify emerging resistance patterns, as observed in RT and PR. Population-wide genomic analyses help reveal the presence of selective pressures and viral adaptation processes to guide strategies to bypass resistance and reinstate effective treatment.

**Keywords: HIV-1 polymerase (pol), non-clade B drug resistance mutations (DRMs), naturally occurring polymorphisms (NOPs), CRF02_AG, antiretroviral imprints, genomic plasticity, treatment intensification in Cameroon, reverse transcriptase inhibitors (RTI) versus integrase strand transfer inhibitors (INSTI)**

OPEN ACCESS

***Edited by:***
*Taisuke Izumi,*
*University of the Sciences,*
*United States*

***Reviewed by:***
*Yuta Hikichi,*
*National Cancer Institute at Frederick,*
*United States*
*George William Carnell,*
*University of Cambridge,*
*United Kingdom*

****Correspondence:***
*Ralf Duerr*
*ralf.duerr@nyulangone.org*

*†These authors share last authorship*

***Specialty section:***
*This article was submitted to*
*Virology,*
*a section of the journal*
*Frontiers in Microbiology*

***Received:*** *10 November 2021*
***Accepted:*** *27 December 2021*
***Published:*** *09 February 2022*

***Citation:***
*Bimela JS, Nanfack AJ, Yang P, Dai S, Kong X-P, Torimiro JN and Duerr R (2022) Antiretroviral Imprints and Genomic Plasticity of HIV-1 pol in Non-clade B: Implications for Treatment.*
*Front. Microbiol. 12:812391.*
*doi: 10.3389/fmicb.2021.812391*

## INTRODUCTION

Combinational antiretroviral therapy (cART) has significantly slowed the AIDS pandemic and reduced the incidence of HIV infections (UNAIDS, 2020). However, treatment intensification exerts selective pressure and drives viral adaptation resulting in the emergence of HIV-1 drug resistance mutations (DRMs). The selection of drug-resistant viruses is based on HIV's high mutation rate, followed by the generation of a genetically diverse pool of HIV quasispecies in each patient, and the selection and outgrowth of the fittest variants, adapted to the given condition (Feder et al., 2021). Under antiretroviral treatment, there is a strong purifying selection for drug-resistant viruses, which can be traced by viral genomic analyses of the drug target's respective genomic regions, primarily found in the *pol* region (Feder et al., 2021). Drug resistance mutations represent a major barrier to effective therapy (Hamers et al., 2018). People infected with HIV-1 who have acquired DRMs are less likely to achieve viral suppression, are more likely to discontinue treatment, acquire new DRMs, and experience virological failure or death (WHO, 2019). Also, naturally occurring polymorphisms (NOPs) can modulate the magnitude of drug resistance, paving the way to developing DRMs, or have intrinsic resistance potential themselves (Wainberg and Brenner, 2012). NOPs are most abundant but least studied in regions like West and Central Africa where highly diverse non-B clade viruses circulate (Wainberg and Brenner, 2012; Hemelaar et al., 2020). Non-clade B infections make up most HIV-1 infections worldwide, with genetic variation among clades up to 35% (Hemelaar et al., 2020). Naturally occurring polymorphisms, their diversity and plasticity, cART-associated viral adaptations, and the growing proportion of DRMs are areas of intense clinical interest and still unfolding. Comprehensive, population-wide studies are required to leverage a deeper understanding of these mutations and NOPs, including effective prevention and treatment strategies. We provided such an analysis for Cameroon, a West-Central African country with a population of ∼27 million and ∼3.7% HIV prevalence in 2021, and we considered all functional, non-clonal reverse transcriptase (RT), protease (PR), and integrase (IN) sequences that have been deposited to the LANL database (8,130 out of ∼10,600 submitted sequences in total) (**Figure 1**).

cART programs in sub-Saharan Africa are based on WHO guidelines that have traditionally recommended reverse transcriptase and protease inhibitors (RTIs, PIs) for first- and second-line therapy, and only recently have begun transitioning to integrase strand transfer inhibitor (InSTI) usage in some countries (UNAIDS, 2017b). Effective access to cART in Cameroon started in 2004 following the WHO/UNAIDS "3 by 5" initiative to provide three million people living with HIV (PLHIV) in low and middle-income countries with life-prolonging cART by the end of 2005 (WHO, 2003). Before 2004, only a few generic antiretroviral drugs [lamivudine (3TC), zidovudine (ZDV), stavudine (d4T), and nevirapine (NVP)] were available at a very low scale in main cities such as the capital Yaoundé (Bourgeois et al., 2005; Aghokeng et al., 2011). Until May 2007, treated patients had to pay for cART (US $23–$100 monthly), laboratory tests (US $58–$85 per viral load assay and $19–$27 per CD4 cell count), and physician's consultation ($1.5–$15), thus limiting the number of people accessing cART (Laurent et al., 2006; Boyer et al., 2009). cART usage in Cameroon expanded in June 2007 with free access to cART for eligible patients dependent on CD4 count-based guidelines, and since 2016 to all HIV-infected individuals following the implementation of the "test and treat" UNAIDS's initiative (WHO, 2005; National AIDS Control Committee, 2015). The scale-up of cART considerably improved the lives of PLHIV globally, including sub-Saharan Africa, but West-Central Africa remains far behind in all three categories of the UNAIDS 90-90-90 targets, i.e., diagnosis, treatment, and viral suppression (UNAIDS, 2017a). Furthermore, the implementation of cART has accelerated the development of DRMs. The overall prevalence of drug-resistant strains in cART-naïve and treated individuals has increased dramatically, from 2.2 and 40.7% in 2010, reaching > 10% and 60% in recent years, respectively (Nanfack et al., 2015, 2017; WHO, 2019). Cameroon and other West-Central African countries have mainly used a triple cocktail of two nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) and one non-NRTIs (NNRTIs) as first-line treatment supported by PIs in the second line (WHO, 2015). Although targeting the same protein (RT), NRTIs and NNRTIs exert differential drug pressure, which results in complementary DRM profiles (**Supplementary Table 1**). For example, RT mutations M184V/I and T215F/Y are linked with NRTIs (e.g., lamivudine, emtricitabine, zidovudine, or stavudine) and K103N/S with NNRTIs (e.g., nevirapine or efavirenz). The intensive use of InSTIs (Boyer et al., 2009; Landman et al., 2014), particularly dolutegravir (DTG), effectively started in 2020, whereas in preceding years, it was distributed in 10 treatment centers only out of more than 160 functional treatment units. InSTIs potently suppress viral load in diverse HIV-1 clade infections, yet subtype-specific differences in efficacy and acquisition of DRMs exist (Wainberg and Brenner, 2012), as has become known for other antiretroviral drug classes (Theys et al., 2019; Stanford University HIV Drug Resistance Database, 2020).

While computational methods have been developed for RTIs, PIs, and more recently, for InSTIs, to quantify the genetic barrier to the acquisition of DRMs (Gotte, 2012; Theys et al., 2019), population-based time-series analyses are needed to show the actual adaptive changes to the viral genomic landscape in countries under treatment pressure. Here, we present such a study for Cameroon, a country undergoing progressive cART scale-up since 2004 and known as the epicenter of the HIV disease where an immense diversity of HIV strains exists (Hahn et al., 2000; Hemelaar et al., 2020). Among a high number of subtypes, circulating, and unique recombinant forms (CRFs, URFs), the globally most abundant recombinant form, CRF02_AG, is predominant (Nanfack et al., 2017; Hemelaar et al., 2020). To understand the emergence and impact of NOPs and DRMs, we use computational and structural methods to compare HIV-1 *pol* genomic plasticity in RT, PR, and IN regions and dissect the cross-sectional changes in HIV-1 strains in Cameroon over time.

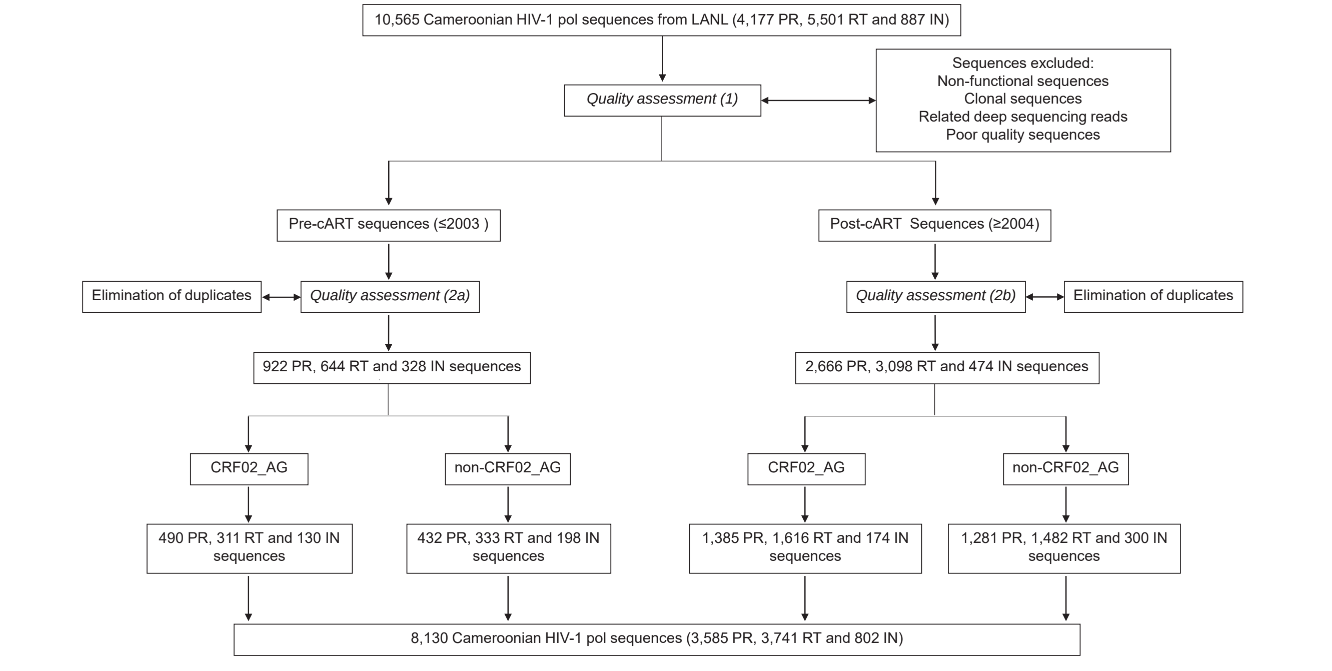

10,565 Cameroonian HIV-1 pol sequences from LANL (4,177 PR, 5,501 RT and 887 IN)
Quality assessment (1)
Sequences excluded:
Non-functional sequences
Clonal sequences
Related deep sequencing reads
Poor quality sequences
Pre-cART sequences (≤2003 )
Post-cART Sequences (≥2004)
Elimination of duplicates
Quality assessment (2a)
Quality assessment (2b)
Elimination of duplicates
922 PR, 644 RT and 328 IN sequences
2,666 PR, 3,098 RT and 474 IN sequences
CRF02_AG
non-CRF02_AG
CRF02_AG
non-CRF02_AG
490 PR, 311 RT and 130 IN sequences
432 PR, 333 RT and 198 IN sequences
1,385 PR, 1,616 RT and 174 IN sequences
1,281 PR, 1,482 RT and 300 IN sequences
8,130 Cameroonian HIV-1 pol sequences (3,585 PR, 3,741 RT and 802 IN)

**FIGURE 1 |** Study flow diagram. Summary of sequence filtering and quality assessment of Cameroonian RT, PR, and IN sequences.

# MATERIALS AND METHODS

## Study Design and Sample Selection

In an ecological analysis, 8,130 HIV-1 RT (3,741), PR (3,585), and IN (802) nucleotide sequences from Cameroonian PLHIV were downloaded from the Los Alamos National Laboratory (LANL) HIV sequence database and studied after a multi-step quality check and selection process to eliminate non-functional, clonal, and poor quality sequences (**Figure 1** and **Supplementary Methods**). Ethical or IRB clearance was not required given the ecological study design with sequences downloaded from public databases.

## Structural Modeling and Structural Stability Prediction

Structural modeling was done using UCSF Chimera v1.13.1 (Pettersen et al., 2004), ICM-Pro (Molsoft) (Abagyan et al., 1994), and SWISS-MODEL server (Bertoni et al., 2017) to determine the potential impact of the quaternary structure and charge distribution on the presence/emergence of polymorphisms, differences in drug binding, selective pressure, and resistance. In addition, the Cartesian_ddg application (Park et al., 2016; Leman et al., 2020) from Rosetta version 2020.28.61328 was used to predict the effects of mutations on the stability of protein/drug complexes, the latter known to potentially affect drug binding and resistance (Barouch-Bentov and Sauer, 2011). The $\Delta\Delta G$ scores were estimated as differences in mean scores for ten independent runs for every mutant and wild-type protein-drug complex structure (**Supplementary Methods**).

## Role of the Funding Source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

## Sample Size Calculations and Statistical Analysis

Mann-Whitney and Kruskal Wallis tests were done to compare mutational frequencies among RT, PR, and IN sequences, which achieved $\geq$ 93.8% power to detect a 0.5 standard deviation (SD) between groups of $\geq$ 99 sequences (5% error). Statistical analyses and power calculations were done using the R/RStudio ggplot2 package (RStudio Team, 2015), Excel (2016), G*Power v.3.1.9.4 (Erdfelder et al., 1996), and GraphPad Prism 8.0 (La Jolla California United States) (**Supplementary Methods**). The threshold for significance was $P < 0.05$. Chord diagrams were generated using the circlize, and correlograms using the corrplot and RColorBrewer packages in program R/RStudio.

# RESULTS

## CRF02_AG Predominance in HIV-1 Reverse Transcriptase, Protease, and Integrase Sequences From Cameroon Pre- and Post-initiation of Combinational Antiretroviral Therapy

Cross-sectional genetic and phylogenetic analyses showed that overall, the HIV-1 clade distribution remained comparable over

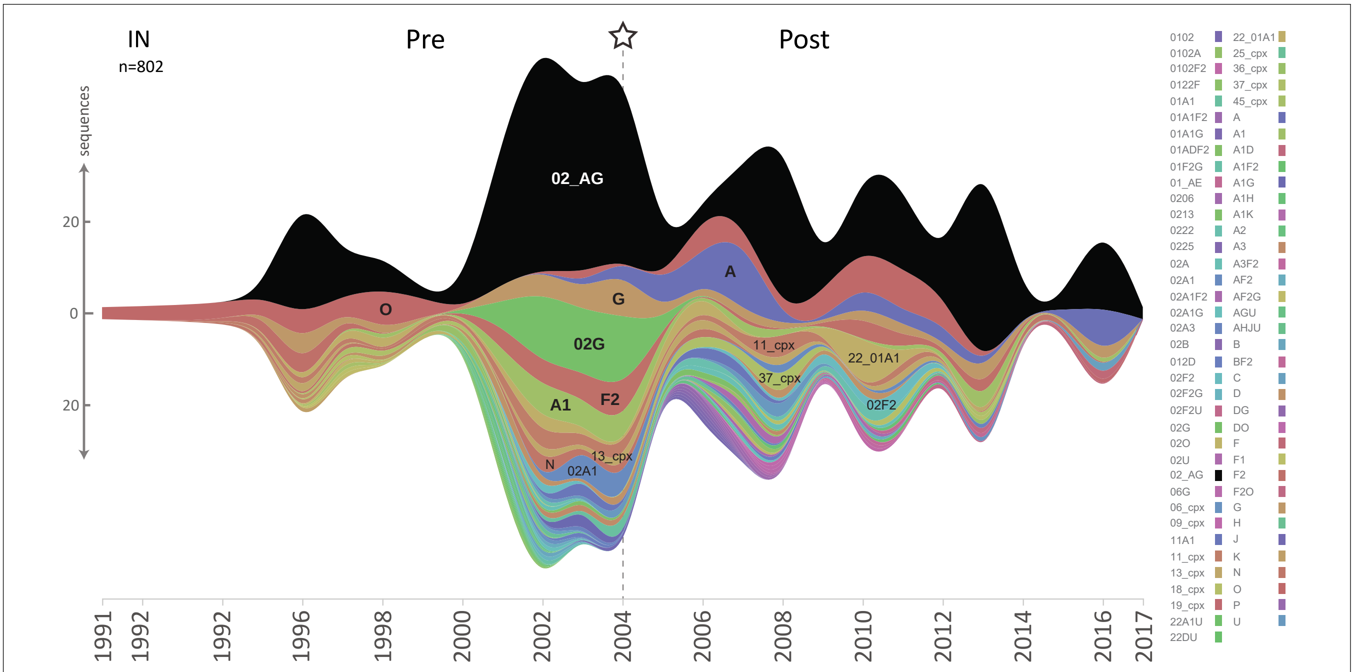


**FIGURE 2 |** HIV-1 lineage distribution in Cameroon according to HIV-1 *pol* IN sequences from 1991 until 2020. Stream graph of lineage distribution of Cameroonian HIV-1 *pol* IN sequences (y-axis). All available IN full-length sequences (HxB2 position bp 4,230–5,093, *n* = 802) from the LANL database are shown, as of November 3rd, 2020, after excluding non-functional, poor-quality, duplicate, and clonal sequences. HIV-1 subtypes, recombinant forms, and groups are color-coded according to the legend to the right, and the most prevalent lineages are also annotated in the graph. The star and dashed line indicate the year when combinational antiretroviral treatment was implemented in Cameroon.

time in the studied RT, PR, and IN *pol* regions, with CRF02_AG being the predominant clade (**Figures 2**, **3** and **Supplementary Figures 1–3**). However, distinct fluctuations in clade frequencies over time were observed, attributable to sampling disparities, evolutionary, epidemiological, and virological factors. Therefore, to avoid confounding factors by clade differences between study time points, we focused for the most part on CRF02_AG in our statistical analysis.

## The Changing *pol* Mutational Landscape Post-implementation of Combinational Antiretroviral Therapy in Cameroon

A comparison of RT, PR, and IN mutation frequencies in Cameroonian CRF02_AG sequences, using respective CRF02_AG consensus sequences as a reference, revealed slightly higher baseline (years $\leq$ 2003) mutation frequencies in RT (mean 3.15%; 13.88 mutations/440 residues in RT) than in PR (2.83%; 2.80/99) and IN (2.07%; 5.95/288). Moreover, when these numbers were compared with the mutation frequencies in the post-cART era (2004–2020), we observed a strongly significant increase in RT ($P < 0.0001$), in contrast to only moderate changes in PR ($P = 0.07$) and IN ($P = 0.04$) (**Figure 3C**, **Supplementary Figure 4**, and **Supplementary Table 1**). In addition, most amino acid sites ($\geq$60%) in RT (262/440) and PR (61/99) underwent an increase in mutation levels, whereas in IN, a large proportion remained unchanged (143/288; 50%) (**Supplementary Figure 4**). As a result, the most significant increase in mutations per site between the pre- and post-cART era occurred in RT followed by PR and IN regions, which portrays their relative importance as target proteins regarding the treatment protocols from between 2004 and 2020 (**Figure 3C** and **Supplementary Figure 4**).

## Site-Specific Mutational Analyses in Reverse Transcriptase, Protease, and Integrase Reveal the Greatest Drug Resistance Mutation Plasticity in Reverse Transcriptase and Protease

To assess Cameroon's treatment-associated mutational changes in more detail, we calculated site-specific amino acid abundances for every position in RT, PR, and IN. We distinguished canonical DRM sites from all remaining sites, and among the latter, we screened all variable (NOP) sites for significantly increasing or decreasing mutations, called treatment-associated mutations (**Figures 4**, **5** and **Supplementary Table 1**). Focusing on clade CRF02_AG, more than half of the known DRM sites in RT (27/41) exhibited a significant change (26 increasing, 1 decreasing) in mutation frequencies post-cART, of which eleven were strongly significant (FDR $< 0.01$). Only eight (five with FDR $< 0.01$) of the 33 known DRM sites, exhibited a significant change in PR. Most pronounced increases ($>$10%) were recorded for RT DRMs M184X ($\Delta$ = 22.16%), K103X ($\Delta$ = 15.16%), and T215X ($\Delta$ = 10.17%), and PR DRM L63X ($\Delta$ = 15.18%). In contrast to RT and PR, none of the 24 known DRM sites in

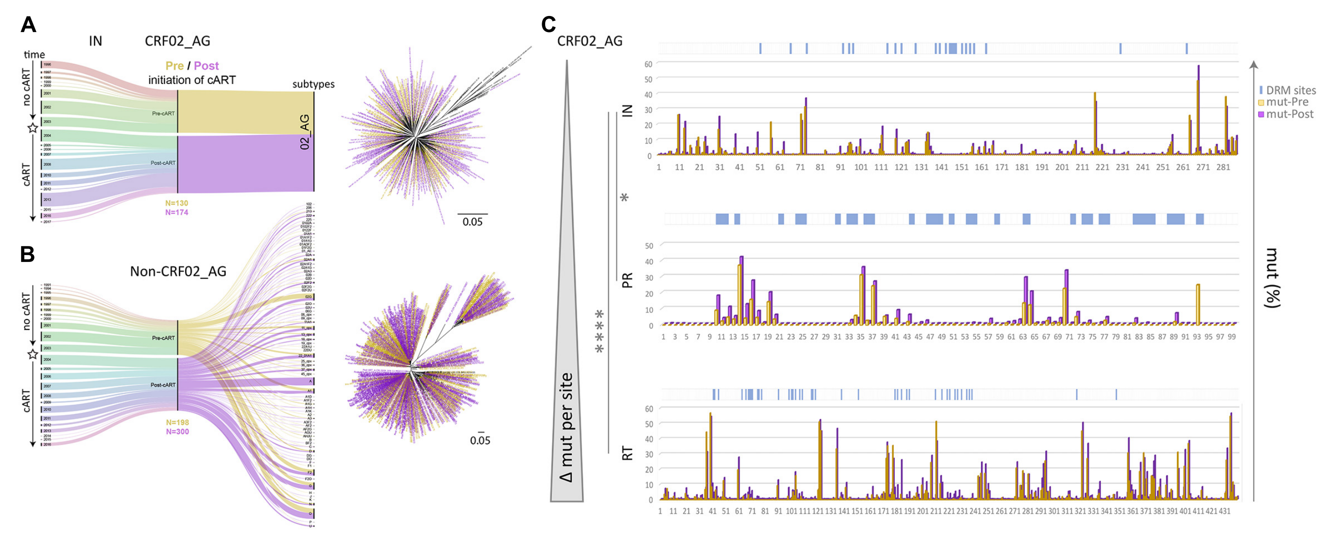


**FIGURE 3 |** Data segregation, phylogenetic analysis, and comparison of drug resistance mutations and naturally occurring polymorphisms in Cameroonian HIV-1 *pol* IN, PR, and RT sequences before and after regional implementation of cART. **(A,B)** HIV-1 IN sequences from Cameroonian HIV-1-infected individuals (as in **Figure 2**) were segregated into CRF02_AG **(A)** and non-CRF02_AG data sets **(B)**. The data set composition by sampling year and lineage/subtype is summarized in alluvial diagrams. Asterisks indicate the subcategorization of sequences collected pre- and post-implementation of cART in Cameroon (2004) along the timeline; sequences are colored yellow (Pre) and purple (Post), respectively. Pre and Post sample numbers are indicated below the plots. Phylogenetic placement of Pre and Post sequences is shown in maximum-likelihood RAxML trees with the same yellow/purple color code. The scale indicates a 5% genetic distance. **(C)** Comparison of site-specific frequencies of mutations (mut) including naturally occurring polymorphisms (NOP) in Cameroonian HIV-1 *pol* IN, PR, and RT sequences before and after regional implementation of cART. CRF02_AG consensus sequences (derived from the pre-cART data sets) served as references to summarize all amino acid mutations per site. Their relative frequencies (%) are compared side-by-side for Pre and Post data sets in bar graphs. Locations of canonical DRMs (according to the Stanford drug resistance database, November 2020) are indicated with blue ticks on top of the charts. The bar charts are sorted from top to bottom according to increasing mutational difference from Pre to Post per site. Stars indicate statistical differences in a Kruskal Wallis test with Dunn's multiplicity correction (* < 0.05, **** < 0.0001).

IN showed a significant change. Among the silent IN mutation landscape, only three accessory mutations (L74X, Q95X, and G163X) increased slightly post-cART (Δ = 5.2%, 3.7%, and 2.8%, respectively). These mutations do not affect the second-generation InSTIs dolutegravir and bictegravir and have little effect (reduced susceptibility) on the first-generation InSTIs elvitegravir and raltegravir (**Figure 4**).

## Emerging Treatment-Associated Mutations in Reverse Transcriptase and Protease as an Imprint of the Regionally Dominant Combinational Antiretroviral Therapy Regimens

To determine the impact of cART scale-up on the HIV-1 NOP landscape and the emergence of treatment-associated mutations, we summarized all significantly changing NOPs over time in RT, PR, and IN (**Figure 5** and **Supplementary Figures 5–7**). Comparable to the patterns of emerging canonical DRMs, we observed the greatest plasticity of novel, treatment-associated mutations in RT (29 significant sites: 20 increasing, 9 decreasing), followed by PR (13 sites: 11 increasing, 2 decreasing), and IN (7 sites: 5 increasing, 2 decreasing). Strongly significant changes (FDR < 0.01) were only observed in RT (6x) and PR (7x). The strongest increases (>10%) were at V135X (Δ = 12.82%), I292X (Δ = 15.56%), and T377X (Δ = 14.96%) in RT, and G16X (Δ = 11.05%) and K70X (Δ = 10.44%) in PR. In IN, no site reached a 10% increase in treatment-associated mutations, M50X (Δ = 9.80%) being closest.

## Linked Emergence of Major Nucleotide Reverse Transcriptase Inhibitor and Non-nucleotide Reverse Transcriptase Inhibitor Mutations in Cameroonian HIV-1 CRF02_AG Viruses

To study whether there was a linkage between the presence of certain mutations, we performed a correlation analysis focusing on mutations that significantly increased from the pre- to post-cART period in Cameroon (**Figure 6** and **Supplementary Figure 8**). In line with the higher total and relative number of mutations in RT vs. PR and also post-cART vs. pre-cART, linked mutations were most frequent in RT sequences post implementation of cART (RT: 152 vs. 24, PR: 17 vs. 4 when comparing post-cART with pre-cART periods, respectively). Furthermore, most of these mutations were positively correlated (RT post: 135, pre: 22; PR: post: 13, pre: 4), whereas a smaller part was inversely correlated (RT, post: 17, pre: 2; PR: post:4, pre:0). Of interest, the three RT mutations that most prominently increased after the regional implementation of cART (>10%) strongly correlated with each other despite being related with different drug classes, i.e., NRTIs (M184X, T215X) and NNRTIs (K103X).

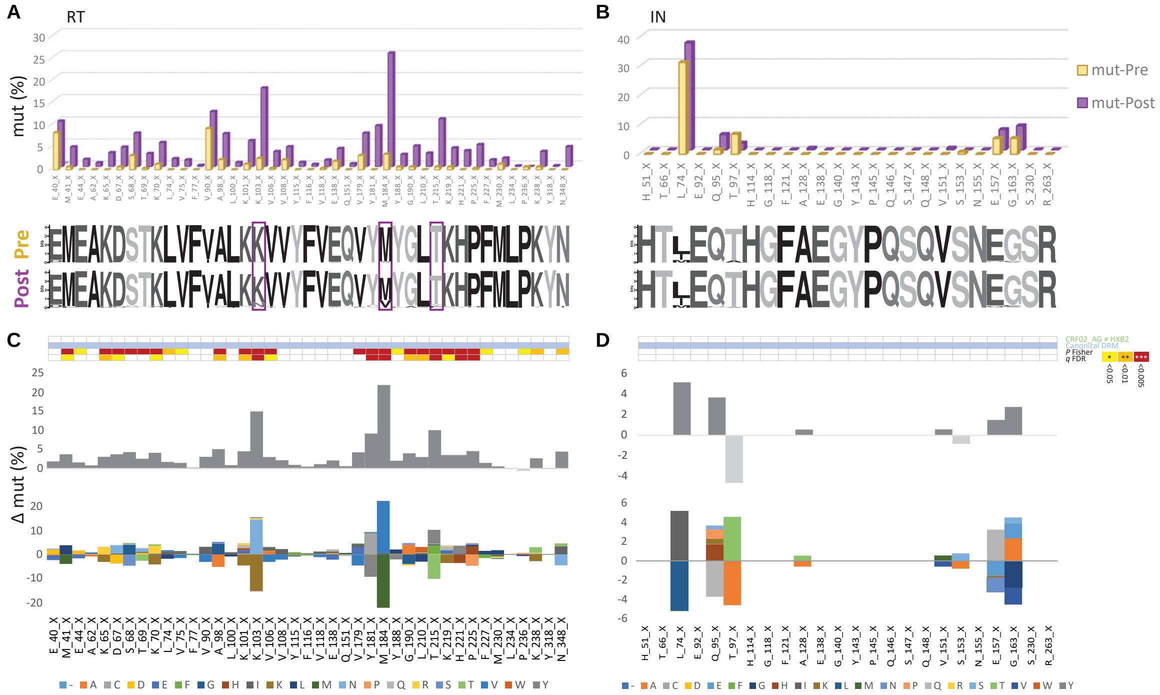


**FIGURE 4 |** Canonical drug resistance mutations of HIV-1 RT and IN sequences from Cameroon before and after regional implementation of cART. **(A,B)** Comparison of site-specific frequencies at canonical drug resistance mutation (mut) sites in Cameroonian HIV-1 *pol* RT (left) and IN sequences (right) before (yellow, Pre) and after (purple, Post) regional implementation of cART. CRF02_AG consensus sequences (derived from pre-cART data sets) served as references to call mut variants per site. The dominant (consensus) amino acid is indicated for each site, followed by the position in RT. X indicates any mutation/minority variant. Below the bar chart, weblogos of amino acid occurrences per site are shown for both Pre and Post data sets. Sites at which DRM frequencies increased by more than 10% from pre- to post-cART period are boxed. **(C,D)** Same selection of all canonical DRM sites in RT and IN [as in **(A,B)**, referring to blue annotations in **Figure 3**]. On the y-axis, the difference in mut percentage (Δ mut) between Post and Pre is indicated for each site, with increasing mut frequencies from Pre to Post shown as positive values (dark gray bars) and decreasing frequencies shown as negative values (light gray bars). The mirror bar chart below indicates all amino acid (aa) changes according to the bottom's aa color code. The 4-row color strips on top indicate differences between CRF02_AG consensus sequences and HxB2 (green), sites of canonical drug resistance mut (DRM) sites (blue), and statistically significant differences between Pre and Post in Fisher Exact tests (*P* values) and false discovery rates (FDR, *q* values), according to the legend to the right.

This is presumably due to the NRTI/NNRTI combinational treatment protocol that has been in place in Cameroon for most of the study time, suggesting that in most patients NRTIs and NNRTIs exerted drug pressure at the same time. Besides the mutations mentioned above, M41X, D67X, K70X, G190X, and K219X were involved in strongly linked RT mutation clusters. The linkage of mutations was substantially weaker in PR. It involved linkages between I15X and I64X, G16X and K70X, K41X and L63X, and L63X and M89X. The linkage of a few RT mutations in the pre-cART period is indicative for either an introduction of drug-resistant viruses into Cameroon, a low-level circulation of NRTIs/NNRTIs and exertion of drug pressure even before 2004, and/or a mutual impact and combined effects of mutations on genomic stability or protein function (epistasis).

## Plasticity of Reverse Transcriptase and Protease Mutations in Non-CRF02_AG Viruses

Besides CRF02_AG, there has traditionally been a broad diversity of HIV-1 subtypes circulating in Cameroon (**Figure 2** and **Supplementary Figures 1, 2**). Under the caveat of subtype-specific differences and a slightly different subtype coverage in pre- and post-cART periods, we analyzed whether common structural or sequence patterns of mutations crystallized in Non-CRF02_AG viruses after cART was implemented in Cameroon (**Figure 7** and **Supplementary Table 1**). Whereas decreasing mutations (gray) were scattered across large parts of the RT and PR protein structures, presumably due to subtype bias between the study periods, the few increasing mutations in RT and PR (purple) had a clustered appearance around the drug-binding sites. In addition to emerging treatment-associated mutations (**Figure 7A**, purple), there was a high number of canonical drug resistance mutations enriched among the bulk of Non-CRF02_AG sequences, which mainly applied to RT (**Figure 7B**, blue), but not for IN (**Figure 7C** left, all non-significant).

## Reverse Transcriptase and Protease Treatment-Associated Mutations Exhibit Differential Structural Patterns and Effects on Protein Stability

In the quaternary proteins, most canonical DRMs in RT and PR cluster around the drug-binding sites, as shown for the significantly increasing DRMs post-cART in CRF02_AG

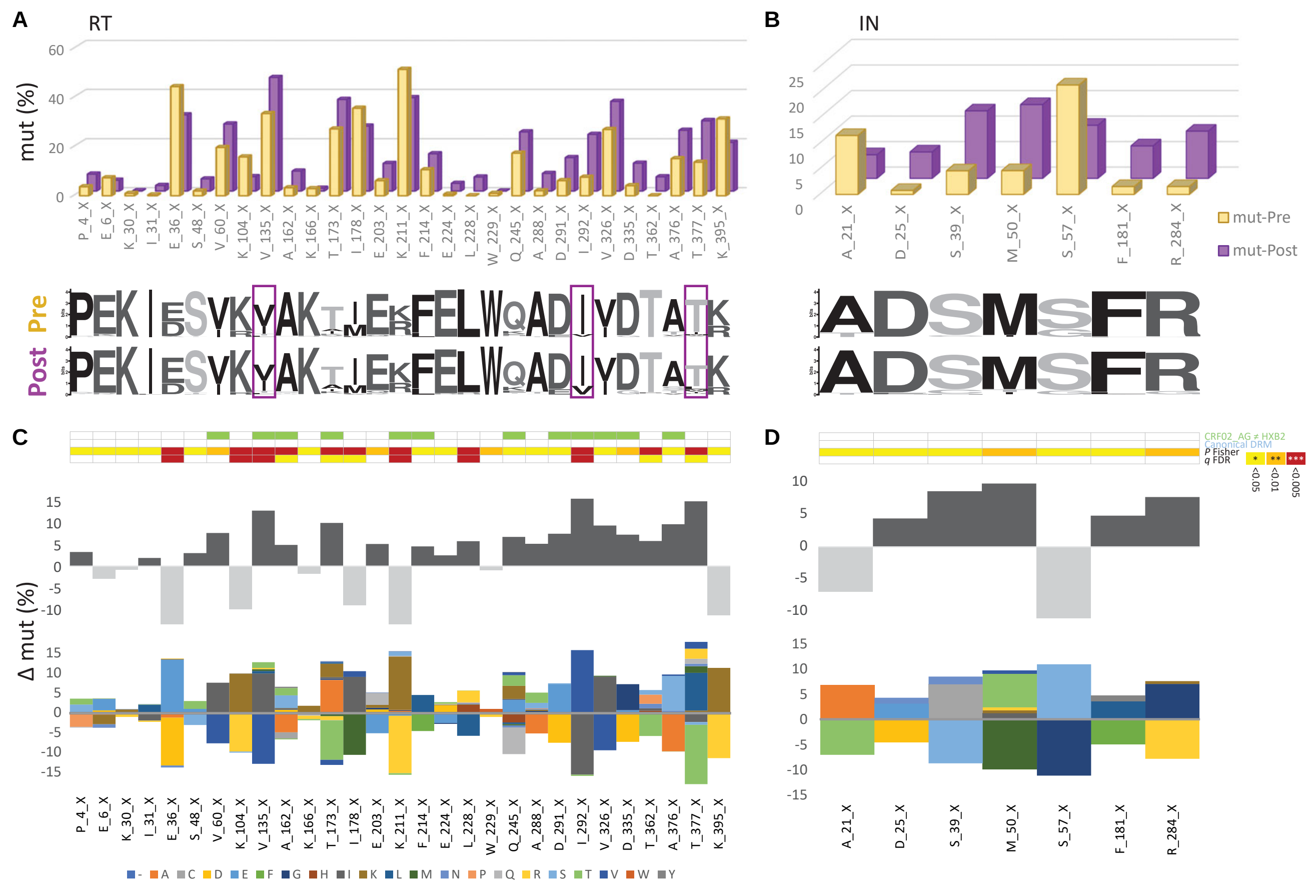


**FIGURE 5 |** The emergence of treatment-associated mutations in RT, but not in IN of Cameroonian HIV-1 CRF02_AG sequences during years of cART scale-up. **(A,B)** Comparison of site-specific frequencies in HIV-1 *pol* RT (left) and IN sequences (right) before (yellow, Pre) and after (purple, Post) regional implementation of cART. All sites are shown that have not been linked with canonical drug resistance and that exhibit changes in mutation (mut) frequencies from to Pre to Post yielding *P* values < 0.05 [see also **(C,D)**]. CRF02_AG consensus sequences (derived from pre-cART data sets) served as references to call mut variants per site. The dominant (consensus) amino acid is indicated for each site, followed by the position in RT. X indicates any mutation/minority variant. Below the bar chart, weblogos of amino acid occurrences per site are indicated from both Pre and Post data sets. Sites at which mutation frequencies increased by more than 10% from the pre- to post-cART period are boxed. **(C,D)** Same selection of RT and IN sites as in **(A,B)**. On the y-axis, the difference in mut percentage (Δ mut) between Post and Pre is indicated for each site, with increasing mut frequencies from Pre to Post shown as positive values (dark gray bars) and decreasing frequencies shown as negative values (light gray bars). The mirror bar chart below indicates all amino acid (aa) changes according to the aa color code at the bottom. The 4-row color strips on top indicate differences between CRF02_AG consensus sequences and HxB2 (green), sites of canonical drug resistance mut (DRM) sites (blue), and statistically significant differences between Pre and Post in Fisher Exact tests (*P* values) and false discovery rates (FDR, *q* values), according to the legend to the right.

(**Figure 8A** and **Supplementary Figure 6A**). The comparative study of emerging treatment-associated mutations revealed a slightly more widespread distribution within the proteins, but still in proximity to the drug-binding region; for example, in RT, restricted to the protein half where drug binding occurs (**Figure 8A** and **Supplementary Figure 6B**). Both canonical DRMs and treatment-associated mutations in RT and PR can have stabilizing, neutral, or destabilizing effects. Strikingly, > 10 RT treatment-associated mutations have strong destabilizing effects, which are known as potential mechanisms of drug resistance (**Figure 8B**, **Supplementary Figure 6C**, and **Supplementary Tables 2–4**) (Barouch-Bentov and Sauer, 2011). In contrast to RT and PR mutations, the less significant IN treatment-associated mutations appear randomly spread across the IN monomer, possibly due to the lower/absent drug pressure exerted by InSTIs in the studied years, and have hardly any destabilizing effect (**Supplementary Figure 7**).

## A Time-Series Analysis of *pol* Mutations Reveals Differential Mutational Plasticities

Having determined significant changes of DRMs and novel, treatment-associated mutation in Cameroonian RT and PR sequences post-cART, we aimed to assess the cross-sectional mutation profiles over time (**Figure 8C** and **Supplementary Figures 9–12**). As a result, we observed differential mutation profiles that included subsequently increasing/decreasing mutations, plateauing mutations, and transient mutation peaks or nadirs. The mutations were mostly based on replacements

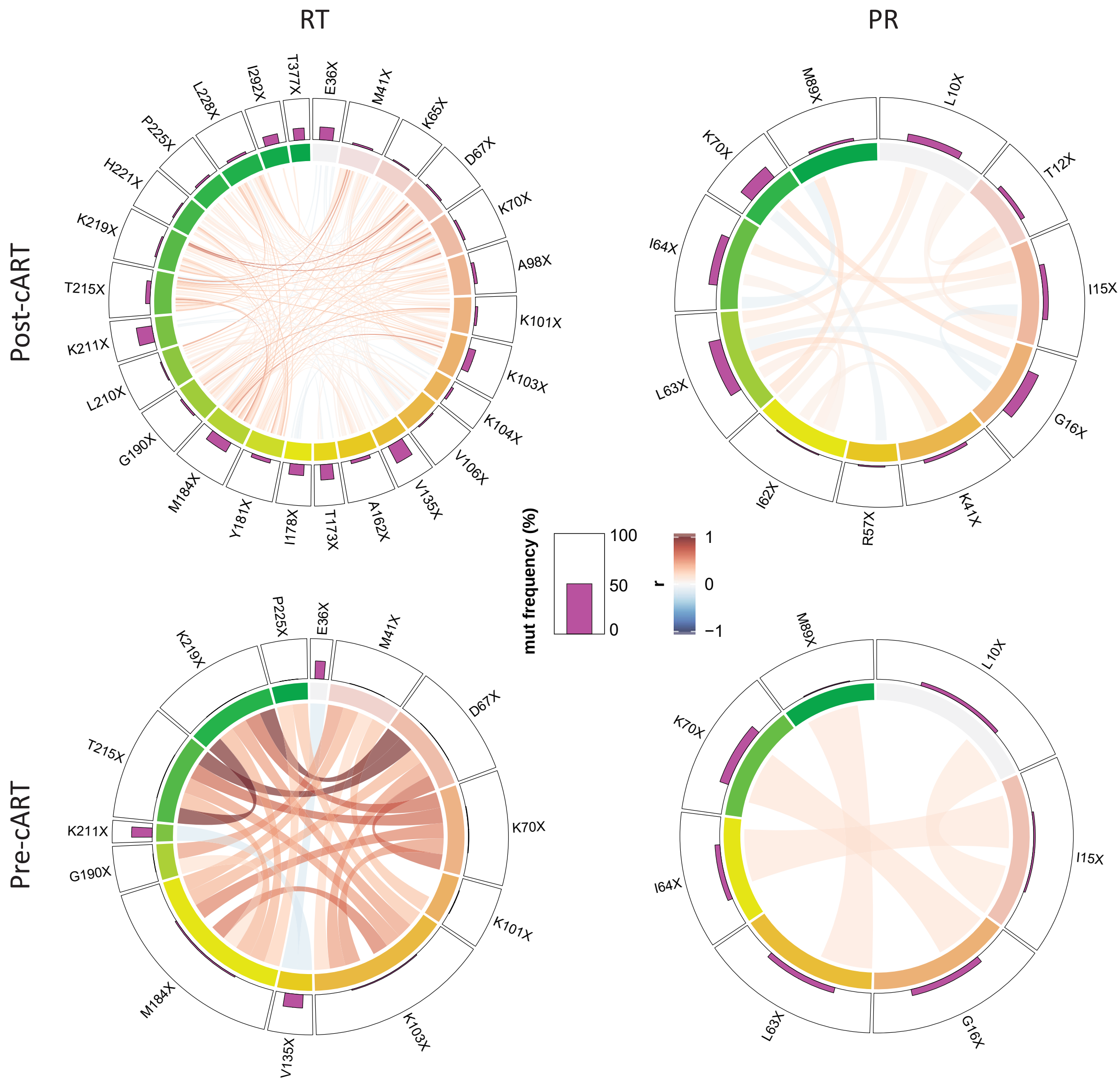


**FIGURE 6 |** Linked emergence of mutations in CRF02_AG *pol* RT and PR. Chord diagram illustrating the network of linear correlations among mutations in RT and PR that changed significantly (increase or decrease) from the period before (pre-cART) to after (post-cART) implementation of combinational antiretroviral treatment in Cameroon (2004). The bar plot on the outer track displays the mutation frequencies of the mutations in the respective period. Pairwise correlations are shown as chords between connected variables, i.e., mutations, in the center of the plot. Chords are color-coded according to the magnitude of the correlation coefficient (r); chord width inversely corresponds to the P-value. Two-tailed Spearman rank tests were performed and P values were adjusted for multiple comparisons using the Benjamini-Hochberg method. Among the full set of linear correlations (see **Supplementary Figure 8**), only significant links/chords are shown, and mutations without a significant link ($P < 0.05$) to another mutation were removed.

by one distinct amino acid (**Figures 4**, **5** and **Supplementary Figure 5**); however, a few sites exhibited a broader range of mutations with fluctuating amino acid frequencies, as evident for RT treatment-associated mutation T362X (**Figure 8C**).

## DISCUSSION

This study provides a thorough overview of HIV *pol* mutations and genotypic drug resistance in an entire population. In an ecological analysis of pooled viral genomic data from Cameroon where HIV prevalence stands at ∼3.7% in 2021, we used computational and structural methods to assess the genomic plasticity of HIV-1 *pol* over time and its implication on treatment. Treatment exerts selective pressure on the swarm of viruses present in every patient and the entire population and drives viral adaptation resulting in the emergence of HIV-1 DRMs. The cross-sectional nature of our study on a representative set of 8,130 sequences from Cameroon (all LANL-deposited RT, PR, and IN sequences that are functional and non-clonal) enabled us to

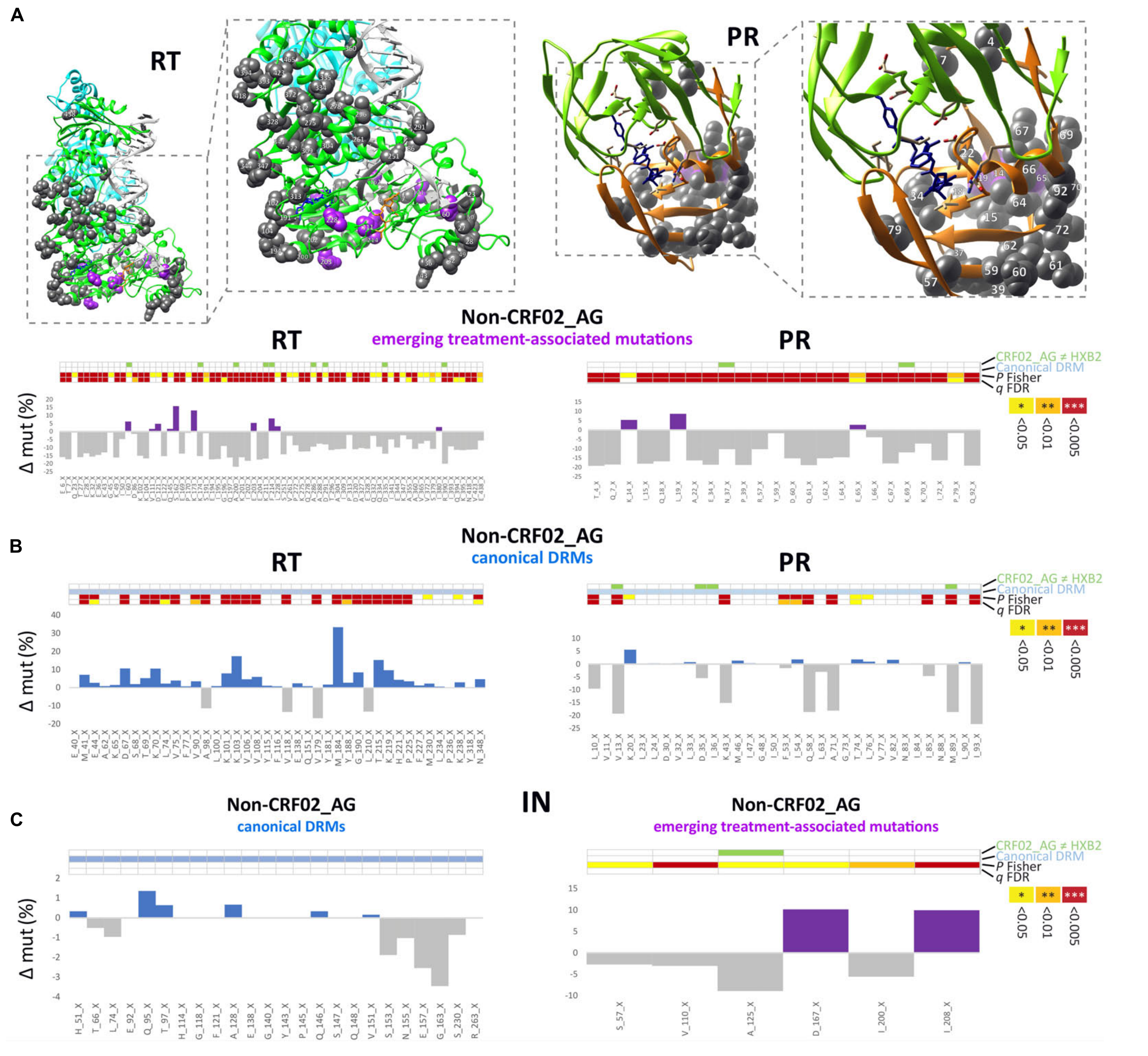


**FIGURE 7 |** Structural and statistical analysis of emerging mutations in Non-CRF02_AG. **(A)** Sites of significantly increasing or decreasing (P Fisher < 0.05) treatment-associated mutations in RT (left) and PR (right) are projected onto RT and PR structures. Calculations of site-specific mutation increases/decreases in Non-CRF02_AG between pre- and post-cART periods are shown below. Detailed views of the drug-binding regions with annotated aa sites are shown in boxes to the right. Treatment-associated mutation residues are displayed as magenta or gray spheres, according to a significant increase or decrease from pre- to post-cART, respectively ($P < 0.05$). **(B)** Calculation of site-specific mutation differences in Non-CRF02_AG for canonical DRM sites. Bars are displayed in blue or gray according to a significant increase or decrease from pre- to post-cART, respectively ($P < 0.05$). **(C)** Calculation of site-specific mutation differences in Non-CRF02_AG for the IN region, both for canonical DRM sites (left) and emerging treatment-associated mutations (right). The 4-row color strips indicate differences between CRF02_AG consensus sequences and HXB2 (green), sites of canonical drug resistance mut (DRM) sites (blue), and statistically significant mutation differences between pre- and post-cART periods in Fisher Exact tests (P values) and false discovery rates (FDR, q values), according to the legend to the right.

compare sufficient numbers of mutations and NOPs statistically over time. Fluctuating sample numbers and the inability to follow and assess individuals longitudinally were limitations. Our study revealed high plasticity in HIV-1 *pol* on the population level, which appears to be profoundly shaped by regionally applied cART protocols. Effective cART in Cameroon started in 2004. Hereafter, substantial scale-ups occurred in 2007 when cART became free of charge for eligible patients (CD4 count < 200 cells/mm$^3$) and with the "test and treat" initiative implemented in 2016. Cameroon's HIV treatment guidelines have primarily relied on regimens with two NRTIs and one NNRTI, which resulted in an increase in pre-treatment NRTI/NNRTI DRM rates up to >10%, and in patients failing first or second-line cART up to >60% (Nanfack et al., 2015, 2017; WHO,

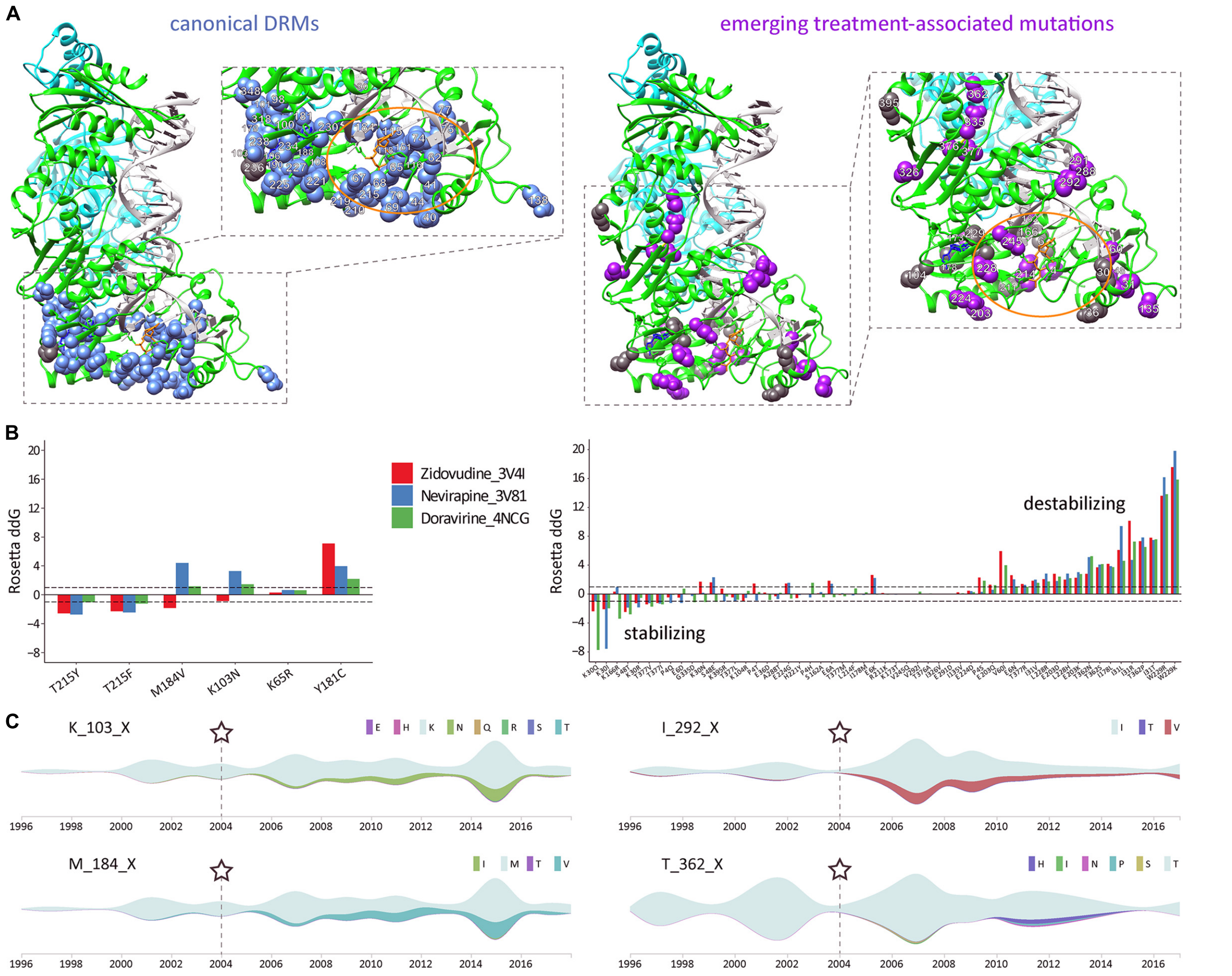


**FIGURE 8 |** Structural and time-series analysis of DRMs and emerging treatment-associated mutations in CRF02_AG *pol* RT. **(A)** Sites of significantly increasing canonical drug resistance mutations (DRMs) (left) and emerging treatment-associated mutations (right), as identified in **Figures 4**, **5**, are projected onto a complex RT structure. Detailed views of the drug-binding regions with annotated aa sites are shown in boxes to the right, and an orange oval highlights the active center. The RT models were generated using crystal structures of RT with DNA and the NRTI AZT-TP (PDB 3V4I) and RT with DNA and the NNRTI nevirapine (PDB 3V81). The nevirapine (blue) and AZT-TP (orange) molecules were placed together for illustration purposes (using structural overlay in Chimera). DRM residues are displayed as blue or gray spheres, according to a significant increase or decrease from pre- to post-cART periods, respectively ($P < 0.05$, according to **Figure 4**). Accordingly, treatment-associated mutation residues are displayed as magenta or gray spheres, according to a significant increase or decrease from pre- to post-cART, respectively ($P < 0.05$, according to **Figure 5**). **(B)** The effect of selected canonical DRMs (most prevalent in Cameroon and/or highest mutational scoring) and all significantly emerging CRF02_AG treatment-associated mutations on three different published RT protein structures were analyzed with the Cartesian ddg application (Rosetta). ddG values > 1 and < 1 are characteristic for destabilizing and stabilizing mutations, respectively. Mutations are listed from left to right according to increasing destabilizing effects. **(C)** Time-series analysis of significantly increasing mutations in HIV-1 CRF02_AG *pol* RT during cART scale-up in Cameroon. Streamgraphs in silhouette mode display mutations among the studied sequence on the y-axis along the timeline on the x-axis. The gray-green color indicates the absence of mutations (and the presence of the dominant aa residue). According to the legend to the right, other colors indicate the presence of mutations/minority variants. The RT aa site and its dominant/consensus aa are indicated to the left of each streamgraph. Shown is a selection of two canonical drug resistance mutation sites (left) and two emerging treatment-associated mutation sites (right) with a significant increase in mutations over time. Asterisks and dashed lines mark the time point of cART implementation in Cameroon.

2019), jeopardizing the success of national HIV management. Nevirapine was included in most first-line regimens in Cameroon between 2004 and 2016 and was replaced in 2016 by efavirenz in the preferential first-line regimen in resource-constrained settings (tenofovir disoproxil fumarate/lamivudine/efavirenz) (National AIDS Control Committee, 2015; WHO, 2015). The long-term use of NRTIs and NNRTIs explains the high proportion of RTI DRMs observed in our study. Starting from low levels of RT/PR-associated DRMs in the cART-naïve population (≤2004), a significant increase occurred at multiple protein residues post-implementation of cART. RT mutations M184X, K103X, and T215X rose most significantly, both in CRF02_AG (**Figure 4**) and non-CRF02_AG (**Figure 7** and **Supplementary Table 1**), which is in line with recent

DRM monitoring studies in Cameroon and worldwide (Nanfack et al., 2015, 2017; WHO, 2019). These mutations strongly affect the efficacy of NRTIs lamivudine/emtricitabine (M184V/I) and zidovudine/stavudine (T215F/Y), and NNRTIs nevirapine/efavirenz (K103N/S). Notably, these mutations were significantly linked among the study sequences (**Figure 6**), implying that combinational NRTI/NNRTI treatment exerts simultaneous selection pressure and induces mutations in different RT regions at comparable rates depending on the applied drugs. The functional consequence of the co-existence of drug resistance mutations remains to be studied in further detail. Based on the recent introduction of InSTIs in Cameroon, the absence of DRMs to InSTIs was expected.

Antiretroviral drugs, including PIs, NRTIs/NNRTIs, and InSTIs, have mainly been tested for efficacy against subtype B viruses. Although most of these drugs are expected to act on diverse subtypes, genetic differences can impact drug resistance pathways or kinetics of DRM development (Theys et al., 2019). Although only marginally assessed, reduced susceptibility to antiretroviral drugs has been described in CRF02_AG infections (Wainberg and Brenner, 2012). Reduced PI susceptibility at baseline has been associated with subsequent virological failure on lopinavir/ritonavir monotherapy in antiretroviral-naïve patients harboring CRF02_AG viruses (Sutherland et al., 2015). In the DAYANA trial, a simplified regimen of tenofovir plus lopinavir/ritonavir used for early treatment attained poor viral suppression (defined as HIV-1 RNA viral load $\geq$ 100 copies/ml between weeks 24 and 96) in a CRF02_AG study population (Landman et al., 2014). In IN, the G118R mutation was associated with NOPs at codons 74 and 118 in clades C and CRF02_AG and conferred resistance against raltegravir and DTG (Malet et al., 2011; Brenner et al., 2016). In addition to canonical DRMs, we found emergent CRF02_AG polymorphisms accumulating during cART scale-up. NOPs have been reported to alter or impair susceptibility to RTIs, PIs, and even InSTIs (Wainberg and Brenner, 2012). Comparable to the observed DRM imprints upon cART in Cameroon, the NOP landscape changed most strikingly in RT, followed by PR and IN. It implies that resistance issues might be involved besides compensatory mutations and evolutionary/epidemiological trends. Protein destabilization is a potential drug resistance mechanism (Barouch-Bentov and Sauer, 2011; Chang and Torbett, 2011; Sheik Amamuddy et al., 2018); however, as our DRM data show, there is no strict association with resistance, and only a subset of DRMs exert destabilizing effects. Consequently, emerging treatment-associated mutations showed a differential pattern. Notably, a set of >10 emerging RT treatment-associated mutations had strong destabilizing potential. For example, RT T362X mutations significantly increased post-cART, and both T362N/S were shown to be strongly destabilizing (**Figures 5**, **6**). T362 is located at the DNA-interacting RT connection domain, and mutations have been associated with NNRTI treatment, possibly affecting RNAse H activity (Julias et al., 2003). In the IN region, we observed a high baseline occurrence of low-resistance mutation L64M, known as a specific feature of CRF02_AG (Rhee et al., 2003). L64M further increased post-cART, though not significantly.

The scarcity of available HIV-1 IN sequences from Cameroon after the roll-out of DTG-based regimens in Africa (2017) limits our sequence-based conclusions on InSTI treatment in Cameroon. Between 2017 and November 2020, only three complete and two fragmented IN sequences have been deposited to LANL. The latter two are from the same HIV-1 CRF18_cpx-infected individual who developed multi-drug InSTI resistance in 2017 based on G140A, Q148R, and E157Q DRMs, which further aggravated in 2019 additionally involving E138K/Q and S147G (Fokam et al., 2020). Our structural simulations suggest that CRF02_AG IN has a comparable quaternary structure with clade B, including surface and drug-binding site charge distribution (**Supplementary Figures 13, 14**). Due to DTG's high genetic barrier to resistance, the emergence of DRMs in treatment-naïve patients is extremely rare, which renders DTG highly effective across clades (Rhee et al., 2019). Increased usage of InSTIs including the long-acting cabotegravir for maintenance therapy and pre-exposure prophylaxis is anticipated (Swindells et al., 2020). Emerging IN drug resistance is an imminent threat requiring broad and widespread monitoring efforts across Cameroon and beyond (Inzaule et al., 2018; Lubke et al., 2019). Particular attention should be paid to patients that previously used Raltegravir-containing regimens as in the case of DTG and Darunavir/r multi-drug resistance in HIV-1 CRF18_cpx infection described recently in Cameroon (Fokam et al., 2020). Furthermore, recent data suggest a reduced InSTI efficacy in patients with DRMs in RT (Siedner et al., 2020), stressing the importance of studying full *pol.* Even beyond *pol*, there have been reports of InSTI resistance caused by five mutations in the *nef* region (Malet et al., 2017). In addition, mutations in *gag* have been shown to contribute to PI resistance, and findings suggest a tight interdependency between Gag structural proteins and the protease during the development of PI resistance (Clavel and Mammano, 2010; Codoñer et al., 2017; Datir et al., 2020). Most recently, mutations in the envelope glycoprotein (Env) have been associated with resistance to antiretrovirals, including the InSTI DTG (Van Duyne et al., 2019; Hikichi et al., 2021). Consequently, future studies should aim at analyzing DRMs and NOPs over the full genome and in the context of clinical drug resistance data to detect drug resistance across drug classes more comprehensively and inform clinical management more accurately. The low number of full-genome CRF02_AG sequences from Cameroon (61 deposited in the LANL database) and the lack of associated clinical data impeded such analyses for Cameroon at this time. In summary, the better characterization of DRMs and NOPs including changes over time in the context of diverse HIV-1 clades and applied cART regimens is paramount to understanding the underlying mechanisms of emerging and evolving antiretroviral drug resistance. The association between genomic imprints and usage of antiretroviral drugs targeting the respective *pol* regions underline the ability of ecological analyses to reveal viral adaptation processes and underlying purifying selection pressures, to eventually guide clinical management and targeted therapies. Our results support the applicability of InSTIs for cART in Cameroon but stress the necessity of tight surveillance of DRMs when increasing drug pressure is exerted. The current study focused on CRF02_AG and the comparison of pre- and post-cART periods in Cameroon. Future studies will need to intensify longitudinal characterizations involving different clades and regions of the world under consideration of the applied treatment protocols to dissect HIV-1 *pol* and full-genome adaptations and their clinical implications on a global scale. Phenotypic testing of emerging treatment-associated mutations in different clades with or without co-occurring DRMs will reveal their direct and indirect impact on drug resistance.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

JNT and RD conceived research goals and analyses. JSB, PY, SD, X-PK, and RD performed formal analyses. RD acquired funding for the project. PY, SD, X-PK, and RD developed methodologies, designed and/or implemented computer codes. AJN, JNT, and RD supervised the research activity and validated the research output. JSB and RD verified the underlying data and prepared figures and tables. JSB, AJN, and RD wrote the manuscript. All authors reviewed and edited the manuscript.

## FUNDING

This study was supported by the Fogarty International Center (FIC), National Institutes of Health (NIH) grant TW009604, and by the Merck Investigator Studies Program (MISP) grant #57483. The authors declare that this study received funding from the Merck Investigator Studies Program (MISP), grant #57483. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## ACKNOWLEDGMENTS

We thank Yicheng Guo and Zizhang Sheng for their assistance in structural analyses and Flavia Camacho for administrative assistance. We further wish to acknowledge the software support of New York University's Data Services, Bobst Library, and, in particular, the expertise shared by Patrick Chester and Senior Academic Technology Specialist Denis Rubin.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.812391/full#supplementary-material

## REFERENCES

Abagyan, R., Totrov, M., and Kuznetsov, D. (1994). ICM - a new method for protein modeling and design - applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* 15:488.

Aghokeng, A. F., Kouanfack, C., Laurent, C., Ebong, E., Atem-Tambe, A., Butel, C., et al. (2011). Scale-up of antiretroviral treatment in sub-Saharan Africa is accompanied by increasing HIV-1 drug resistance mutations in drug-naive patients. *AIDS* 25, 2183–2188. doi: 10.1097/QAD.0b013e32834bbbe9

Barouch-Bentov, R., and Sauer, K. (2011). Mechanisms of drug resistance in kinases. *Expert Opin. Investig. Drugs* 20, 153–208.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* 7:10480. doi: 10.1038/s41598-017-09654-8

Bourgeois, A., Laurent, C., Mougnutou, R., Nkoué, N., Lactuock, B., Ciaffi, L., et al. (2005). Field assessment of generic antiretroviral drugs: a prospective cohort study in Cameroon. *Antiviral Ther.* 10, 335–341.

Boyer, S., Marcellin, F., Ongolo-Zogo, P., Abega, S., Nantchouang, R., Spire, B., et al. (2009). Financial barriers to HIV treatment in yaounde, cameroon: first results of a national cross-sectional survey. *Bull. World Health Organ.* 87, 279–287. doi: 10.2471/blt.07.049643

Brenner, B. G., Thomas, R., Blanco, J. L., Ibanescu, R., Oliveira, M., Mesplède, T., et al. (2016). Development of a G118R mutation in HIV-1 integrase following a switch to dolutegravir monotherapy leading to cross-resistance to integrase inhibitors. *J. Antimicrob. Chemother.* 71, 1948–1953. doi: 10.1093/jac/dkw071

Chang, M. W., and Torbett, B. E. (2011). Accessory mutations maintain stability in drug-resistant HIV-1 protease. *J. Mol. Biol.* 410, 756–760. doi: 10.1016/j.jmb.2011.03.038

Clavel, F., and Mammano, F. (2010). Role of gag in HIV resistance to protease inhibitors. *Viruses* 2, 1411–1426. doi: 10.3390/v2071411

Codoñer, F. M., Peña, R., Blanch-Lombarte, O., Jimenez-Moyano, E., Pino, M., Vollbrecht, T., et al. (2017). Gag-protease coevolution analyses define novel structural surfaces in the HIV-1 matrix and capsid involved in resistance to Protease Inhibitors. *Sci. Rep.* 7:3717. doi: 10.1038/s41598-017-03260-4

Datir, R. P., Kemp, S., El Bouzidi, K., Mlcochova, P., Goldstein, R. A., Breuer, J., et al. (2020). In vivo emergence of a novel protease inhibitor resistance signature in HIV-1 matrix. *mBio* 11, e2036–e2020. doi: 10.1128/mBio.02036-20

Erdfelder, E., Faul, F., and Buchner, A. (1996). GPOWER: a general power analysis program. *Behav. Res. Meth. Instrum.* 28, 1–11.

Feder, A. F., Harper, K. N., Brumme, C. J., and Pennings, P. S. (2021). Understanding patterns of HIV multi-drug resistance through models of temporal and spatial drug heterogeneity. *eLife* 10:e69032. doi: 10.7554/eLife.69032

Fokam, J., Takou, D., Semengue, E. N. J., Teto, G., Beloumou, G., Dambaya, B., et al. (2020). First case of dolutegravir and darunavir/r multi drug-resistant hiv-1 in cameroon following exposure to raltegravir: lessons and implications in the era of transition to Dolutegravir-based regimens. *Antimicrob. Resist. Infect. Control.* 9:143. doi: 10.1186/s13756-020-00799-2

Gotte, M. (2012). The distinct contributions of fitness and genetic barrier to the development of antiviral drug resistance. *Curr. Opin. Virol.* 2, 644–650. doi: 10.1016/j.coviro.2012.08.004

Hahn, B. H., Shaw, G. M., De Cock, K. M., and Sharp, P. M. (2000). AIDS as a zoonosis: scientific and public health implications. *Science* 287, 607–614. doi: 10.1126/science.287.5453.607

Hamers, R. L., Rinke de Wit, T. F., and Holmes, C. B. (2018). HIV drug resistance in low-income and middle-income countries. *Lancet HIV* 5, e00588–e596.

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2020). Global and regional epidemiology of HIV-1 recombinants in 1990 - 2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781. doi: 10.1016/S2352-3018(20)30252-6

Hikichi, Y., Van Duyne, R., Pham, P., Groebner, J. L., Wiegand, A., Mellors, J. W., et al. (2021). Mechanistic analysis of the broad antiretroviral resistance conferred by hiv-1 envelope glycoprotein mutations. *mBio* 12, e3134–e3120. doi: 10.1128/mBio.03134-20

Inzaule, S. C., Hamers, R. L., Noguera-Julian, M., Casadellà, M., Parera, M., Rinke de Wit, T. F., et al. (2018). Primary resistance to integrase strand transfer inhibitors in patients infected with diverse HIV-1 subtypes in sub-Saharan Africa. *J. Antimicrob. Chemother.* 73, 1167–1172. doi: 10.1093/jac/dky005

Julias, J. G., McWilliams, M. J., Sarafianos, S. G., Alvord, W. G., Arnold, E., and Hughes, S. H. (2003). Mutation of amino acids in the connection domain of human immunodeficiency virus type 1 reverse transcriptase that contact the template-primer affects RNase H activity. *J. Virol.* 77, 8548–8554. doi: 10.1128/jvi.77.15.8548-8554.2003

Landman, R., Koulla-Shiro, S., Sow, P. S., Ngolle, M., Diallo, M., Guèye, N. F. M., et al. (2014). Evaluation of four tenofovir-containing regimens as first-line treatments in cameroon and senegal: the ANRS 12115 DAYANA Trial. *Antiviral Ther.* 19, 51–59. doi: 10.3851/IMP2675

Laurent, C., Kouanfack, C., Vergne, L., Tardy, M., Zekeng, L., Noumsi, N., et al. (2006). Antiretroviral drug resistance and routine therapy, cameroon. *Emerg. Infect. Dis.* 12, 1001–1004. doi: 10.3201/eid1206.050860

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., et al. (2020). Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nat. Methods* 17, 665–680. doi: 10.1038/s41592-020-0848-2

Lubke, N., Jensen, B., Huttig, F., Feldt, T., Walker, A., Kaiser, R., et al. (2019). Failure of dolutegravir first-line art with selection of virus carrying R263K and G118R. *N. Engl. J. Med.* 381, 887–889. doi: 10.1056/NEJMc1806554

Malet, I., Fourati, S., Charpentier, C., Morand-Joubert, L., Armenia, D., Wirden, M., et al. (2011). The HIV-1 integrase G118R mutation confers raltegravir resistance to the CRF02_AG HIV-1 subtype. *J. Antimicrob. Chemother.* 66, 2827–2830. doi: 10.1093/jac/dkr389

Malet, I., Subra, F., Charpentier, C., Collin, G., Descamps, D., Calvez, V., et al. (2017). Mutations located outside the integrase gene can confer resistance to hiv-1 integrase strand transfer inhibitors. *mBio* 8, e922–e917. doi: 10.1128/mBio.00922-17

Nanfack, A. J., Agyingi, L., Noubiap, J. J., Ngai, J. N., Colizzi, V., and Nyambi, P. N. (2015). Use of amplification refractory mutation system PCR assay as a simple and effective tool to detect HIV-1 drug resistance mutations. *J. Clin. Microbiol.* 53, 1662–1671. doi: 10.1128/JCM.00114-15

Nanfack, A. J., Redd, A. D., Bimela, J. S., Ncham, G., Achem, E., Banin, A. N., et al. (2017). Multimethod longitudinal HIV drug resistance analysis in antiretroviral-therapy-naive patients. *J Clin. Microbiol.* 55, 2785–2800. doi: 10.1128/JCM.00634-17

National AIDS Control Committee (2015). *National Guideline on the Prevention and Management of HIV in Cameroon*. Available online at: https://www.childrenandaids.org/sites/default/files/2018-05/Cameroon_NatGuidelinesHIV_2015.pdf (accessed October 19, 2020).

Park, H., Bradley, P., Greisen, P. Jr., Liu, Y., Mulligan, V. K., Kim, D. E., et al. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J. Chem. Theory Comput.* 12, 6201–6212. doi: 10.1021/acs.jctc.6b00819

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Rhee, S. Y., Grant, P. M., Tzou, P. L., Barrow, G., Harrigan, P. R., Ioannidis, J. P. A., et al. (2019). A systematic review of the genetic mechanisms of dolutegravir resistance. *J. Antimicrob. Chemother.* 74, 3135–3149. doi: 10.1093/jac/dkz256

RStudio Team (2015). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc.

Sheik Amamuddy, O., Bishop, N. T., and Tastan Bishop, O. (2018). Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics. *Sci. Rep.* 8:17938. doi: 10.1038/s41598-018-36041-8

Siedner, M. J., Moorhouse, M. A., Simmons, B., Oliveira, T. D., Lessells, R., Giandhari, J., et al. (2020). Reduced efficacy of HIV-1 integrase inhibitors in patients with drug resistance mutations in reverse transcriptase. *Nat. Commun.* 11:5922. doi: 10.1038/s41467-020-19801-x

Stanford University HIV Drug Resistance Database (2020). *Mutation Prevalence According to Subtype and Treatment*. Available online at: https://hivdb.stanford.edu/cgi-bin/MutPrevBySubtypeRx.cgi (accessed October 3, 2020).

Sutherland, K. A., Ghosn, J., Gregson, J., Mbisa, J. L., Chaix, M. L., Codar, I. C., et al. (2015). HIV-1 subtype influences susceptibility and response to monotherapy with the protease inhibitor lopinavir/ritonavir. *J. Antimicrob. Chemother.* 70, 243–248. doi: 10.1093/jac/dku365

Swindells, S., Andrade-Villanueva, J. F., Richmond, G. J., Rizzardini, G., Baumgarten, A., Masiá, M., et al. (2020). Long-acting cabotegravir and rilpivirine for maintenance of HIV-1 suppression. *N. Engl. J. Med.* 382, 1112–1123. doi: 10.1056/NEJMoa1904398

Theys, K., Libin, P. J. K., Van Laethem, K., and Abecasis, A. B. (2019). An evolutionary model-based approach to quantify the genetic barrier to drug resistance in fast-evolving viruses and its application to HIV-1 subtypes and integrase inhibitors. *Antimicrob. Agents Chemother.* 63, e00539–e519. doi: 10.1128/AAC.00539-19

UNAIDS (2017a). *Ending AIDS, Progress Towards the 90-90-90 Targets. Global AIDS Update*. Geneva: UNAIDS.

UNAIDS (2017b). *New High-Quality Antiretroviral Therapy to be Launched in South Africa, Kenya and Over 90 Low-and Middle-Income Countries at Reduced Price*. Geneva: UNAIDS.

UNAIDS (2020). *UNAIDS Global AIDS Update 2020 - Seizing the Moment*. Available online at: https://www.unaids.org/sites/default/files/media_asset/2020_global-aids-report_en.pdf (accessed October 19, 2020).

Van Duyne, R., Kuo, L. S., Pham, P., Fujii, K., and Freed, E. O. (2019). Mutations in the HIV-1 envelope glycoprotein can broadly rescue blocks at multiple steps in the virus replication cycle. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9040–9049. doi: 10.1073/pnas.1820333116

Wainberg, M. A., and Brenner, B. G. (2012). The impact of HIV genetic polymorphisms and subtype differences on the occurrence of resistance to antiretroviral drugs. *Mol. Biol. Int.* 2012:256982. doi: 10.1155/2012/256982

WHO (2003). *Global Initiative to Provide Antiretroviral Therapy to 3 Million People with HIV/AIDS in Developing Countries by the End of 2005*. Available online at: http://www.who.int/3by5/publications/documents/en (accessed October 19, 2020).

WHO (2005). *Summary Country Profile for HIV/AIDS Treatment Scale-Up*. Available online at: https://www.who.int/3by5/support/june2005_cmr.pdf (accessed December 9, 2021).

WHO (2015). *Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection*. Available online at: https://apps.who.int/iris/bitstream/handle/10665/208825/9789241549684_eng.pdf?sequence=1 (accessed December, 2021).

WHO (2019). *HIV Drug Resistance Report 2019*. Available online at: https://www.who.int/hiv/pub/drugresistance/hivdr-report-2019/en/ (accessed October 2, 2020).

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design or conduct of the study.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/fmicb.2021.803190

# Incorporating Within-Host Diversity in Phylogenetic Analyses for Detecting Clusters of New HIV Diagnoses

*August Guang[1,2*†], Mark Howison[3†], Lauren Ledingham[4], Matthew D'Antuono[4], Philip A. Chan[4], Charles Lawrence[5], Casey W. Dunn[6] and Rami Kantor[4]*

[1] *Center for Computational Biology of Human Disease, Brown University, Providence, RI, United States,* [2] *Center for Computation and Visualization, Brown University, Providence, RI, United States,* [3] *Research Improving People's Lives, Providence, RI, United States,* [4] *Division of Infectious Diseases, The Alpert Medical School, Brown University, Providence, RI, United States,* [5] *Division of Applied Mathematics, Brown University, Providence, RI, United States,* [6] *Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, United States*

**Background:** Phylogenetic analyses of HIV sequences are used to detect clusters and inform public health interventions. Conventional approaches summarize within-host HIV diversity with a single consensus sequence per host of the *pol* gene, obtained from Sanger or next-generation sequencing (NGS). There is growing recognition that this approach discards potentially important information about within-host sequence variation, which can impact phylogenetic inference. However, whether alternative summary methods that incorporate intra-host variation impact phylogenetic inference of transmission network features is unknown.

**Methods:** We introduce *profile sampling*, a method to incorporate within-host NGS sequence diversity into phylogenetic HIV cluster inference. We compare this approach to Sanger- and NGS-derived *pol* and near-whole-genome consensus sequences and evaluate its potential benefits in identifying molecular clusters among all newly-HIV-diagnosed individuals over six months at the largest HIV center in Rhode Island.

**Results:** *Profile sampling* cluster inference demonstrated that within-host viral diversity impacts phylogenetic inference across individuals, and that consensus sequence approaches can obscure both magnitude and effect of these impacts. Clustering differed between Sanger- and NGS-derived consensus and *profile sampling* sequences, and across gene regions.

**Discussion:** *Profile sampling* can incorporate within-host HIV diversity captured by NGS into phylogenetic analyses. This additional information can improve robustness of cluster detection.

**Keywords: HIV, cluster inference, profile sampling, phylogenetics, next generation sequencing (NGS), near-whole-genome, consensus sequence, transmission disruption**

OPEN ACCESS

***Edited by:***
*Michael M. Thomson,*
*Instituto de Salud Carlos III (ISCIII),*
*Spain*

***Reviewed by:***
*Marcos Perez Losada,*
*George Washington University,*
*United States*
*Brittany Rife Magalis,*
*University of Florida, United States*

****Correspondence:***
*August Guang*
*august_guang@brown.edu*

*† These authors have contributed equally to this work and share first authorship*

***Specialty section:***
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *27 October 2021*
***Accepted:*** *22 December 2021*
***Published:*** *17 February 2022*

***Citation:***
*Guang A, Howison M, Ledingham L, D'Antuono M, Chan PA, Lawrence C, Dunn CW and Kantor R (2022) Incorporating Within-Host Diversity in Phylogenetic Analyses for Detecting Clusters of New HIV Diagnoses. Front. Microbiol. 12:803190. doi: 10.3389/fmicb.2021.803190*

# INTRODUCTION

HIV continues to be a significant cause of morbidity and mortality in the United States (US) (Fauci and Lane, 2020). Public health officials and providers are interested in inferring transmission links between individuals with HIV to inform and improve treatment and prevention approaches (Hogben et al., 2016). In the absence of reliable patient contact histories, phylogenetic analysis of HIV sequence data can and has been used to infer transmission clusters (Leitner et al., 1996), under the assumption that two individuals sharing a most recent common ancestor in a phylogeny are more likely to share or lead to an epidemiological link in the real, unobservable transmission network. The application of molecular epidemiology and cluster inference techniques in public health interventions to disrupt transmission was delineated as one of the four key pillars for ending the HIV epidemic in the US (Fauci et al., 2019).

While historically phylogenetic informativeness of the HIV *pol* genomic region was suggested and contested (Hué et al., 2004; Stürmer et al., 2004), its use is now widespread in cluster inference, often due to the availability of sequences from guideline-recommended routine drug resistance testing, typically performed by commercial Sanger sequencing (Panel on Antiretroviral Guidelines for Adults and Adolescents, 2020). In a recent review of HIV cluster inference, 98 out of 105 (93%) analyzed the *pol* region (Hassan et al., 2017).

The increasing availability of NGS technology has led to longer (across more genes) and deeper (multiple reads that correspond to multiple within-host genomes) sequencing of HIV, and data sets more routinely cover nearly the whole genome at great depth (Voelkerding et al., 2009). Recent evidence suggests improvements in both phylogenetic analysis and cluster inference from longer near-whole-genome HIV sequences obtained with NGS. For example, Yebra et al. (2016) found that the accuracy of phylogenetic reconstruction and cluster inference on simulated sequences improved with longer genomic regions (with the best accuracy from a *gag-pol-env* concatenation). Novitsky et al. (2015) similarly studied effects on cluster inference of using longer genomic regions from near-whole-genome publicly available Sanger sequences and found that the proportion of sequences in clusters increased with longer sequences. Even before the availability of NGS, using longer regions of the HIV genome was shown to improve phylogenetic reconstruction. In one of the earliest studies of HIV sequence data with a known HIV transmission network, Leitner et al. (1996) found that combining data from the *gag* and *env* regions improved the accuracy of phylogenetic reconstruction.

While potential advantages of longer NGS sequences in inferring clusters have been examined (Novitsky et al., 2015; Yebra et al., 2016), advantages of deeper sequencing are less investigated, and whether it can improve HIV molecular clustering inference is unknown. This is due to limitations in established practices of inferring HIV phylogenies across hosts. Researchers often rely on a single consensus sequence for each host that discard all but the majority variant at each site, since most molecular epidemiology approaches require a single fully-resolved sequence per individual in the phylogeny. Accordingly, researchers studying HIV transmission networks discard available information on within-host variation, known to impact phylogenetic inference (Leitner et al., 1996; Leitner, 2019).

The consensus approach, which to date has been employed with Sanger sequencing data in multiple studies of HIV molecular epidemiology (Hassan et al., 2017), carries an underlying statistical assumption of *low relative entropy* (Guang et al., 2016). For HIV, this is equivalent to the strong assumption that a consensus sequence adequately captures all relevant information about HIV diversity within an individual and that variation within hosts has no information about relationships across hosts. While many researchers understand that this assumption is likely wrong and intra-host variation is relevant for phylogenetic analysis of HIV [for a recent review, see Leitner (2019)], in practice researchers have faced limitations in data collection that prevent measuring intra-host variation or in available analysis methods that preserve intra-host variation during alignment and phylogeny. With the advent of long read sequencing technologies for full HIV genomes, obtaining fully resolved sequences that represent the within-host viral population will be possible, but methods to incorporate intra-host variation for transmission cluster analysis will still need to be developed.

Two previous studies have accounted for within-host variation in deeply-sequenced NGS data with coalescent evolutionary models (Romero-Severson et al., 2014; Giardina et al., 2017), but such models still assume a consensus sequence as the observed data. Two other studies introduced methods to use deeply-sequenced HIV data without assuming a consensus, for a different but related epidemiological goal of estimating transmission directionality and identifying multiple infections (Skums et al., 2018; Wymant et al., 2018). Methods also exist that combine haplotype estimation from deeply-sequenced NGS data and phylogenetics (Bendall et al., 2021) as a way to incorporate within-host diversity, but available haplotyping methods have a high computational cost and results are often not sufficiently accurate for cluster analysis (Wymant et al., 2018). Additionally, all aforementioned methods that do not rely on a consensus incorporate within-host diversity by including multiple sequences or tips per sample, which presents difficulties with summarizing or collapsing the resulting phylogenetic tree in order to identify transmission clusters and measure cluster certainty.

In this study, we develop a new method we call *profile sampling* that incorporates within-host HIV genome variation into phylogenetic analyses used to identify transmission clusters. We examine if, and to what extent, incorporation of within-host variation available from deeply-sequenced Illumina-based NGS data provides improved phylogenetic inference and clustering relative to traditional consensus-sequence-based approaches. We focus our analyses on all newly HIV-diagnosed individuals during six months from the largest HIV center in Rhode Island, US.

# MATERIALS AND METHODS

## Data Collection and Sequencing

HIV-1 *pol* Sanger sequences (HXB2 positions 2253-3554), available through clinical care, were collected from the 37 adults (18 years) newly-diagnosed with HIV-1 during the first six months of 2013 and treated at The Miriam Hospital Immunology Center in Providence, Rhode Island, US. Patients at this Center represent ∼80% of the state's HIV epidemic.

In addition, blood was obtained from consenting participants and processed to isolate RNA from plasma ($n$ = 27), and proviral DNA from whole blood ($n$ = 4) or peripheral blood mononuclear cells (PBMC; $n$ = 6). Using Sanger sequencing and Illumina-based NGS, near-whole-genome viral sequences were obtained from one compartment; plasma for participants with detectable viral load and proviral DNA for participants with undetectable viral load or unsuccessful plasma genotyping. The study was approved by the Institutional Review Board at Lifespan, which is the parent health network of The Miriam Hospital.

Total nucleic acids were extracted and an in-house genotyping assay was used to generate the near-whole genome sequence (wgs), based on previously published methods (Nadai et al., 2008; Di Giallonardo et al., 2014). For each sample, two cDNA templates were generated by SuperscriptIII First Strand Synthesis System (Thermofisher, Carlsbad, CA, United States), followed by eight separate nested PCR reactions; these eight amplicons span the near-whole HIV genome. Final amplicon products were sequenced by Sanger using the 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA, United States) and by NGS using Nextera XT DNA Library Prep chemistry (Illumina, San Diego, CA, United States) to generate multiplexed libraries for Illumina's MiSeq platform with 250 base paired-end reads. Sanger consensus sequences were generated manually using Sequencher version 5.2.4 (Gene Codes, Ann Arbor, MI, United States) to confirm degenerate nucleotides. NGS data were processed and demultiplexed using BaseSpace cloud application (Illumina, San Diego, CA, United States). NGS consensus sequences were called at a 20% threshold.

## Profile Sampling

We introduce a new approach for incorporating within-host viral variation into phylogenetic analysis, called *profile sampling* (**Figure 1**). *Profile sampling* builds upon existing methods of phylogenetic and cluster inference by also sampling from within-host viral diversity. We start by aligning each individual's HIV NGS reads using the hivmmer pipeline (Howison et al., 2019), which we developed and now extended to support near-whole-genome HIV data and perform codon-aware alignment within each gene (hivmmer version 0.2.1). A key feature of this pipeline is its use of profile hidden Markov models (HMMs) to model and align collections of HIV sequences. Profile HMMs have been abundantly used for biological sequence analyses and are particularly well-suited to modeling variation in populations of sequences (Eddy, 2004). Briefly, hivmmer performs quality control and error correction in overlapping regions of read pairs using PEAR version 0.9.11 (Zhang et al., 2014), translates them into possible reading frames, aligns them in amino acid space to profile HMMs of all HIV-1 group M reference sequences (Los Alamos National Lab, 2020) using the profile HMM alignment tool HMMER version 3.1b2 (Eddy, 2011), and produces a codon frequency table across the near-whole HIV genome. We refer to this resulting codon frequency table as the individual's HIV *profile*.

Subsequently, we sample 500 fully-resolved sequences from each of the 37 individuals' HIV profile according to the frequency of observed codons in the profile, for a total of 37 × 500 = 18,500 *profile-sampled* sequences. These sequences do not correspond to real strains present in the biological sample, but do capture the empirical distribution of within-host variation at the individual codon level. We note that the sequences do not capture linkage across codons, which is important for the detection and elimination of recombinant HIV strains as part of quality control, but is unessential for phylogenetic analyses. We then collate the 18,500 sequences into 500 *profile-sampled* data sets, by sampling without replacement so that each data set has 37 sequences (one for each individual) and can be used in a phylogenetic analysis with existing methods.

We also use the 18,500 sequences to estimate the within-host diversity for the 37 individuals as the average percent difference in nucleotides across all pairwise comparisons of each individual's 500 *profile-sampled* nucleotide sequences. These pairwise differences are calculated using the Hamming distance (Allam et al., 2011) [also called **p**-distance (Maldarelli et al., 2013; Hassan et al., 2017)].

## Phylogenetic Inference

For *profile sampling*, we perform phylogenetic inference of wgs (HXB2 positions 790-9417) on each of the 500 *profile-sampled* data sets by estimating a multiple sequence alignment with OMM_MACSE version 10.02 (Ranwez et al., 2018) and a maximum-likelihood phylogeny with the GTRCAT model and 100 rapid bootstrap replicates using RAxML version 8.2.12 (Stamatakis, 2014), with HIV-1 group O (GenBank accession L20587.1) as the outgroup. We perform this same phylogenetic inference on three clinically-relevant sub-genomic regions: protease and reverse transcriptase at the beginning of the *pol* gene ("prrt", HXB2 positions 2253-3554), *int* gene (HXB2 positions 4230-5096), and *env* gene (HXB2 positions 6225-8790). The prrt and *int* regions are routinely sequenced in clinical care to detect drug resistance and inform clinical anti-retroviral therapy choices. The *env* region is sequenced to genotypically infer viral tropism and co-receptor usage. Cluster inference is performed on all phylogenies using Cluster Picker (Ragonnet-Cronin et al., 2013) with a threshold of 99% bootstrap support.

In addition to *profile sampling*, we infer phylogenies with similar tools, regions and parameters for the NGS consensus sequences at the 20% threshold and the Sanger sequences. We perform cluster inference on all consensus phylogenies using a similar method as for *profile sampling*. We do not impose a genetic distance threshold because empirically-justified thresholds that are comparable across the near-whole-genome and the *int* and *env* regions have not to our knowledge been established. This approach of using only bootstrap criteria for cluster detection is consistent with methods commonly used in the broader HIV cluster analysis literature (Hassan et al., 2017).

We investigate the impact of within-host diversity on phylogenetic topology and evolutionary distance estimates in the sub-genomic and near-whole-genome regions. To examine variation in topology, we first calculate pairwise geodesic distance (Billera et al., 2001; Owen and Provan, 2011) among the 500

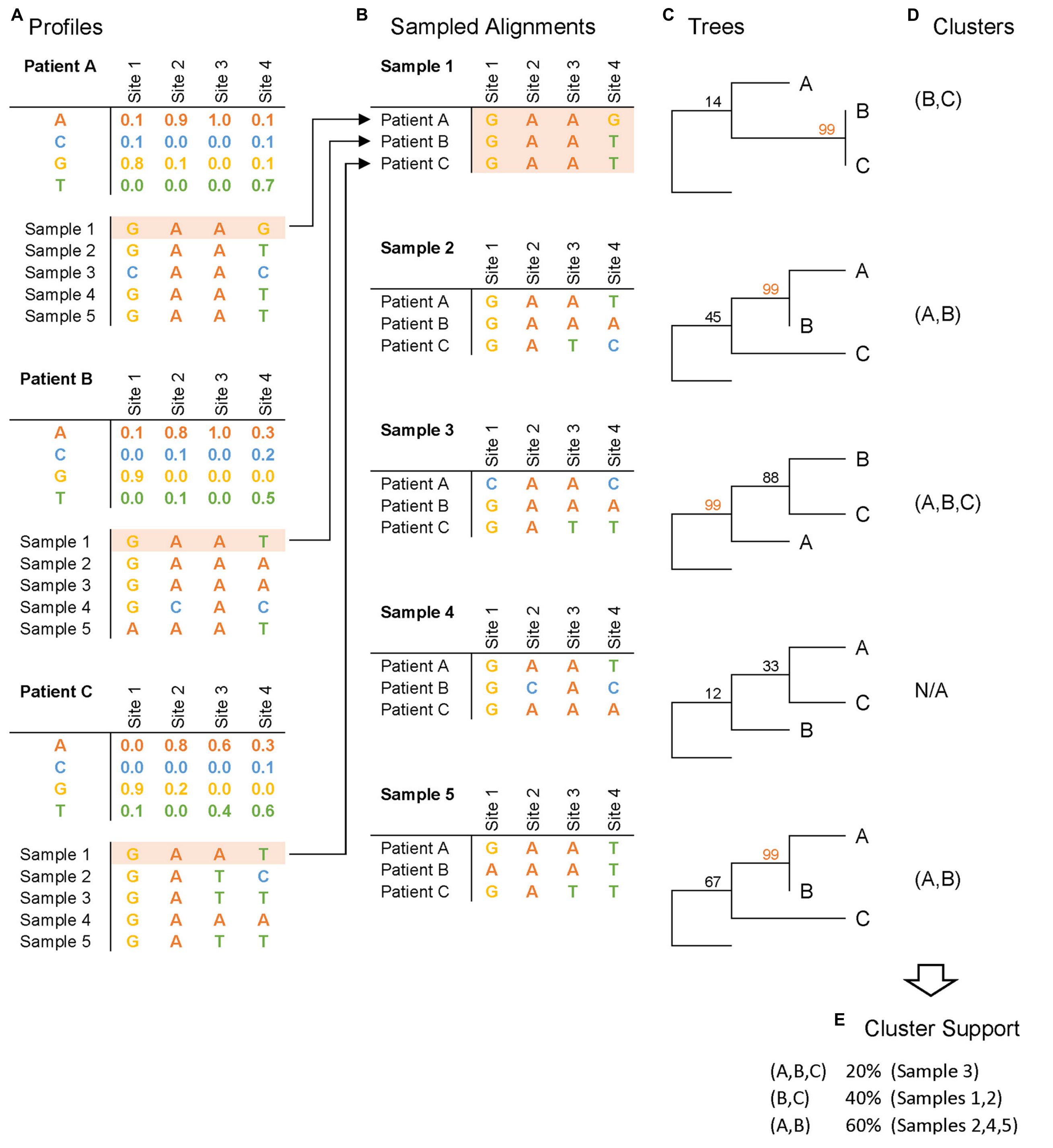


**FIGURE 1 |** Profile sampling pipeline. This schematic figure depicts the four steps of the *profile sampling* process, illustrated here with 5 samples per patient: **(A)** NGS-derived frequencies at each HIV genome site for each patient are generated, and synthetic sequences are sampled from these frequency tables to summarize intra-host variation; **(B)** sampled sequences are collated across patients to construct sampled alignments; **(C)** phylogenetic trees are inferred with bootstrap support from the alignments; **(D)** clusters are inferred based on phylogenetic bootstrap support (illustrated here with bootstrap support ≥ 99); and **(E)** cluster support is measured as the frequency that a cluster is inferred across samples.

phylogenies from the profile samples, as well as phylogenies from the NGS consensus and Sanger sequences. Then we perform multi-dimensional scaling on the resulting distance matrix to visualize topological space in two dimensions. Next, to examine variation in estimated evolutionary distance, we sum the branch lengths within each phylogeny across all branches and across only tip branches and visualize the distribution of these branch length sums. These analyses establish to what extent phylogenies from consensus sequences (which are point estimates) summarize the underlying variation in two important aspects of the phylogeny: estimated topology and estimated evolutionary distance.

Finally, we examine the clusters that are detected in phylogenies of NGS consensus sequences versus Sanger sequences, and across the four genomic regions and their *profile sampling* support, using the frequency that a cluster appears across the 500 *profile-sampled* phylogenies. We refer to this value as the *profile-sampled* support and note that it can be conceived as analogical to the conventional bootstrap support for evaluating robustness of an individual phylogeny's topology but extends that feature to evaluating robustness of cluster detection using within-host sequence variation.

All analysis source code is available from https://github.com/kantorlab/hiv-profile-sampling.

## RESULTS

### Profile Sampling Estimates of Within-Host Diversity

**Figure 2** shows the estimated within-host percent diversity in each examined genomic region across individuals, ordered by *env*, which we expected *a priori* to be the most variable region. The largest estimated diversity is in *env* for individual MC28 (3.9%), and *env* has the overall largest estimated diversity range (0.2–3.9%, mean 1.5%). The other regions have ranges of 0.2–1.9% (mean 0.9%) for prrt, 0.1–2.0% (mean 0.9%) for *int*, and 0.2–2.6% (mean 1.2%) for wgs. Such within-host estimations are not feasible with conventional consensus Sanger or NGS approaches, although methods such as phyloscanner and HAPHPIPE that utilize deeper NGS sequencing to build phylogenies with multiple tips per sample, are able to also quantify within-host diversity (Wymant et al., 2018; Bendall et al., 2021).

### Phylogenetic Estimates Are Sensitive to Within-Host Diversity

**Figure 3** demonstrates multi-dimensional scaling on the *profile-sampled* phylogenies and the phylogenies from Sanger and NGS consensus sequences *within* each genomic region. The *profile sampling* approach reveals for each genomic region a multi-modal topological space in which phylogenies inferred from both Sanger and NGS consensus sequences are outliers; a result that is confirmed by multi-dimensional scaling *across* all regions (**Figure 4**). A key difference between the consensus and *profile-sampled* sequences is that consensus sequences contain ambiguous nucleotides at sites with $\geq$2 nucleotides by Sanger Sequencing base calling or with $\geq$20% frequency for NGS. In contrast, *profile-sampled* sequences by construction have no ambiguous sites, and ambiguity is instead incorporated into analyses through frequency of the ambiguous nucleotides across the 500 samples.

**Figure 5** shows the distribution of branch length sums across compared phylogenies. Overall, estimates are larger in *env* and wgs, and smaller when restricting to only tip branches. In some cases, consensus phylogenies provide an adequate summary of the distribution (as in the phylogeny of the NGS consensus sequence for tip branches for wgs). In other cases, consensus phylogenies have estimates that are outliers in the distribution (as in the phylogenies from NGS and Sanger consensus sequences for all branches in wgs and *env*).

Taken together, the heterogeneity between the phylogenetic results from *profile sampling* and consensus-inferred point estimates demonstrate that within-host virus sequence diversity impacts the inference of virus phylogeny across individuals, and that the consensus approach to handling ambiguity and collapsing within-host sequence variation can obscure both the magnitude and effect of these impacts.

### *Profile-Sampled* Cluster Support Differs by Sequencing Depth and Genomic Region

Combining the results of all examined methods (Sanger, NGS consensus, NGS *profile sampling*) and genomic regions (prrt, int, env, wgs) there were overall 12 identified clusters among the 37 participants. Seven clusters had two members, four had three members, and one had five members. **Figure 6** demonstrates comparison of cluster detection by examined methods and genomic regions. Some clusters had consistently high support (>75%) across all regions (e.g., MC25/MC26/MC52 and MC14/MC59). Other clusters had higher support in certain regions (e.g., MC17/MC20/MC21 in *env* and wgs). Eight clusters across different regions were detected by *profile sampling* but not by consensus methods, while all clusters detected by consensus methods were detected by *profile sampling*. One larger cluster, MC37/MC41/MC47/MC53/MC56, was detected only by *profile sampling* with the wgs dataset.

By providing previously-unavailable cluster support that considers within-host "deep" viral variation, *profile sampling* in the wgs dataset allowed detection of the largest (all 12) overall number of clusters. The clusters detected in wgs also had the highest overall *profile-sampled* support, as compared to the other genomic regions. The median *profile-sampled* support was 99.8% for wgs, 77.6% for *env*, 41.4% for *int*, and 51.3% for prrt. The *profile-sampled* support for wgs was significantly larger than for *int* (*p*-value = 0.005, Dunn's test of multiple comparisons using paired rank sums with Holm-Bonferroni correction) and prrt (*p*-value = 0.010), but not significantly larger than for *env* (*p*-value = 0.092).

The phylogenies of NGS consensus sequences detected only six clusters in prrt, seven in *int*, seven in *env*, and seven in wgs (**Figure 7**). The phylogenies of Sanger sequences detected only four clusters in prrt, six in *int*, seven in *env*, and seven in wgs (**Figure 8**). Only one cluster (MC25/MC26/MC52) was consistently detected across phylogenies from NGS and Sanger consensus sequences, and across all regions.

The median *profile-sampled* support for clusters detected by Sanger consensus sequences (green and yellow cells, **Figure 6**) was 60.2%, not different than for those detected by NGS consensus sequences (72.5%; orange and yellow cells in **Figure 6**;

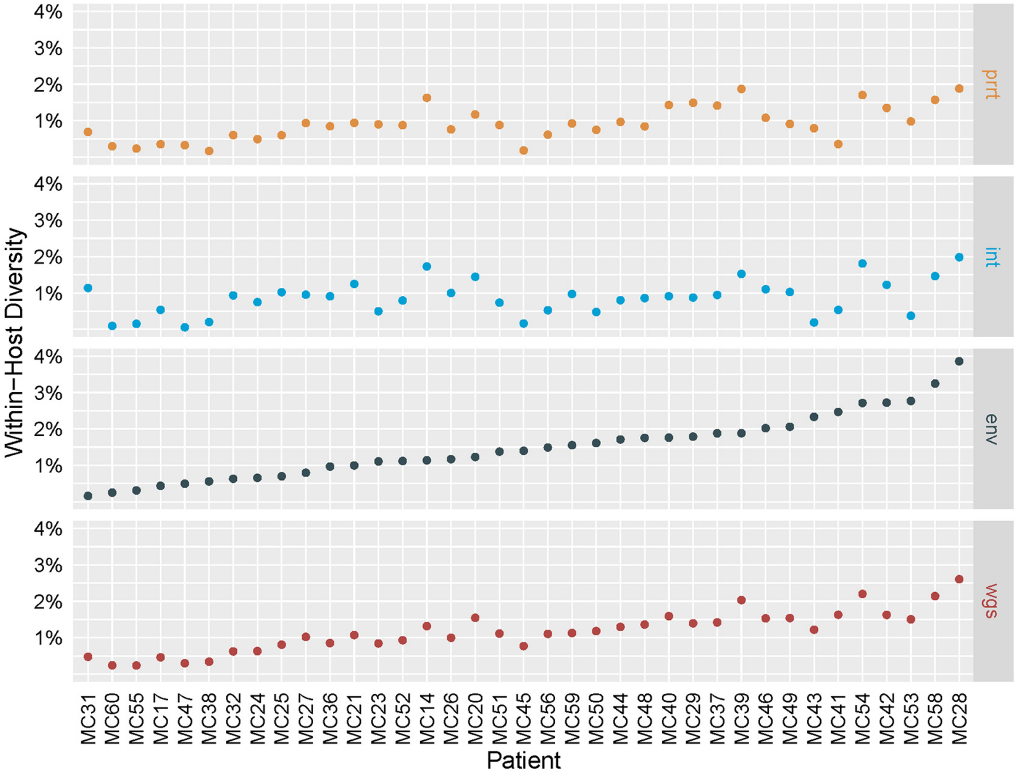


**FIGURE 2 |** Intra-host genetic diversity by genomic region. Intra-host genetic diversity (*Y* axis; defined as the average percent difference across all pairwise comparisons of the 500 *profile-sampled* nucleotide sequences for an individual) of the four examined genomic regions (gray boxes on the right) in the 37 sampled individuals (*X* axis) is highest in *env* for most individuals and lies within the range of previously reported values.

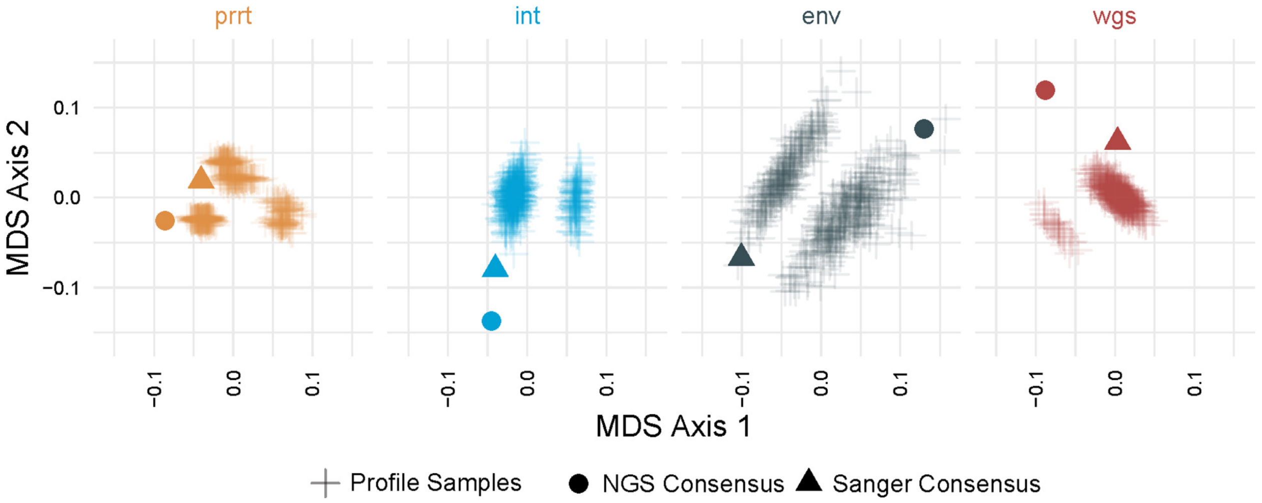


**FIGURE 3 |** Multi-dimensional scaling (MDS) of pairwise geodesic distance among maximum-likelihood phylogenies from the *profile sampling* approach within genomic regions. MDS Axis 1 (*X* axis) and Axis 2 (*Y* axis) show that the space of inferred phylogenies is multi-modal for all genomic regions. The phylogenies from NGS and Sanger consensus sequences (dot and triangle) are point estimates that do not capture the full variation in phylogenies that can be inferred from deeply-sequenced NGS data (plus signs) in all examined genomic regions (colors).

*p*-value = 0.415, Wilcoxon signed-rank test). Totaling the clusters detected across the four regions, phylogenies of Sanger consensus sequences detected fewer clusters (27) than phylogenies of NGS consensus sequences (31) or *profile sampling* (43); and detected fewer clusters in each region except *env*. Cluster support values were higher for clusters detected by phylogenies of both NGS and

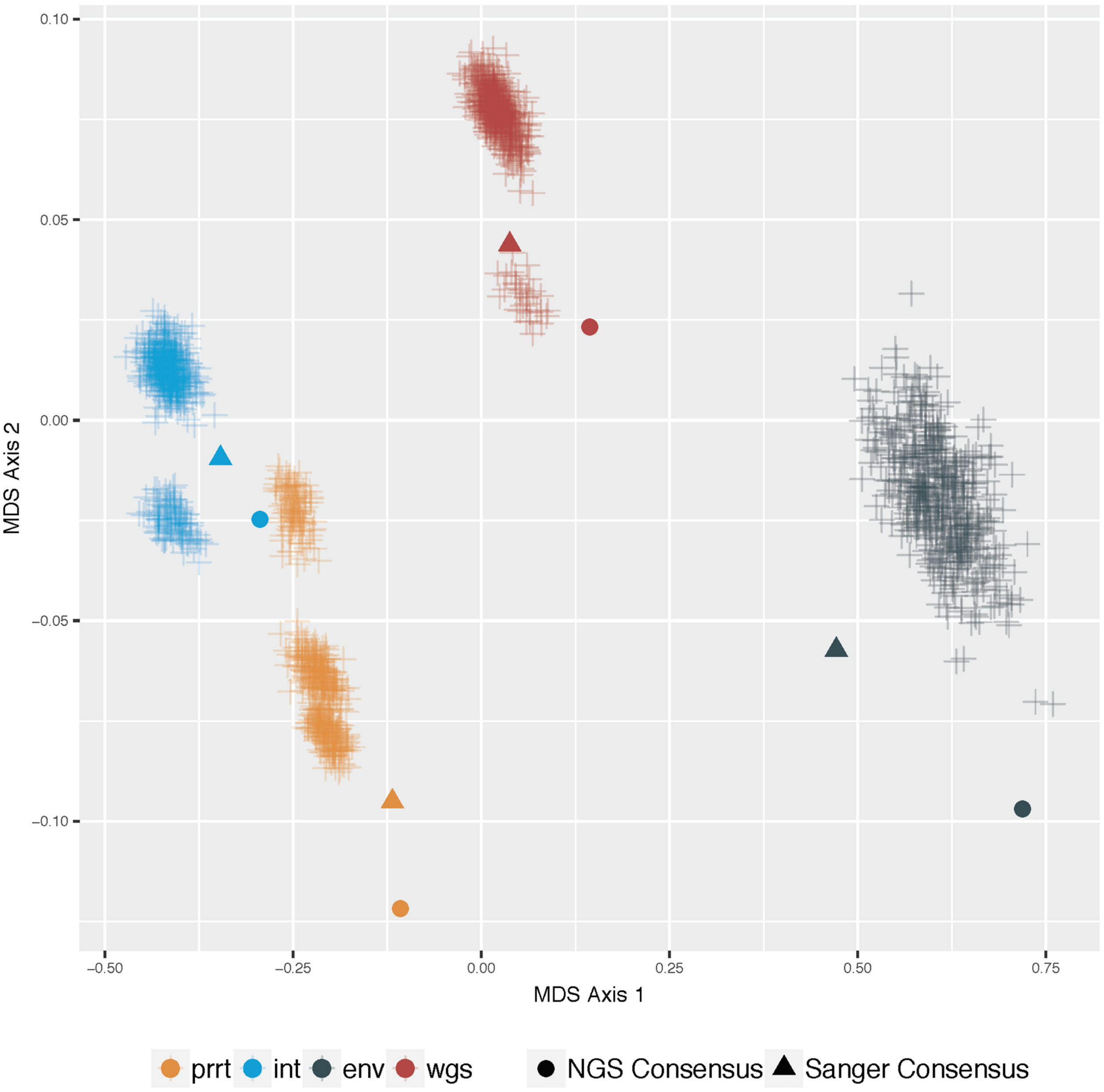


**FIGURE 4 |** Multi-dimensional scaling (MDS) of pairwise geodesic distance among maximum-likelihood phylogenies from the *profile-sampling* approach across all genomic regions. MDS Axis 1 (*X* axis) and Axis 2 (*Y* axis) show that the space of inferred phylogenies is multi-modal for all genomic regions. The phylogenies from consensus sequences (dot and triangle) are point estimates that do not capture the full variation in phylogenies that can be inferred from deeply-sequenced NGS data (plus signs) in all examined genomic regions (colors).

Sanger sequences (yellow cells, **Figure 6**; median cluster support 98.5%) than those detected only by one or the other (orange or green cells, **Figure 6**; median cluster support 68.1%).

## DISCUSSION

Current phylogenetic approaches to inference of HIV transmission clusters utilize consensus sequences to summarize within-host sequence variation. We introduce a different summarization strategy, *profile sampling,* that preserves the within-host sequence variation provided by the deeper sequencing that is now widely available. In a dataset of all newly HIV-diagnosed individuals over six months at the largest HIV center in Rhode Island, United States, deeper sequencing provided by NGS and incorporated by the newly-introduced *profile sampling* captured within-host diversity, revealing clusters detected by *profile sampling* but not by consensus approaches, including one larger cluster found only with *profile sampling* of the wgs. This suggests that routinely used consensus sequence

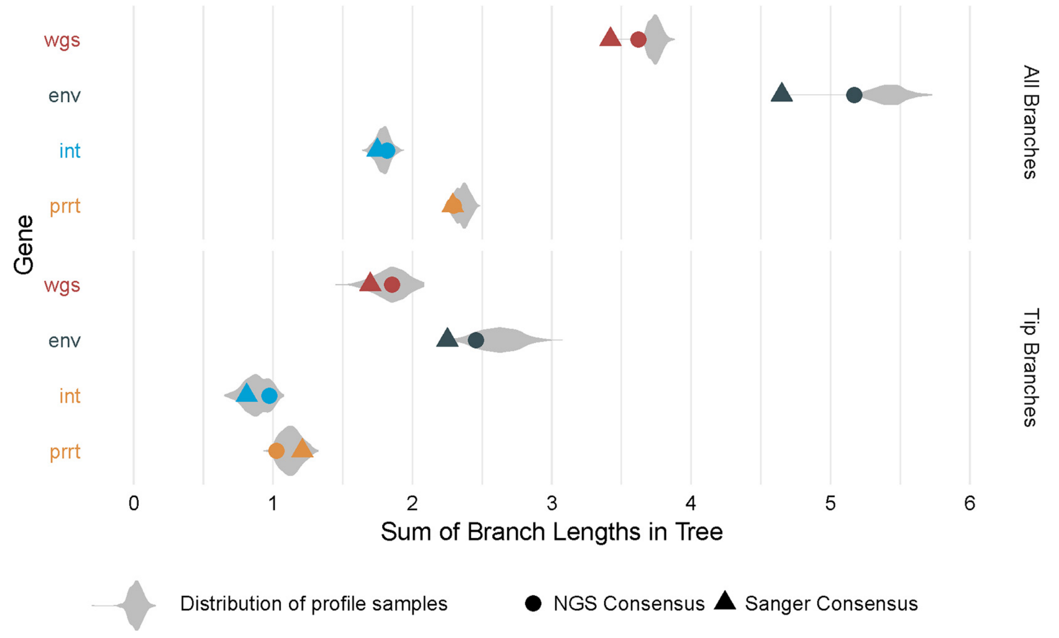


**FIGURE 5 |** Distribution of branch length sums across phylogenies. The figure demonstrates total branch lengths (*X* axis), in each of the *profile-sampled* phylogenies (*Y* axis and colors). The phylogenies from consensus sequences (dot and triangle) can lie at extreme values within these distributions, both when considering the lengths across all branches (top) and the lengths across only the branches at the tips (bottom).

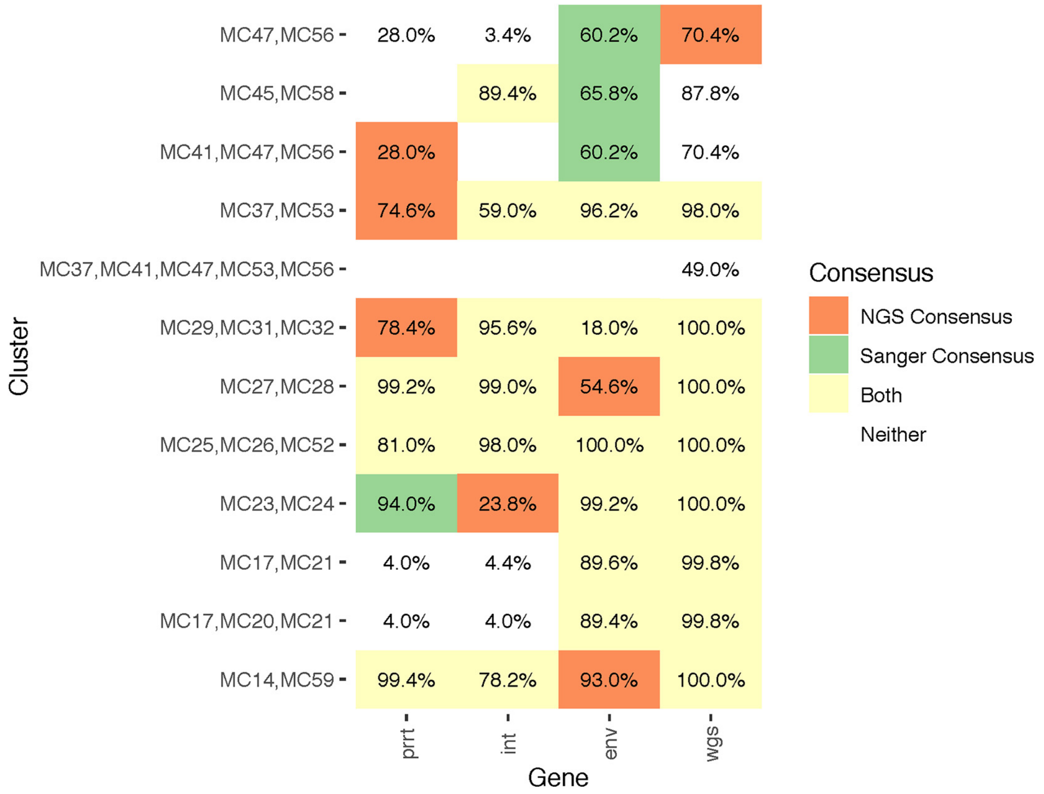


**FIGURE 6 |** Quantitative differences in *profile-sampled* cluster support across genomic regions. The figure illustrates the clusters and their subclusters (*Y* axis) identified by Sanger versus NGS consensus sequences (colors; see legend) across genomic regions (*X* axis). Numeric values indicate cluster support from the *profile sampling* method. A blank cell indicates that the cluster was not detected in that genomic region.

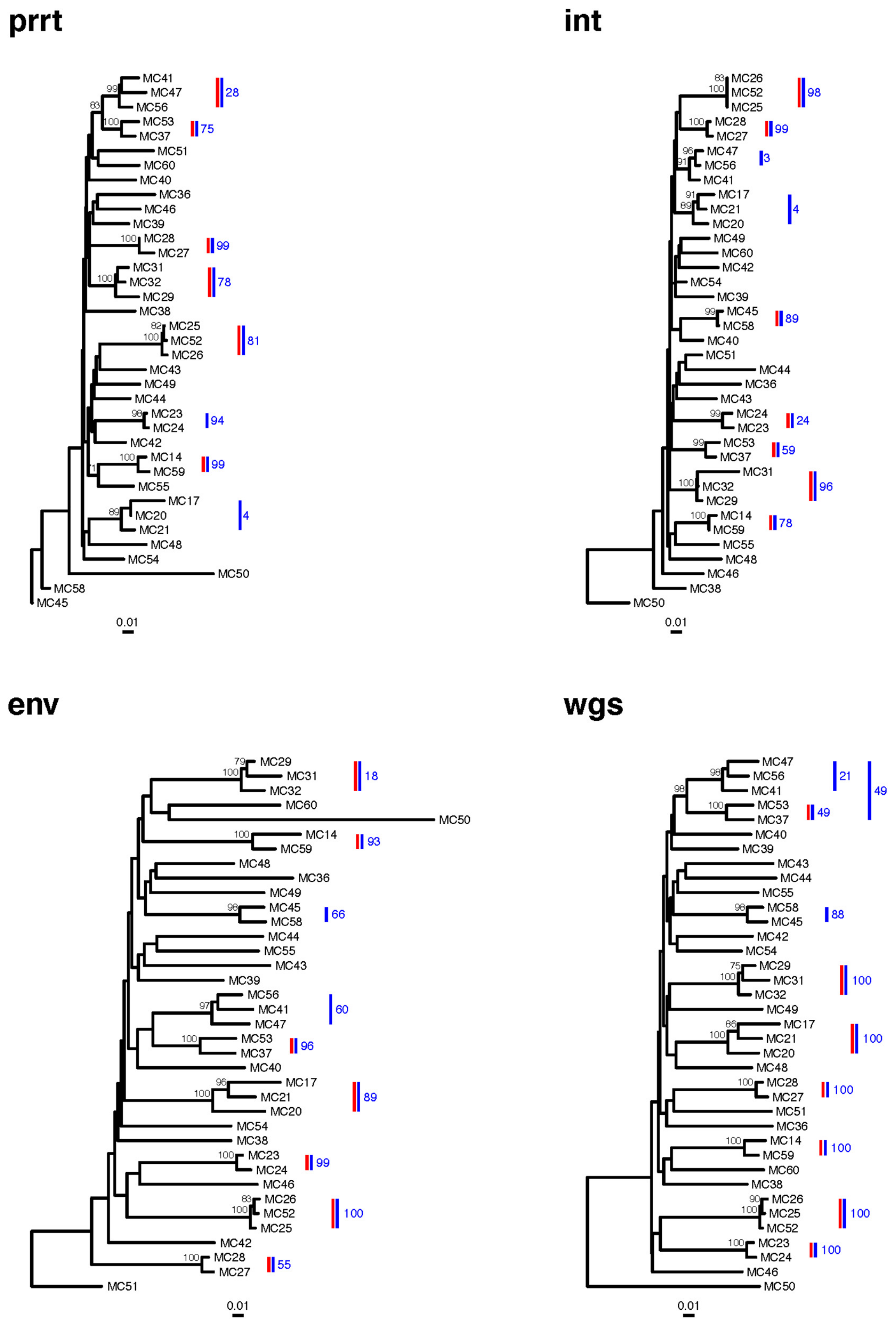


**FIGURE 7 |** Next-generation sequencing (NGS) consensus sequence phylogenetic trees of 37 new HIV diagnoses in RI according to genomic region. The figure demonstrates clusters in phylogenetic trees from four genomic regions (prrt-protease reverse transcriptase; int-integrase; env-envelope; wgs-whole genome sequence). Clusters (≥99% bootstrap support) inferred from the phylogenies of NGS consensus sequences (vertical red bars) differ across genomic regions. The largest number of clusters was inferred from *int, env,* and wgs, and the smallest number from prrt. *Profile sampling* detected additional clusters (vertical blue bars) and provided a bootstrap-like measure of cluster support (annotation to blue bars). Bootstraps > 70% are shown to the left of the relevant node. Trees are rooted by an HIV-1 group O sequence, which is omitted from the plots.

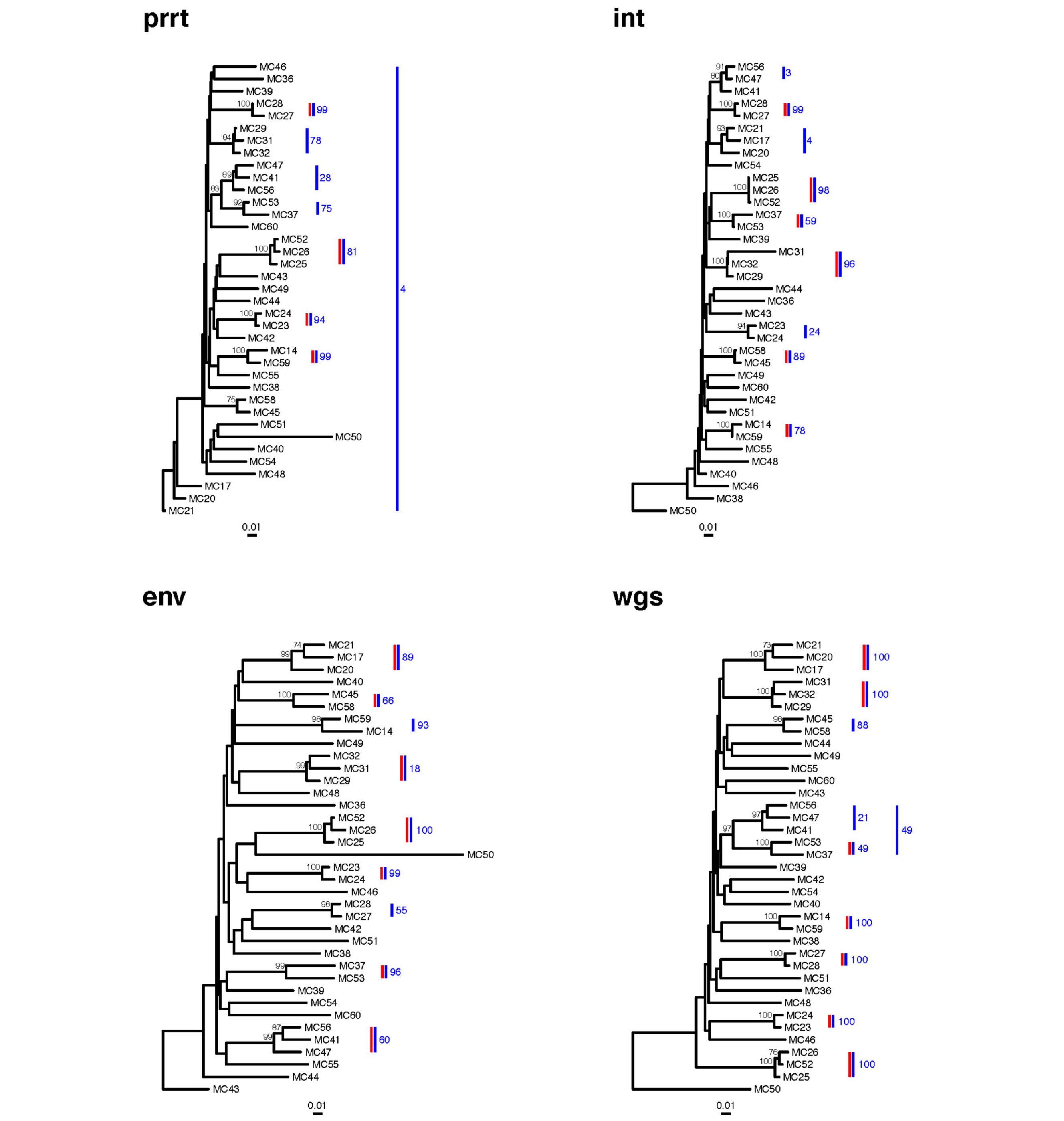


**FIGURE 8 |** Sanger sequence phylogenetic trees of 37 new HIV diagnoses in RI according to genomic region. The figure demonstrates clusters in phylogenetic trees from four genomic regions (prrt-protease reverse transcriptase; int-integrase; env-envelope; wgs-whole genome sequence). Clusters (≥99% bootstrap support) inferred from the phylogenies of Sanger consensus sequences (vertical red bars) differ across genomic regions. The largest number of clusters was inferred from env and wgs, and the smallest number from prrt. *Profile sampling* detected additional clusters (vertical blue bars) and provided a bootstrap-like measure of cluster support (annotation to blue bars). Bootstraps > 70% are shown to the left of the relevant node. Trees are rooted by an HIV-1 group O sequence, which is omitted from the plots.

approaches discard potentially relevant information present in NGS data, and that considering this additional information in phylogenetic analysis may improve the robustness of HIV cluster detection. *Profile sampling* can thus provide a new quantitative measure of cluster confidence with potential applications to public health activities. Such activities could be better justified in scenarios where clusters triggering them have high cluster support from deep-sequenced data, though this was not addressed here and needs to be demonstrated.

*Profile sampling* complements well-established bootstrapping methods, and in some senses is orthogonal to them. Phylogenetic inference depends on a sequence alignment, where each row corresponds to a single host and each column corresponds to a given genomic position. Bootstrapping resamples columns of the matrix with replacement, giving an indication of how consistent signal is across genomic positions. *Profile sampling*, on the other hand, resamples each site in the alignment given the sequence diversity observed within each host. This gives an indication of how consistent signal is across HIV genomes within each host. Variation in within-host diversity could be due to a variety of biological factors, like viral mutation, effective viral population size, and time since infection, as well as technical factors like sequencing depth and sequencing errors. Not accommodating this variation could lead to overconfidence in results or missed clinically relevant phylogenetic signals.

Although the standard practice of collapsing within-host diversity into a single consensus sequence simplifies downstream analyses, the results presented here demonstrate that this practice discards potentially relevant biological results and may mislead phylogenetic analysis and resulting epidemiological consequences. For example, public health activities triggered by phylogenetic inference of HIV molecular clustering to inform and improve prevention and treatment interventions can be affected (Peters et al., 2016). In our data, clusters vary in their *profile-sampled* support, and consensus approaches can fail to detect clusters supported by deep-sequenced data, as in the case of the largest cluster, which was detected by *profile sampling*, not by consensus approaches. As data acquisition increasingly shifts to NGS approaches (Kantor, 2021), it is important to compare results of larger datasets from new methods to the more common conventional Sanger *pol* consensus sequences.

Much of the enthusiasm about shifting from Sanger sequencing to NGS has been due to reducing costs and the ability to more easily collect data on the entire HIV genome rather than few genes. Our results suggest that much benefit from NGS may also come from its greater sequencing depth, capturing viral sequence variation within individuals. This benefit can only be realized, though, if this variation is propagated to phylogenetic analyses, such as by *profile sampling* introduced here, rather than being collapsed to a consensus sequence, as is conventionally done. We suggest creating a profile that captures that variation, performing multiple phylogenetic analyses on sequences sampled from the profile, and then summarizing the phylogenetic analyses. This summary method can also take advantage of the output from long-read sequencing technologies which are able to provide fully resolved sequences from the viral population and which are starting to replace short-read NGS sequencing in a number of HIV labs. Future tools could incorporate the variation directly into the phylogenetic inference process itself (Leitner, 2019).

In our comparison of cluster inference across genomic regions, we found that fewer clusters were detected overall in prrt and *int* compared to *env* and wgs. Prior studies of clustering from Sanger consensus sequences present mixed results on prevalence of clustering across genomic regions. Some studies found concordant clustering across *gag-env* (Han et al., 2009) and *gag-pol-env* (Kaye et al., 2008; English et al., 2011), while others found fewer clusters in *pol* than in *env* (Kapaata et al., 2013), or in *gag-env* than in *pol* (Ndiaye et al., 2013). The additional information available in deep-sequenced NGS data, along with cluster support measures provided by *profile sampling*, could help resolve differences, as suggested here. In our data, not only were more clusters detected in the near-whole length genome, but those clusters also had higher cluster support as measured by deep-sequenced NGS data. While prior studies demonstrated better accuracy of cluster inference on simulated NGS sequences when using wgs (Yebra et al., 2016) and that the proportion of Sanger sequences in clusters increased with longer sequence regions (Novitsky et al., 2015), we have demonstrated here that deeply-sequenced, near-whole length NGS data can be used with *profile sampling* to detect clusters undetectable by consensus approaches. The impact of this approach for public health remains to be determined.

One limitation of our study is the small number of participants and the short timeframe they were enrolled in. However, participants represent a dense temporal sampling, and comprise all newly HIV-diagnosed individuals in a six-month period at the largest HIV center in Rhode Island, in which 80% of the state's people with HIV are cared for. The overall size of the HIV epidemic in Rhode Island was estimated as 2,396 individuals in 2016 (Rhode Island Department of Health, 2019), but NGS data for this population are not currently available beyond those presented here. In future work, we will apply *profile sampling* to larger NGS data sets, to assess cluster inference concordance between Sanger and NGS data, and its impact on public health actions to halt HIV transmission. Additionally, we do not know the real number of clusters or the true transmission chains, a limitation with all studies on HIV transmission networks. Our construction of HIV profiles from NGS data is also limited by the accuracy of the NGS assays themselves. The codon frequencies in the profiles may be biased measures of the true within-host diversity because of biases in PCR amplification, as well as a variety of technical factors related to next-generation sequencing and analysis [see Howison et al. (2019) for a detailed discussion]. Sequencing protocols such as Primer ID (Jabara et al., 2011; Zhou et al., 2020) have been introduced to reduce and correct for these biases, and should be considered in the future.

In conclusion, the true HIV transmission network is unknown, but phylogenetic analysis and cluster inference are promising tools for aiding clinicians and public health officials in better understanding and in disrupting HIV transmission (Fauci et al., 2019). Most current phylogenetic approaches do not fully utilize the information on within-host diversity available in deep-sequenced, near-whole-genome NGS data. As NGS data sets are increasingly available and become more representative of HIV epidemics, we suggest that the additional information they measure has the potential to improve the robustness of HIV molecular cluster inference, the impact of which needs to be further investigated.

## DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of patient privacy concerns. Requests to access the datasets should be directed to RK at rkantor@brown.edu.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board at Lifespan. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

AG, MH, PC, CD, and RK contributed to the conception of the study. LL, MD'A, PC, and RK contributed to data collection and organization. AG and CL created the method. AG and MH contributed to implementation, analysis and visualization. AG, MH, CD, and RK contributed to writing the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

## FUNDING

This work was supported by the National Institutes of Health (R01AI136058, K24AI134359, P30AI042853 and P20GM109035) and a Brown University DEANS Award.

## ACKNOWLEDGMENTS

We thank Dr. Mia Coetzer for assistance with next-generation sequencing and feedback on an earlier draft of the manuscript. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

## REFERENCES

Allam, O., Samarani, S., and Ahmad, A. (2011). Hammering out HIV-1 incidence with Hamming distance. *AIDS* 25, 2047–2048. doi: 10.1097/QAD.0b013e32834bac66

Bendall, M. L., Gibson, K. M., Steiner, M. C., Rentia, U., Pérez-Losada, M., and Crandall, K. A. (2021). HAPHPIPE: haplotype reconstruction and phylodynamics for deep sequencing of intrahost viral populations. *Mol. Biol. Evol.* 38, 1677–1690. doi: 10.1093/molbev/msaa315

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* 27, 733–767.

Di Giallonardo, F., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., et al. (2014). Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res.* 42:e115. doi: 10.1093/nar/gku537

Eddy, S. R. (2004). What is a hidden Markov model? *Nat. Biotechnol.* 22, 1315–1316.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

English, S., Katzourakis, A., Bonsall, D., Flanagan, P., Duda, A., Fidler, S., et al. (2011). Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants. *Retrovirology* 8:54. doi: 10.1186/1742-4690-8-54

Fauci, A. S., and Lane, H. C. (2020). Four decades of HIV/AIDS – much accomplished, much to do. *N. Engl. J. Med.* 383, 1–4. doi: 10.1056/NEJMp1916753

Fauci, A. S., Redfield, R. R., Sigounas, G., Weahkee, M. D., and Giroir, B. P. (2019). Ending the HIV epidemic: a plan for the United States. *JAMA* 321:844. doi: 10.1001/jama.2019.1343

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., and Leitner, T. (2017). Inference of transmission network structure from HIV phylogenetic trees. *PLoS Comput. Biol.* 13:e1005316. doi: 10.1371/journal.pcbi.1005316

Guang, A., Zapata, F., Howison, M., Lawrence, C. E., and Dunn, C. W. (2016). An integrated perspective on phylogenetic workflows. *Trends Ecol. Evol.* 31, 116–126. doi: 10.1016/j.tree.2015.12.007

Han, Z., Leung, T. W., Zhao, J., Wang, M., Fan, L., Li, K., et al. (2009). A HIV-1 heterosexual transmission chain in Guangzhou, China: a molecular epidemiological study. *Virol. J.* 6:148. doi: 10.1186/1743-422X-6-148

Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjörnsson, J. (2017). Defining HIV-1 transmission clusters based on sequence data. *AIDS* 31, 1211–1222. doi: 10.1097/QAD.0000000000001470

Hogben, M., Collins, D., Hoots, B., and O'Connor, K. (2016). Partner services in sexually transmitted disease prevention programs: a review. *Sex. Transm. Dis.* 43, S53–S62. doi: 10.1097/OLQ.0000000000000328

Howison, M., Coetzer, M., and Kantor, R. (2019). Measurement error and variant-calling in deep Illumina sequencing of HIV. *Bioinformatics* 35, 2029–2035. doi: 10.1093/bioinformatics/bty919

Hué, S., Clewley, J. P., Cane, P. A., and Pillay, D. (2004). HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18, 719–728. doi: 10.1097/00002030-200403260-00002

Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20166–20171. doi: 10.1073/pnas.1110064108

Kantor, R. (2021). Next generation sequencing for HIV-1 drug resistance testing-a special issue walkthrough. *Viruses* 13:340. doi: 10.3390/v13020340

Kapaata, A., Lyagoba, F., Ssemwanga, D., Magambo, B., Nanyonjo, M., Levin, J., et al. (2013). HIV-1 subtype distribution trends and evidence of transmission clusters among incident cases in a rural clinical cohort in Southwest Uganda, 2004–2010. *AIDS Res. Hum. Retroviruses* 29, 520–527. doi: 10.1089/AID.2012.0170

Kaye, M., Chibo, D., and Birch, C. (2008). Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. *J. Acquir. Immune Defic. Syndr.* 49, 9–16. doi: 10.1097/QAI.0b013e318180c8af

Leitner, T. (2019). Phylogenetics in HIV transmission: taking within-host diversity into account. *Curr. Opin. HIV AIDS* 14, 181–187. doi: 10.1097/COH.0000000000000536

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., and Albert, J. (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. U.S.A.* 93, 10864–10869. doi: 10.1073/pnas.93.20.10864

Los Alamos National Lab (2020). *HIV Database [Internet]*. Available online at: http://www.hiv.lanl.gov/ (accessed February 9, 2020).

Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., Polis, M. A., et al. (2013). HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *J. Virol.* 87, 10313–10323. doi: 10.1128/JVI.01225-12

Nadai, Y., Eyzaguirre, L. M., Constantine, N. T., Sill, A. M., Cleghorn, F., Blattner, W. A., et al. (2008). Protocol for nearly full-length sequencing of HIV-1 RNA from plasma. *PLoS One* 3:e1420. doi: 10.1371/journal.pone.0001420

Ndiaye, H. D., Tchiakpe, E., Vidal, N., Ndiaye, O., Diop, A. K., Peeters, M., et al. (2013). HIV type 1 subtype c remains the predominant subtype in men having sex with men in Senegal. *AIDS Res. Hum. Retroviruses* 29, 1265–1272. doi: 10.1089/aid.2013.0140

Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V., and Essex, M. (2015). Importance of viral sequence length and number of variable and informative sites in analysis of HIV clustering. *AIDS Res. Hum. Retroviruses* 31, 531–542. doi: 10.1089/AID.2014.0211

Owen, M., and Provan, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 8, 2–13. doi: 10.1109/TCBB.2010.3

Panel on Antiretroviral Guidelines for Adults and Adolescents (2020). *Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV*. Available online at: https://clinicalinfo.hiv.gov/sites/default/files/guidelines/documents/AdultandAdolescentGL.pdf (accessed February 9, 2020).

Peters, P. J., Pontones, P., Hoover, K. W., Patel, M. R., Galang, R. R., Shields, J., et al. (2016). HIV infection linked to injection use of Oxymorphone in Indiana, 2014–2015. *N. Engl. J. Med.* 375, 229–239. doi: 10.1056/NEJMoa1515195

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J., et al. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinformatics* 14:317. doi: 10.1186/1471-2105-14-317

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol. Biol. Evol.* 35, 2582–2584. doi: 10.1093/molbev/msy159

Rhode Island Department of Health (2019). *HIV Progress [Internet]*. Available online at: https://health.ri.gov/data/hiv/ (accessed December 12, 2019).

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., and Leitner, T. (2014). Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Mol. Biol. Evol.* 31, 2472–2482. doi: 10.1093/molbev/msu179

Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., et al. (2018). QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics* 34, 163–170. doi: 10.1093/bioinformatics/btx402

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stürmer, M., Preiser, W., Gute, P., Nisius, G., and Doerr, H. W. (2004). Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. *AIDS* 18, 2109–2113. doi: 10.1097/00002030-200411050-00002

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55, 641–658. doi: 10.1373/clinchem.2008.112789

Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., et al. (2018). PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity. *Mol. Biol. Evol.* 25, 719–733. doi: 10.1093/molbev/msx304

Yebra, G., Hodcroft, E. B., Ragonnet-Cronin, M. L., Pillay, D., Brown, A. J. L., PANGEA_HIV Consortium, et al. (2016). Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. *Sci. Rep.* 6:39489. doi: 10.1038/srep39489

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina paired-end read merger. *Bioinformatics* 30, 614–620. doi: 10.1093/bioinformatics/btt593

Zhou, S., Sizemore, S., Moeser, M., Zimmerman, S., Samoff, E., Mobley, V., et al. (2020). Near real-time identification of recent HIV transmissions, transmitted drug resistance mutations, and transmission networks by MPID-NGS in North Carolina. *J. Infect. Dis.* 223, 876–884. doi: 10.1093/infdis/jiaa417

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 25 February 2022
doi: 10.3389/fmicb.2022.821006

# Molecular Epidemiology and Transmission Dynamics of the HIV-1 Epidemic in Ethiopia: Epidemic Decline Coincided With Behavioral Interventions Before ART Scale-Up

*Dawit Assefa Arimide*[1,2], *Luis Roger Esquivel-Gómez*[3], *Yenew Kebede*[4], *Sviataslau Sasinovich*[1], *Taye Balcha*[1], *Per Björkman*[1], *Denise Kühnert*[3] *and Patrik Medstrand*[1]*

[1] *Department of Translational Medicine, Lund University, Malmo, Sweden,* [2] *TB/HIV Department, Ethiopian Public Health Institute, Addis Ababa, Ethiopia,* [3] *Transmission, Infection, Diversification and Evolution Group, Max-Planck Institute for the Science of Human History, Jena, Germany,* [4] *Africa Centre for Disease Prevention and Control, Africa Union Commission, Addis Ababa, Ethiopia*

OPEN ACCESS

***Edited by:***
*Kok Keng Tee,*
*University of Malaya, Malaysia*

***Reviewed by:***
*Teiichiro Shiino,*
*National Center for Global Health and Medicine, Japan*
*Edson Delatorre,*
*Federal University of Espirito Santo, Brazil*
*Kledoaldo Lima,*
*Hospital das Clínicas, Universidade Federal de Pernambuco, Brazil*

****Correspondence:***
*Patrik Medstrand*
*patrik.medstrand@med.lu.se*

***Specialty section:***
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *23 November 2021*
***Accepted:*** *24 January 2022*
***Published:*** *25 February 2022*

***Citation:***
*Arimide DA, Esquivel-Gómez LR, Kebede Y, Sasinovich S, Balcha T, Björkman P, Kühnert D and Medstrand P (2022) Molecular Epidemiology and Transmission Dynamics of the HIV-1 Epidemic in Ethiopia: Epidemic Decline Coincided With Behavioral Interventions Before ART Scale-Up. Front. Microbiol. 13:821006. doi: 10.3389/fmicb.2022.821006*

**Background:** Ethiopia is one of the sub-Saharan countries hit hard by the HIV epidemic. Previous studies have shown that subtype C dominates the Ethiopian HIV-1 epidemic, but the evolutionary and temporal dynamics of HIV-1 in Ethiopia have not been closely scrutinized. Understanding the evolutionary and epidemiological pattern of HIV is vital to monitor the spread, evaluate and implement HIV prevention strategies.

**Methods:** We analyzed 1,276 Ethiopian HIV-1 subtype C polymerase (*pol* sequences), including 144 newly generated sequences, collected from different parts of the country from 1986 to 2017. We employed state-of-art maximum likelihood and Bayesian phylodynamic analyses to comprehensively describe the evolutionary dynamics of the HIV-1 epidemic in Ethiopia. We used Bayesian phylodynamic models to estimate the dynamics of the effective population size ($N_e$) and reproductive numbers ($R_e$) through time for the HIV epidemic in Ethiopia.

**Results:** Our analysis revealed that the Ethiopian HIV-1 epidemic originated from two independent introductions at the beginning of the 1970s and 1980s from eastern and southern African countries, respectively, followed by epidemic growth reaching its maximum in the early 1990s. We identified three large clusters with a majority of Ethiopian sequences. Phylodynamic analyses revealed that all three clusters were characterized by high transmission rates during the early epidemic, followed by a decline in HIV-1 transmissions after 1990. $R_e$ was high (4–6) during the earlier time of the epidemic but dropped significantly and remained low ($R_e < 1$) after the mid-1990. Similarly, with an expected shift in time, the effective population size ($N_e$) steadily increased until the beginning of 2000, followed by a decline and stabilization until recent years. The phylodynamic analyses corroborated the modeled UNAIDS incidence and prevalence estimates.

**Conclusion:** The rapid decline in the HIV epidemic took place a decade before introducing antiretroviral therapy in Ethiopia and coincided with early behavioral, preventive, and awareness interventions implemented in the country. Our findings highlight the importance of behavioral interventions and antiretroviral therapy scale-up to halt and maintain HIV transmissions at low levels ($R_e < 1$). The phylodynamic analyses provide epidemiological insights not directly available using standard surveillance and may inform the adjustment of public health strategies in HIV prevention in Ethiopia.

**Keywords: effective reproductive number, effective population size, birth–death model, phylodynamic, HIV-1 epidemic, transmission cluster, behavioral intervention, Ethiopia**

## INTRODUCTION

The human immunodeficiency virus type 1 (HIV-1) is one of the most devastating infectious diseases in human history (UNAIDS, 2020). At the end of 2020, an estimated 38 million people were living with HIV/AIDS worldwide. Sub-Saharan Africa, the region where HIV-1 emerged during the 1920s, remains the most affected region, accounting for close to 70% of people living with HIV worldwide (Faria et al., 2014; UNAIDS, 2020). Despite the large-scale roll-out of antiretroviral treatment (ART), HIV incidence remains high, mainly in sub-Saharan Africa (UNAIDS, 2020). Ethiopia is one of the many sub-Saharan countries that was severely affected by the HIV epidemic.

HIV-1 is classified into four phylogenetically distinct groups: M (main), N (non-M, non-O), O (outlier), and P (pending), each representing different zoonotic cross-species transmissions of simian immunodeficiency viruses from non-human primates to humans (Sharp and Hahn, 2011; Faria et al., 2014; Giovanetti et al., 2020). Group M is the most prevalent, accounts for more than 95% of all the HIV-1 infections and is divided into 10 subtypes (A–D, F–H, and J–L), more than 102 different circulating recombinant forms (CRFs), and numerous unique recombinant forms (URFs; Hemelaar et al., 2019; Giovanetti et al., 2020). Subtype C is currently the dominant HIV-1 subtype and is responsible for nearly half all HIV-1 infections globally (Hemelaar et al., 2019). Although found worldwide, no official assignment of subtype C strains into phylogenetic sub-subtypes has been made. However, several distinct genetic clades associated with geography have been defined, the southern African clades (C-SA) and the eastern African clade (C-EA). Strains of the C-EA clade and a sub-clade of C-SA, termed C′-ET, are most prevalent in Ethiopia (Thomson and Fernandez-Garcia, 2011; Arimide et al., 2018).

The first HIV-1 infection and AIDS case report in Ethiopia was in 1984 and 1986, respectively (Lester et al., 1988; Tsega et al., 1988). Initially, the epidemic was concentrated to urban areas and along major commercial routes. Serology surveys revealed high prevalence (17%–55%) among risk populations (e.g., female sex workers: FSWs, long-distance truck drivers: LDTD, and soldiers; Mehret, 1990; Mebret et al., 1990). However, after introduction of antiretroviral therapy (ART) in public health care in 2005, the prevalence among the general population decreased and stabilized at significantly lower levels while the prevalence remained high in risk populations (EPHI, 2014).

The HIV epidemic in Ethiopia is considered a generalized epidemic with heterosexual transmission being the dominant mode of transmission (Kebede et al., 2000). Since 1985, Ethiopia has implemented several community-based HIV prevention programs to improve knowledge about the infection and mode of transmission, and interventions to reduce engagement in risk behavior (Mebret et al., 1990; Okubagzhi and Singh, 2002). However, the epidemiological dynamics and their correlations with introduction of various HIV prevention and interventions programs have not been characterized.

Similar to many low-income countries, epidemiological data regarding HIV from Ethiopia are sparse and incomplete, making surveillance of the HIV epidemic challenging. The increased availability of HIV genetic sequencing data and the development of phylogenetic and phylodynamic tools has enabled the use of molecular epidemiology analysis to describe the transmission dynamics and evolutionary history of HIV (Yusim et al., 2001; Delatorre and Bello, 2012; Mir et al., 2018; Vasylyeva et al., 2019). Previous studies in Ethiopia have shown that subtype C dominates the Ethiopian HIV epidemic and have provided valuable insight into HIV genetic diversity, its origins, and epidemic dynamics, but are limited in study participant numbers and geographic and temporal representation (Abebe et al., 2001a,b; Pollakis et al., 2003; Tully and Wood, 2010; Delatorre and Bello, 2012; Mir et al., 2018). Here, we used HIV-1 subtype C *pol* gene sequences collected from different regions of Ethiopia between 1986 and 2017. We employed state-of-the-art phylogenetic and phylodynamic methods, including both Bayesian coalescent and birth–death modeling, to elucidate evolutionary trajectories and temporal dynamics of the HIV-1 epidemic in Ethiopia.

## MATERIALS AND METHODS

### Baseline HIV-1 Drug Resistance Survey

We conducted a prospective HIVDR survey among antiretroviral-naïve adults in St. Paul General Specialized Hospital located in Addis Ababa, Ethiopia, in 2011. We performed the study according to the WHO-recommended survey methodology (Jordan et al., 2008). Treatment-naïve adults (>18 years) eligible to start ART at the St. Paul Generalized Specialized Hospital were consecutively enrolled. Whole blood specimens were collected and transported to the Ethiopia

Public Health Institute (EPHI), the national HIV laboratory, and WHO-accredited laboratory for viral load testing and HIVDR genotyping.

HIV genotyping was done using an in-house assay as described previously (Arimide et al., 2018). Briefly, a 1,084 base-pair fragment of HIV-1 *pol* (corresponding to positions 2,243–3,326 of HXB2; GenBank Accession Number: K03455) comprising amino acids 6–99 of the protease and 1–251 of the reverse transcriptase was obtained by RT-PCR and nested PCR. The purified PCR fragments were then sequenced and analyzed on the ABI 3500xl Genetic Analyzer (Applied Biosystems, Foster City, CA, United States). Sequence assembly and editing were performed using the RECall V 2.0 HIV-1 sequencing analysis tool (University of British Columbia, Vancouver, Canada; Woods et al., 2012). All sequences reported in this study have been deposited in GenBank under Accession Numbers OL598713-OL598856.

## Study Population and Sequence Dataset

We used the dataset of newly sequenced HIV-1 *pol* sequences and retrieved all publicly available Ethiopian HIV-1 subtype C *pol* sequences (matching pos. 2,243–3,326 relative of HXB2) from the Los Alamos National Laboratory (LANL) HIV Sequence database[1] (Date of access, December 2019). The quality of HIV-1 sequences was verified using the online Quality Control program of the LANL HIV sequence database (see Footnote 1) and sequences with stop codons, frameshifts, and poor quality were removed. We retained only one sequence per patient and selected the earliest sequence for patients with multiple sequences.

We removed duplicate sequences and sequences with potential contamination using the ElimDupes online tool from LANL. Moreover, to identify Ethiopian country-specific transmission clusters, we included a dataset of similar sequences from GenBank by identifying the 10 genetically closest GenBank sequences with BLAST for each Ethiopian HIV-1 subtype C sequence (Altschul et al., 1990; Mount, 2007). We only included sequences of 950 nucleotides or longer with known isolation dates and country of isolation in the analysis, since this 950-bp region has sufficient signal to reconstruct transmission links among infected individuals (Hué et al., 2004).

## HIV-1 Subtyping

Initial explorative HIV-1 subtyping was performed using the online automated subtyping tools REGA v3.0 (Pineda-Peña et al., 2013), COMET v2.2 (Struck et al., 2014), and RIP (Martin et al., 2010). Putative intra-subtype recombinant sequences were detected using jpHMM (jumping profile Hidden Markov Model)[2] (Schultz et al., 2009; Arimide et al., 2018). Only non-recombinant sequences were used for the analysis. Final subtyping was determined by maximum likelihood (ML) phylogenetic tree analysis with subtype reference sequences (Arimide et al., 2018).

## Maximum Likelihood Phylogenetic Analyses

A multiple sequence alignment was obtained using MAFFT V. 7 (Katoh and Standley, 2013) and was then manually edited using BioEdit V7.0.9.0 (Hall, 1999) until a non-redundant codon alignment was obtained. To avoid the effect of drug-induced convergent evolution, positions of identified mutations causing or contributing to HIVDR were removed from the alignment, resulting in a final alignment of 909 bp (Wensing et al., 2016).

The initial ML phylogenetic tree was constructed using an online version of PhyML (Guindon et al., 2010) under the GTR + I + Γ4 (general time-reversible nucleotide substitution model using the estimated proportion of invariable sites and four gamma categories). Heuristic tree search was performed using the SPR branch-swapping algorithm. Branch support was determined with aLRT-SH (approximate likelihood ratio test Shimodaira–Hasegawa-like) implemented in PhyML (Guindon et al., 2010). A branch in the phylogeny with an aLRT-SH value ≥0.9 was considered significant (Guindon et al., 2010; Esbjörnsson et al., 2016). The ML trees were visualized using FigTree v1.4.3 (Rambaut, 2016).

Our initial ML phylogenetic trees were constructed using the combined dataset of all Ethiopian sequences and sequences from the BLAST search. To comprehensively describe the HIV-1 subtype C circulating in Ethiopia, the dataset was divided into two based on phylogenetic branch support, the C-EA and C-ET clades.

## Analysis of Transmission Clusters

Separate transmission cluster analysis was performed for the two data sets using the ML phylogenetic analysis implemented by IQ-TREE under GTR + I + Γ4 as selected as the best fitting substitution model for the dataset using jModelTest v2.1.7 and with 1,000 replicates for the aLRT-SH test (Nguyen et al., 2015). Clusters with an aLRT-SH support ≥0.9 were considered significant (Guindon et al., 2010; Esbjörnsson et al., 2016). A transmission cluster was defined as a cluster in the ML phylogeny from root to tips (Esbjörnsson et al., 2016; Hassan et al., 2017; Sallam et al., 2017; Arimide et al., 2018). Clusters with an aLRT-SH-support of ≥0.9 that had a majority (at least 80%) of Ethiopia sequences were considered an Ethiopian transmission cluster. Transmission clusters were also defined based on their sizes (number of sequences/cluster), into dyads (two sequences), medium-sized clusters/networks (3–14 sequences), and large clusters (≥15 sequences; Aldous et al., 2012; Esbjörnsson et al., 2016).

To determine whether there was phylogenetic clustering by geographic region, viral sequences were grouped into six geographic regions (sequence collection location). The strength of association between the geographic location and the phylogeny was determined using two phylogeny–trait association statistics, the parsimony score (PS) and the association index (AI) tests, both of which were implemented in the Bayesian Tip-association Significance testing (BaTS) program (Parker et al., 2008). A significance level of $p < 0.05$ was used in both statistics.

[1] http://www.hiv.lanl.gov

[2] http://jphmm.gobics.de/submission_hiv

## Estimating Temporal Signal

For each cluster, we assessed the temporal signal of the data sets by performing root-to-tip genetic distance using TempEst (Rambaut et al., 2016). Clusters that had a positive correlation between genetic diversity and time were considered for further analysis.

## Estimating Viral Phylodynamic History

The birth–death skyline model (BDSKY; Stadler et al., 2012, 2013) implemented in BEAST2 v 2.6.2 was used to quantify epidemic growth through time described by changes in the effective reproductive number ($R_e$) which is the average number of secondary infections from an infected individual at any given time during the epidemic (Bouckaert et al., 2019; Vasylyeva et al., 2019). We used a lognormal distribution prior, LogNorm (0,1), for the effective reproductive number with the upper bound of 10, and a LogNorm (0,1) prior for the become uninfectious rate ($\delta$) in units per year (i.e., the inverse of the time duration of being infectious in a unit of years). We used $\delta = 0.2$, corresponding to a 5-year duration of the infectious period, as the mean of the distribution. In order to account for the uneven number of sequences per year, we employed a different sampling probability (*rho*) prior for each year, using a *beta* distribution with a mean equal to the number of samples divided by the reported number of HIV cases in the country for that year. We estimated the change in $R_e$ for six equally spaced intervals between the time to most recent common ancestor (tMRCA) and the most recent sampling year.

Phylodynamic analyses were also performed using the Bayesian Skygrid coalescent tree prior, implemented in BEAST 1.10,4 (Gill et al., 2013; Suchard et al., 2018; Hill and Baele, 2019), to estimate changes in effective population size ($N_e$) through time and estimate the population growth rates ($r$, years$^{-1}$) by using a logistic growth coalescent tree prior. Analyses were performed using the GTR + I + Γ4 nucleotide substitution model. The temporal scale of the evolutionary process was estimated using a relaxed uncorrelated molecular clock model with an underlying lognormal distribution with normal priors. This allowed the estimation of the evolutionary rate ($\mu$, nucleotide substitutions per site per year, s/s/y), the age of the most recent common ancestor (tMRCA, years), and the phylodynamic parameters.

For each of the two phylodynamic approaches, we ran three independent Markov Chain Monte Carlo (MCMC) chains until all associated parameters converged to ensure good mixing (ESS > 200) after discarding the first 10% of the MCMC chains. The convergence of the MCMC was inspected visually and by calculating the ESS for each parameter using Tracer v 1.7.5 (Rambaut et al., 2018). We used LogCombiner to combine the different independent results (log and corresponding tree file) from the multiple chains (Drummond et al., 2012). We used the bdskytools package[3] in R to plot the results of the BDSKY analysis.

## Ethical Approval

We obtained ethical approval from the Research and Ethical Clearance Committee of EPHI and the National Health Research Ethics Review Committee of the Ministry of Science and Technology of Ethiopia. All participants for the baseline HIV drug resistance survey provided written informed consent to participate in the study.

# RESULTS

## Study Population and Initial Phylogenetic Analysis

We retrieved 1,132 Ethiopian *pol* sequences from LANL, collected from different parts of the country from 1986 to 2017. Additionally, we included 144 HIV-1 *pol* sequences from the baseline HIVDR survey. The combined dataset ($n = 1{,}276$ sequences) contained 399 putative recombinant sequences, which were removed from further analysis. We further included a dataset of similar sequences from GenBank by identifying the 10 genetically closest GenBank sequences with BLAST for each of the 877 non-recombinant Ethiopian HIV-1 subtype C sequences in the study. The final combined dataset contained 1,333 non-recombinant HIV-1 subtype C *pol* sequences (877 Ethiopian and 456 global), which were used for phylogenetic analysis. The ML phylogenetic tree identified two distinct and well-supported clades, the C-EA and C′-ET clades (**Figure 1**). Among the 877 Ethiopian sequences included in the analysis, the C-EA clade represented 567 (65.0%) of the sequences, while 310 (35.0%) belonged to the C′-ET clade.

Most of the Ethiopian C-EA sequences were found in one large cluster (aLRT = 0.87), and only 33 Ethiopian C-EA sequences fell outside this cluster. Sequences of the global dataset intermixed with the Ethiopian sequences and represented sequences obtained most frequently ($N = 67$, 33.3%) in other East African countries, North America, and Europe. Sequences from Burundi dominated the basally located sequences. In the case of the second major clade, the majority of the Ethiopian C′-ET sequences (95.5%) formed a well-supported sub-clade (aLRT = 0.92), branching off from the basally located sequences. Southern African countries' sequences were intermixed ($N = 121$, 47.1%) with the Ethiopian C′-ET sequences, but they were most prominent at the base of the clade.

## Transmission Cluster Analysis

We inferred transmission clusters by separate ML phylogenetic tree analyses of the two clades (C-EA and C′-ET). For C-EA, the ML phylogenetic tree contained a total of 810 sequences (with reference sequences) and identified two large well-supported clusters of 259 (C-EA-259) and 148 sequences (C-EA-148). The C-EA-259 cluster (aLRT = 0.93) contained 213 Ethiopian (82.2% of the sequences of the cluster) and 46 non-Ethiopian sequences, collected 1988–2017, and the C-EA-148 cluster (aLRT = 0.95) contained 124 Ethiopian (83.8%) sequences and 24 non-Ethiopian sequences, collected 1996–2017 (**Figure 2**). Moreover, we identified six networks (medium-sized clusters) and nine dyads.

[3] https://github.com/laduplessis/bdskytools

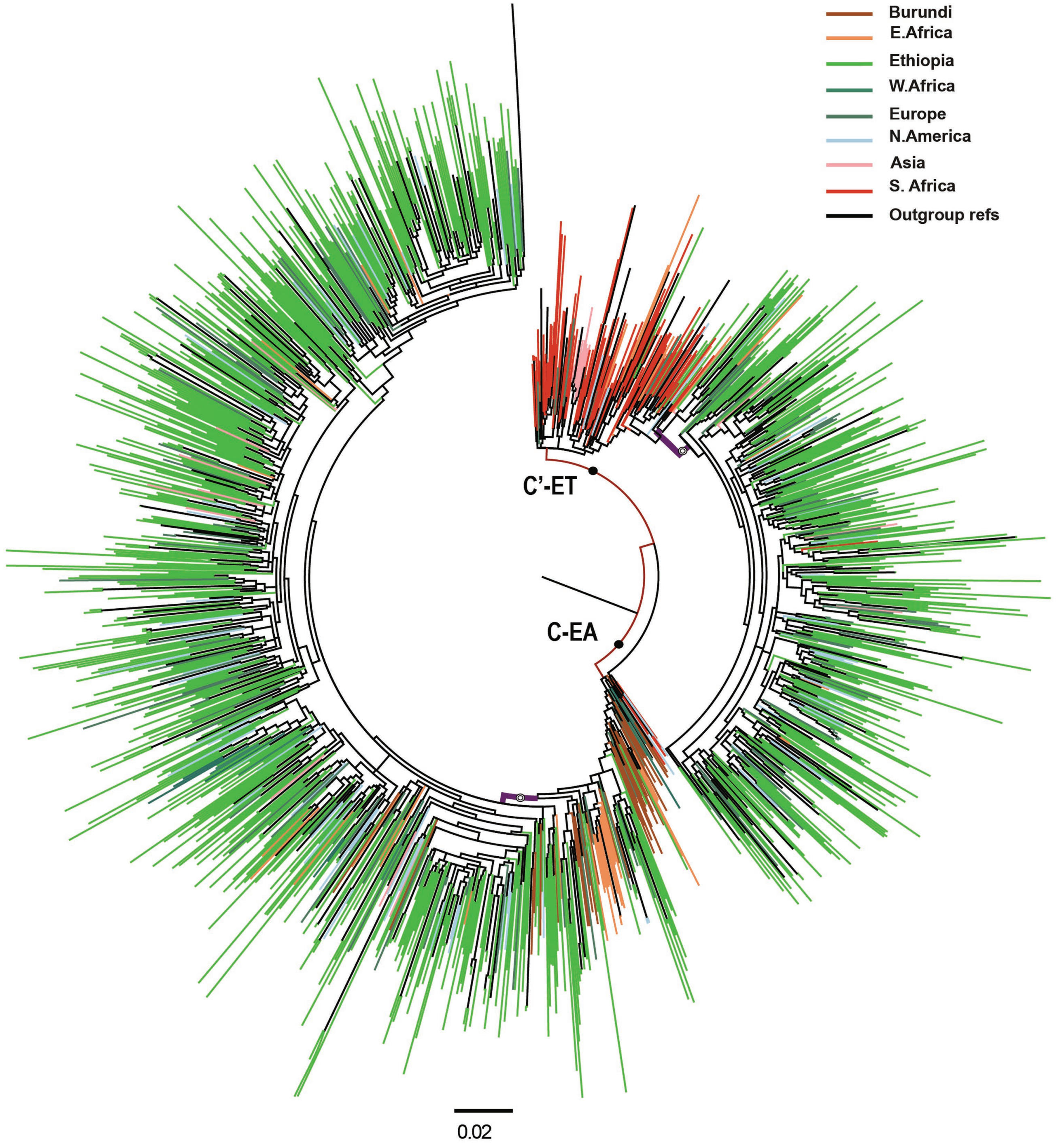


**FIGURE 1** | Maximum likelihood (ML) phylogenetic tree of HIV-1 subtype C pol sequences ($n = 1,333$). Maximum likelihood phylogenetic tree constructed using 877 Ethiopian subtype C *pol* sequences collected 1986–2017 and 456 global subtype C *pol* sequences. Colored tips are according to the geographic origin of sequences, as indicated in the legend in the top right corner. Branches defining the major clades (C-EA and C′-ET) are those indicated with a filled circle, having branch support (aLRT-SH) >0.9. The Ethiopian clusters are indicated by open circles, having branch support (aLRT-SH) >0.9. The scale bar represents 0.02 substitutions/site.

Among the 580 C′-ET sequences, we identified one well-supported (aLRT = 0.96) large cluster containing 153 sequences [C′-ET-153; 124 Ethiopian (81.0% of the sequences in the cluster)], and 29 non-Ethiopian sequences collected 1995–2017 (**Figure 3**). Since we had no information about the associated risk behavior of the respective individuals, we could not associate cluster

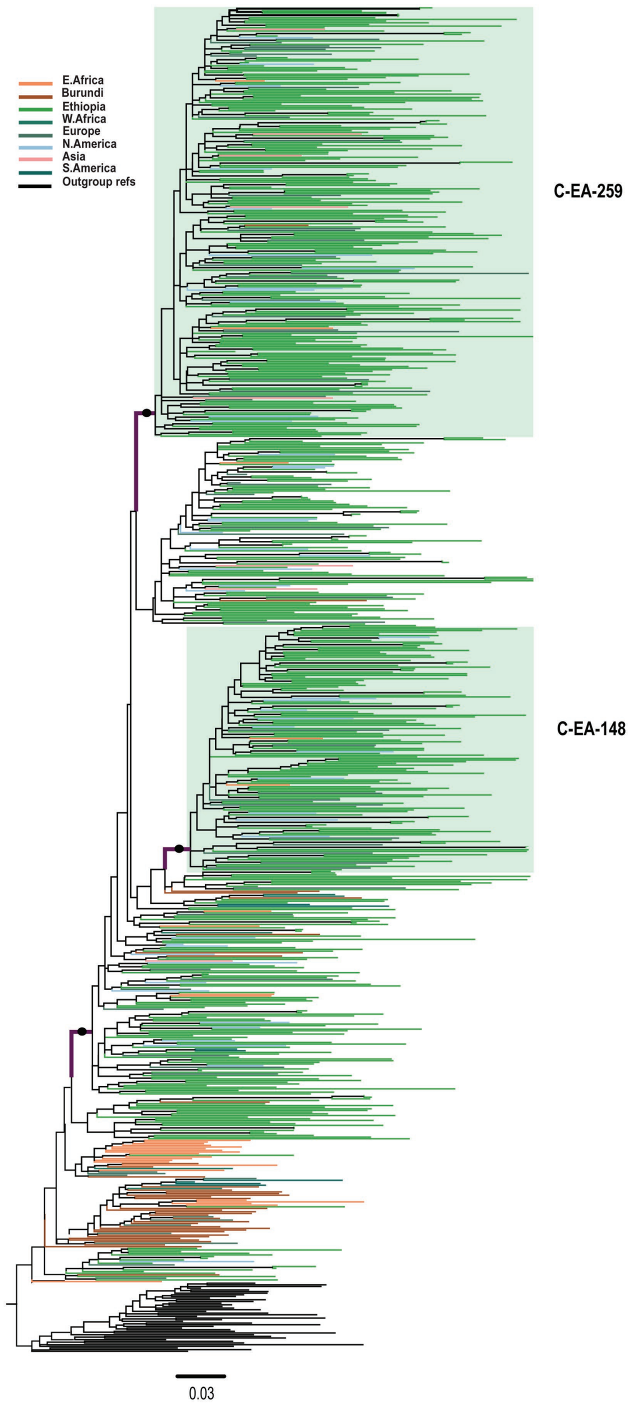


**FIGURE 2 |** Maximum likelihood phylogenetic tree of HIV-1 C-EA clade *pol* sequences ($n = 810$). Maximum likelihood phylogenetic tree constructed using 567 Ethiopian subtype C *pol* sequences collected between 1986 and 2017, 201 global subtype C *pol* sequences, and 42 reference sequences. Colored tips are according to the geographic origin of sequences, as indicated in the legend in the top left corner. The C-EA-259 and C-EA-148 clusters are highlighted in green, corresponding to a branch support aLRT-SH >0.9 and >80% Ethiopian sequence. The filled circle defining the C′-ET clade represents an aLRT-SH >0.9. The tree was rooted using the C′-ET reference sequences. The scale bar represents 0.03 nucleotide substitutions per site.

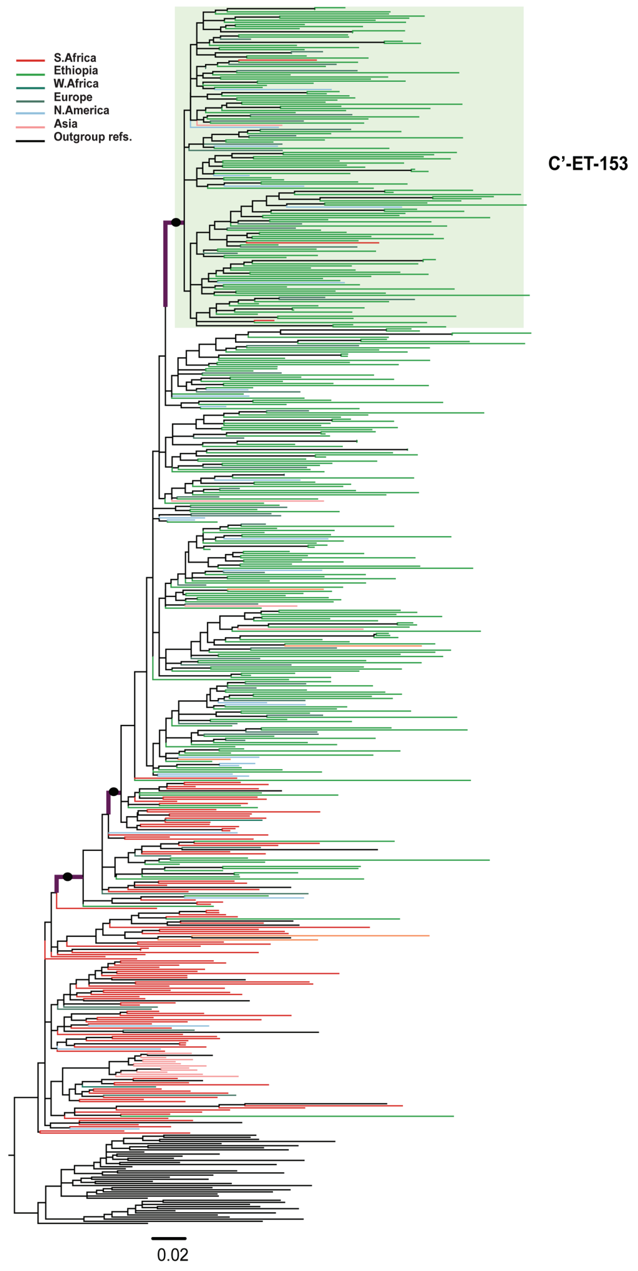


**FIGURE 3 |** Maximum likelihood phylogenetic tree of HIV-1 C′-ET clade *pol* sequences ($n = 580$). Maximum likelihood phylogenetic tree constructed using 310 Ethiopian subtype C *pol* sequences collected 1986–2017, 227 global subtype C *pol* sequences, and 43 reference sequences. Colored tips are according to the geographic origin of sequences, as indicated in the legend in the top left corner. The C′-ET-153 cluster, highlighted in a green shade, is defined by branch support (aLRT-SH) of >0.9 and >80% Ethiopian sequences. The branches with filled circles possibly defining the C′-ET clade have branch support (aLRT-SH) >0.9. The tree was rooted using the C-EA reference sequences. The scale bar represents 0.02 nucleotide substitutions per site.

formation with risk behavior. However, geographic location was available for analysis but was not associated with cluster formation.

## Evolutionary Rates and Dates of HIV-1 Subtype C in Ethiopia

To comprehensively describe the Ethiopian HIV-1 epidemic, we further performed analyses on the three large clusters (C-EA-259, C-EA-148, and C′-ET-153). Root-to-tip analysis indicated a temporal signal in the three data sets (correlation coefficient of 0.53, 0.43, 0.47 for C-EA-259, C-EA-148, and C′-ET-153, respectively).

First, we estimated the tMRCA of the transmission clusters and the C-EA clade. The clade tMRCA represents the date of the origin of the circulating subtype C clade in the region. In contrast, the estimated tMRCA of the Ethiopian transmission clusters should approximate the introductions and local spread of the viral strains in the country (Dalai et al., 2009; Esbjornsson et al., 2011). Based on the inferred tMRCA, the posterior median estimates of C-EA-259 (1975, 95% HPD: 1970–1979) and C-EA-148 (1976, 95% HPD: 1963–1985) were older than the estimate for the C′-ET transmission cluster (1983, 95% HPD: 1975–1988; **Table 1**). The median root tMRCAs for C-EA clade was estimated to be 1971 (95% HPD: 1966–1976).

**TABLE 1** | Population dynamics and evolutionary estimates for subtype C cluster in Ethiopia.

| | Subtype C clade/Cluster | | |
|---|---|---|---|
| | C-EA-259 | C-EA-148 | C′-ET-153 |
| Sequences from Ethiopia (n) | 213 | 124 | 124 |
| Range of collection (year) | 1988–2017 | 1996–2017 | 1995–2017 |
| Mean coefficient of variation | 0.24 | 0.23 | 0.33 |
| Median evolutionary substitution rate (95% HPD)[1] | 1.76 (1.49–2.00) | 1.74 (1.11–2.40) | 1.83 (1.33–2.32) |
| Median year of tMRCA (95% HPD) | 1975 (1970–1979) | 1976 (1963–1985) | 1983 (1975–1988) |
| Median rate of population growth (95% HPD)[2] | 0.66 (0.51–0.81) | 0.61 (0.38–0.86) | 0.80 (0.53–1.10) |
| Median epidemic doubling time (years) (95% HPD)[3] | 1.05 (0.86–1.36) | 1.12 (0.81–1.82) | 0.86 (0.63–1.31) |
| Maximum effective reproductive number ($R_e$)[4] | 6.13 (95% HPD, 3.53–10.14) | 3.93 (95% HPD, 1.88–7.07) | 4.88 (95% HPD, 2.57–8.57) |
| Basic reproductive number ($R_0$)[5] | 4.30 (95% HPD: 3.55–4.05) | 4.05 (95% HPD: 2.90–5.25) | 5.00 (95% HPD: 3.65–6.50) |
| Median become uninfectious rate (95% HPD)[6] | 0.13 (0.06, 0.20) | 0.19 (0.08, 0.29) | 0.21 (0.09, 0.33) |

[1] *Median number of substitutions/site/year × $10^{-3}$.*

[2] *Median population growth rate (r) per year, determined in BEAST v1.10.4 using a logistic tree prior.*

[3] *The time (years) required to double the effective number of infections (λ), calculated as $\lambda = \ln(2)/r$, where r is the population growth rate.*

[4] *$R_e$ (effective reproductive number) which reflect the average number of secondary infections from an infected individual at any given time during the epidemic.*

[5] *Basic reproductive number ($R_0$) which reflects the average number of infections generated by an infected individual in a population where all individuals are susceptible to infection calculated by using the formula $R_0 = rD + 1$(56) (where r is the population growth rate and D is the average duration of infectiousness period).*

[6] *Become uninfectious rate, which reflects the inverse of the time duration of being infectious, in the unit of years.*

The median estimated evolutionary rate was in the range $1.76–1.82 \times 10^{-3}$ substitutions/site/year for the three clusters, with overlapping 95% HPD intervals (**Table 1**), indicating no significant difference of evolutionary rates among the three clusters.

## Temporal Dynamics of Viral Transmission

Direct estimation of the temporal dynamics of the effective reproductive number, $R_e$, was performed using the BDSKY model. The BDSKY analysis assumed a piecewise constant $R_e$, changing over six equidistance intervals between the tMRCA and the most recent sampling. The $R_e$ showed similar dynamics for the three clusters. From the start until the beginning of the 1990s, $R_e$ remained consistently high ($R_e > 1$) and dropped below the epidemiological threshold ($R_e < 1$) at the mid-1990s and remained below one till recent years (**Figures 4A–C**). In all three clusters, we observed the maximum $R_e$ values (4-6) during the early period (before 1990) of the epidemic (**Table 1**).

The posterior median estimates of the become non-infectious rate obtained for each cluster were of 0.13 (95% HPD: 0.06–0.20) for C-EA-259, 0.19 (95% HPD: 0.08–0.29) for C-EA-148, and 0.21 (95% HPD: 0.09–0.33) for C′-ET-153, which translates to an infectious period of ~8 years for C-EA-259 and ~5 years for the other two clusters. Despite the longer infectious period estimated for C-EA-259, the overlapping HPD indicates no significant differences among clusters.

We performed different sensitivity analyses to explore the robustness of our BDSKY estimates. We used different values for the mean of the become non-infectious rate prior (δ) going from 6 months to 10 years (*δ*: 2, 1, 0.5, 0.2, 0.125, and 0.1). We also performed the sensitivity analysis by estimating the effective reproductive number for six and ten equally spaced intervals between tMRCA and the most recent sample. We obtained similar results for all analyses. $R_e$ was consistently >1 for the early period until the early 1990s, followed by a decline to $R_e < 1$ after the mid-1990s.

We further performed phylodynamic analysis using the Bayesian Skygrid model to estimate the temporal characteristics of the HIV-1 epidemic in Ethiopia. We analyzed the three Ethiopian clusters and estimated the change in the effective population size ($N_e$) through time, representing the change in the total number of infections contributing to new cases. The Bayesian skygrid inference revealed a rapid increase in $N_e$ for all the three clusters from the initial introduction period until shortly before the year 2000, followed by a decline and stabilization in $N_e$ until recent years (**Figures 4D–F**). We also determined the population growth rate (*r*), the rate of increase in the effective population size with time, using the logistic growth model of the coalescent parametric model. The median growth rate was 0.66, 0.61, and 0.80 year$^{-1}$ for clusters C-EA-259, C-EA-148, and C′-ET-153, respectively, with overlapping HPD intervals (**Table 1**).

We also estimated the mean coalescent-based basic reproductive number ($R_0$) values for each cluster from the logistic growth model using the formula $R_0 = rD + 1$ (Pybus et al., 2001; where r is the population growth rate and D is the average duration of infectiousness period). Assuming an average infectious period

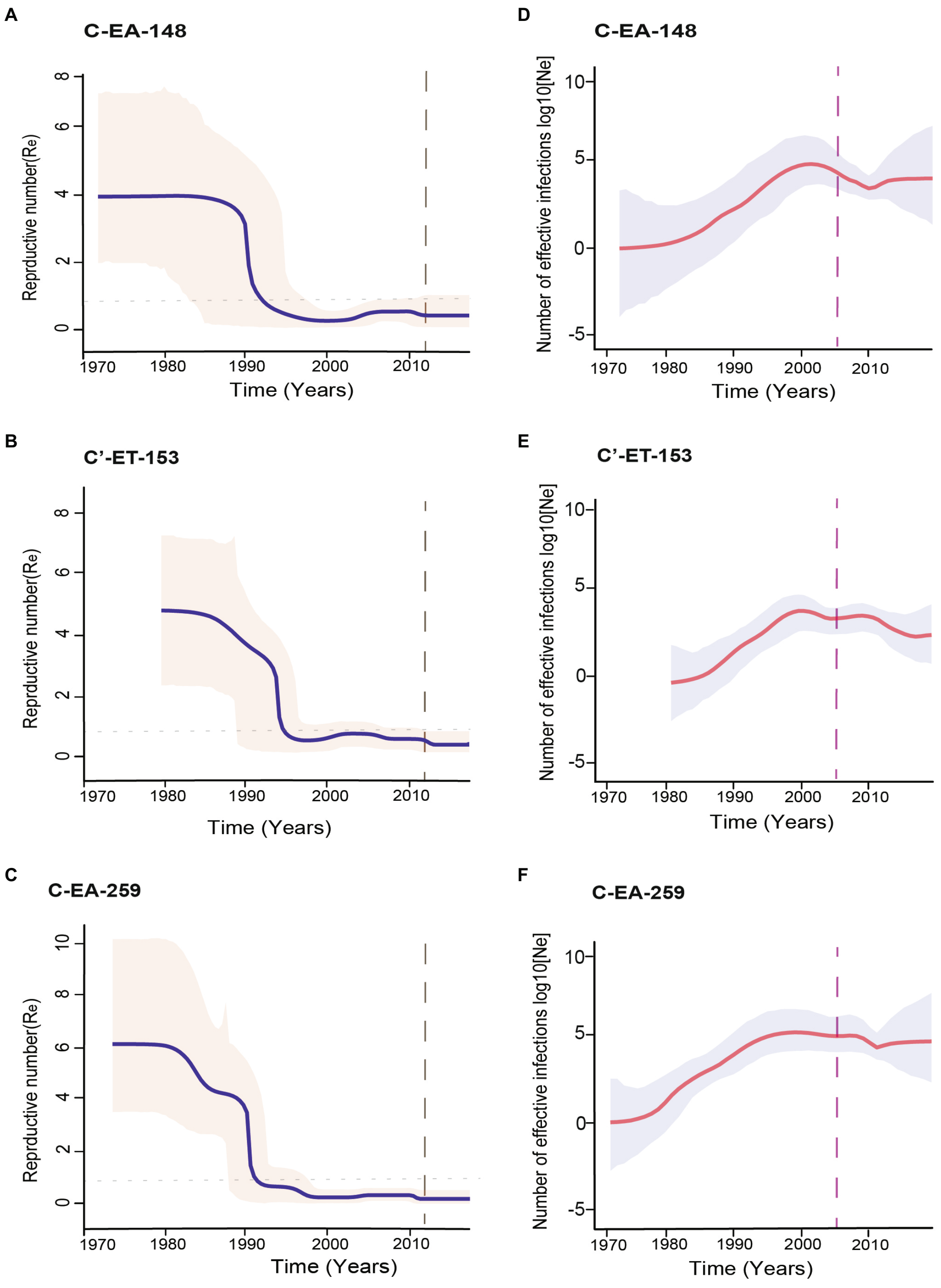


**FIGURE 4 |** Population dynamics of the HIV-1 epidemic in Ethiopia using the three major clusters (C-EA-259, C-EA-148, and C'ET-153). **(A–C)** The temporal dynamics of effective reproductive number ($R_e$) using the Bayesian birth–death model, the median $R_e$ are shown by the continuous blue line, and indicated in a pink shade is the 95% highest probability density (HPD) intervals. The gray dashed line indicates the last coalescent event reported by the lineage through time (LTT) analysis. The horizontally dotted line represents the epidemiological threshold (Re = 1). **(D–F)** The median estimates of the effective population size ($N_e$) over time using the Bayesian skygrid model. The red line shows the median logarithmic effective population size ($N_e$) over viral generation time (*t*), representing effective transmissions, and the gray shade indicates the 95% highest probability density (HPD) intervals. The pink dashed line represents the time of antiretroviral therapy (ART) introduction in Ethiopia.

of 5 years, $R_0$ was in the range 4.0–5.0 for the three clusters, all with overlapping 95% HPD intervals (**Table 1**).

## DISCUSSION

In this study, we analyzed a large dataset of HIV-1 *pol* sequences collected from different regions of Ethiopia during more than 30 years. We used both Bayesian coalescent and birth–death models to characterize the dynamics of the HIV-1 epidemic in the country. Overall, our analysis confirms that strains of two subtype C clades are circulating in Ethiopia, supporting the hypothesis that the HIV-1 epidemic in Ethiopia is the result of at least two independent HIV-1 introductions from eastern and southern African countries (Delatorre and Bello, 2012). Moreover, the phylodynamic analyses revealed that the epidemic dynamics in Ethiopia were characterized by an expanding epidemic growth from the start of the epidemic until the mid-1990s, followed by a sharp decline in HIV-1 transmissions. The decline in $R_e$ occurred many years before introducing ART and coincided with early behavioral, preventive interventions, and public health awareness campaigns implemented in Ethiopia.

$R_e$ is a proxy for HIV incidence and describes the transmission dynamics; $R_e > 1$ means that the epidemic is growing, $R_e < 1$ shows the epidemic is declining, while $R_e = 1$ shows that the epidemic is stabilizing (Stadler et al., 2013). Our phylodynamic analysis showed that the three clusters followed similar epidemic trends. The BDSKY model indicated epidemic growth ($R_e > 1$) from the 1970s to the early 1990s. The basic reproductive number ($R_0$) and mean initial $R_e$ were comparably high for the clusters, indicating an early exponential epidemic growth. Similarly, a high epidemic growth rate was estimated for each cluster (0.61–0.80 $\text{year}^{-1}$) and a steady increase in $N_e$ until the beginning of 2000, highlighting the upward trend of HIV transmissions in Ethiopia during the period.

The exponential epidemic growth observed in our analyses is consistent with retrospective serological data, which showed a massive increase of HIV infections among risk populations in Addis Ababa and cities along the main trading routes in Ethiopia during this early period. An extensive survey on FSWs operating in the main trading routes of Ethiopia in 1988 reported an HIV-1 prevalence between 5.3% and 38.1% (Mehret et al., 1990c). Studies performed in the capital Addis Ababa, 1988–90, showed an increase in prevalence from 25% to 54%, and 13% to 18% among FSWs and LDTDs, respectively (Khodakevich et al., 1990; Mehret et al., 1990a,c; Kebede et al., 2000), and 12%–18% among soldiers 1990–1993. Similarly, an increase in HIV prevalence among pregnant women attending antenatal care clinics (ANC) in Addis Ababa (4.6%–10.5%, 1989–90; Kebede et al., 2000) indicated extensive spread in the population.

During the early years, the rapid epidemic increase was most likely due to lack of awareness of HIV, high mobility among FSWs, high-risk sexual behavior, high STI prevalence among the general population (Desta et al., 1990; Mehret et al., 1990b; Negassa et al., 1990), while no prevention interventions were in place. The increased population movement following considerable urbanization and political instability in the country during this early period might also have contributed to the high HIV prevalence and epidemic spread (Hladik et al., 2006; Esbjornsson et al., 2011).

Although there is a lack of data that can describe the HIV epidemic on a national scale, different studies have shown a decline in new infections since the mid-1990s, corroborating our results (Kebede et al., 2000; Tsegaye et al., 2002; Wolday et al., 2007). The decline in HIV prevalence among young adult women (15–24 years) represents a well-established indicator of epidemic decline. It measures the frequency of relatively recent infections and is less influenced by death (Tsegaye et al., 2002). The HIV prevalence trend among young women (15–24 years) attending ANCs in Addis Ababa between 1995 and 2003 declined significantly from 24.2% to 12.9% (Tsegaye et al., 2002; Wolday et al., 2007). Moreover, there was a sharp decline in HIV prevalence among young blood donors in Addis Ababa and nine other towns during this period (Kebede et al., 2000).

Due to a lack of comprehensive data, it has not been easy to obtain estimates of the national incidence trend in Ethiopia. However, a study done to assess the temporal trend among pregnant women who attended the ANCs in the capital Addis Ababa, assessing >7,000 serum specimens collected 1995–2003, showed a significant decline in the HIV-1 incidence rate (from 7.7% to 2.0%, 1995–2003; Wolday et al., 2007). The reduction was substantial among young ANC attendees (aged 15–19 years), indicating an epidemic decline (7.8% to 0.0%, 1995–2003; Wolday et al., 2007). A mathematical modeling study also demonstrated a substantial reduction in the HIV incidence in Ethiopia after 1995 with an estimated annual decline of 6.3% per year, resulting in a total decrease of 77% between 1990 and 2016 (Deribew et al., 2019).

The early decline in the HIV transmissions observed in our study and documented in serological surveys coincide with the change of sexual behavior, prevention, and better control of other sexually transmitted infections (STIs) achieved through the sustained public education and mobilization campaigns. Ethiopia was one of the first countries in sub-Saharan Africa to introduce a task force to prevent and control HIV/AIDS and STI infections, including a national plan for the HIV epidemic response intervention (Zewdie et al., 1990; Kebede et al., 2000; Kloos and Mariam, 2000; Okubagzhi and Singh, 2002). During the early 1990s, Ethiopia had implemented a wide range of HIV prevention and information programs. Implementation of several behavioral interventions and awareness programs took place using the national media, schools, and public gatherings (Hadgu et al., 1990; Zewdie et al., 1990). These programs mainly focused on sustained health education, risk reduction, condom promotion, and prevention and control of STIs (Zewdie et al., 1990; Okubagzhi and Singh, 2002).

The national survey data on behavioral risk factors in Ethiopia are limited. However, different program reviews (1989–1991) and two nationwide surveys on condom use (1987–1993) revealed that these interventions led to changes in sexual risk behavior and increased knowledge about HIV/AIDS. Moreover, the intervention increased condom use and substantially reduced both non-regular partner and STI (Mehret et al., 1996). Similarly, another study showed condom use increased, and non-regular partners decreased among high school students in Addis Ababa and Gondar in the period after 1990 (Kebede et al., 2000). Moreover, a study among male factory workers in Ethiopia showed a change in sexual risk behavior (Mekonnen et al., 2003). Although it is difficult to quantify the impact of the different interventions on HIV incidence, it is reasonable to assume that the various prevention programs impacted HIV transmissions.

Several other studies outside Ethiopia have reported a significant decline in HIV prevalence after behavioral interventions (Martin, 1987; Hessol et al., 1989; Nelson et al., 1996; Stoneburner and Low-Beer, 2004; Halperin et al., 2011). A study in Uganda and Zimbabwe showed a significant decline in HIV prevalence after 1990, resulting from public health intervention on reduced sexual risk behavior (Stoneburner and Low-Beer, 2004; Halperin et al., 2011). Similarly, behavioral interventions resulted in a substantial decrease in HIV transmissions among MSM in Europe and North America in the mid-1980s and heterosexuals in Thailand in the early 1990s (Martin, 1987; Hessol et al., 1989; Nelson et al., 1996; Hué et al., 2005). In line with our results, a comprehensive review of empirical and modeled HIV incidence trends across 20 countries in Sub-Saharan Africa, 1990–2012, revealed a decline in incidence commenced before introducing ART programs, highlighting the significance of behavioral intervention in reducing HIV transmissions (Taaffe et al., 2014).

The trends of the phylodynamic analyses (**Figure 4**) are in concordance with the UNAIDS HIV incidence and prevalence modeled estimates (**Figure 5**),[4] showing a high incidence and prevalence during the years before 1990–1995, followed by a decline in incidence and stabilization in prevalence.

Thus, our results align well with published serological and epidemiological trends in Ethiopia. The epidemic decline coincides with the timing of behavioral interventions in Ethiopia, suggesting a link between the early decline of HIV spread and behavioral interventions many years before the implementation of ART in the country. However, the introduction of ART, which has proved to successfully suppress HIV replication and reduce the risk of onward transmissions, has significantly contributed to reducing HIV transmission, mortality, and maintaining the epidemic decline (Cohen et al., 2011).

The phylogenetic analysis confirms that strains of two HIV-1 subtype C clades (C′-ET and C-EA) are circulating in Ethiopia, suggesting that the HIV epidemic in Ethiopia arose by at least two independent introductions of founder strains from the eastern and southern African countries, respectively (Pollakis et al., 2003; Delatorre and Bello, 2012). Previous studies have defined several distinct subtype C clades that, in most cases, are associated with geographical regions (Thomson and Fernandez-Garcia, 2011). In the case of the Ethiopian lineages, they represent a southern African clade (where the Ethiopian C-SA sub-clade named C′-ET is more or less confined to Ethiopia) and an eastern African clade (C-EA). Our findings align with previous studies showing a distinct phylogeographic subdivision of the HIV-1 subtype C circulating in east, central, and southern African countries (Abebe et al., 2000; Pollakis et al., 2003; Thomson and Fernandez-Garcia, 2011; Delatorre and Bello, 2012).

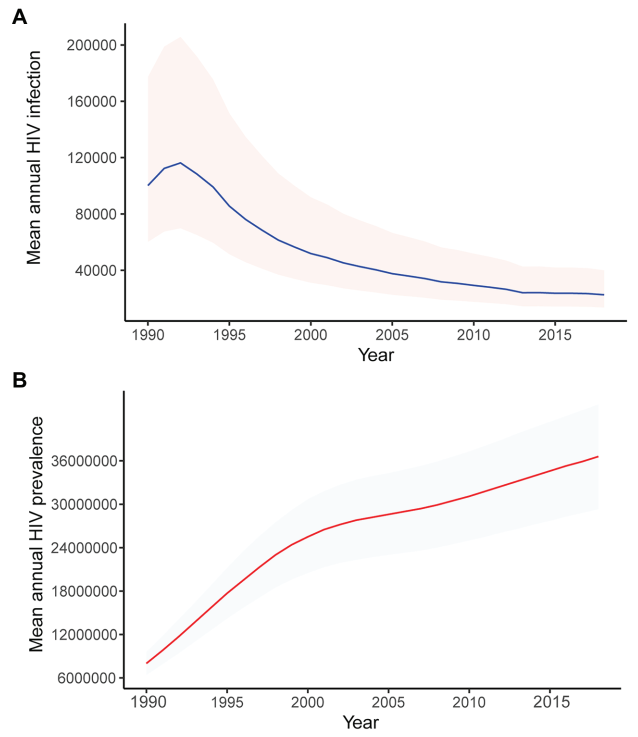


**FIGURE 5 |** Modeled mean annual HIV **(A)** incidence and **(B)** prevalence, 1990–2018. Data obtained from UNAIDS.

The transmission cluster analysis indicated that the Ethiopian sequences formed large clusters, indicative of a few major introductions or expansions in the country. The three Ethiopian clusters described here were mixed regarding collection sites, suggesting intermixing of the HIV epidemic in Ethiopia. Different socio-cultural and behavioral factors might also contribute to cluster formation, and assessing these factors are essential for designing HIV-1 transmission preventive strategies. However, the sequences obtained from the public database do not contain information on risk factors, sociodemographic, and other clinical information. Hence, further analysis on factors associated with cluster formation was not possible to discern in this study.

Sequences of Burundi dominated the basal root of the large monophyletic C-EA clade incorporating more than 90% of the Ethiopian sequences, which is in line with a previous study showing that the C-EA clade likely had its origin in Burundi (Delatorre and Bello, 2012). Moreover, the basal root of the monophyletic clade defining the C′-ET clade was dominated by sequences from southern African countries, possibly reflecting the origin of this clade from southern African countries. However, our analysis could not identify the exact countries. Interconnectivity between populations due to geographic proximity has been an essential factor for the spread of HIV across African countries (Wilkinson et al., 2015, 2016; Faria et al., 2019). However, the large distances and cultural interconnectivity between Ethiopia, Burundi, and southern African countries suggest that other factors were in play. Population

[4] https://aidsinfo.unaids.org

movements (due to unknown reasons) could have played a role in the introduction of HIV-1 subtype C to Ethiopia, similar to those observed in other parts of Africa (Gray et al., 2009).

Estimating the date of origin and timing of transmissions of HIV is essential to understanding the dynamics of HIV spread. Here, integral to our analysis of HIV transmission dynamics, we also obtained the tMRCA of the C-EA and C′-ET clades in Ethiopia. The molecular dating analysis suggested that the introduction of the C-EA clade took place more than a decade before the first reported AIDS case in Ethiopia. The dating is plausible considering that AIDS symptoms typically arise 6–10 years after infection. Moreover, our tMRCA estimates coincide with estimates of the introductions of the C-EA clade in other Eastern Africa countries, including Kenya, Tanzania, and Uganda (Delatorre and Bello, 2012), and are consistent with previous estimates for subtype C introduction in Ethiopia (Delatorre and Bello, 2012; Mir et al., 2018). Notably, this period also coincided with a large population migration from Burundi, which could have played a crucial role in disseminating the C-EA clade to Eastern Africa countries (Delatorre and Bello, 2012).

In contrast, the tMRCA of C′-ET was estimated at the beginning of the 1980s and is likely the result of a single introduction. This period coincided with the years of socio-political changes in the southern African countries and is associated with a steep growth of the HIV epidemic and viral migrations within southern African countries (Wilkinson et al., 2015, 2016).

To our knowledge, this study represents the most comprehensive study concerning the HIV epidemic in Ethiopia to date. It employs a large number of HIV-1 *pol* sequences collected during more than 30 years (1986–2017) from different geographical locations in Ethiopia. Moreover, we used state-of-art phylogenetic and phylodynamic methods to investigate the dynamics of the epidemic. Like many other molecular epidemiology studies, we incorporated HIV-1 *pol* gene sequences deposited in public databases in our analysis. As new HIV infections are recorded, more sequencing will allow to keep track of the ongoing transmission dynamics. The total sampling density was low, mainly due to a generally low sequencing coverage in Ethiopia, compared to the country's total number of infected individuals. Thus, the transmission clusters identified here cannot fully represent Ethiopia's entire HIV-1 transmission networks. Moreover, we based our analysis on HIV-1 *pol* sequences, representing the most sequenced HIV region due to the numerous published HIVDR studies. Although the HIV-1 *pol* fragment has sufficient phylogenetic signal for phylogenetic analysis of HIV (Hué et al., 2004), longer sequences, including whole genome sequences, may have provided a more informative inference of the HIV-1 molecular epidemiology and transmission history. Finally, the sequences used in this analysis lacked associated information, such as clinical, demographic, risk population assignment, or socio-economic data and, hence, we could not perform a detailed analysis of associated risk factors for HIV transmissions in our study.

In summary, we have employed state-of-art phylogenetic and phylodynamic approaches to describe the molecular epidemiology of HIV in Ethiopia. Our findings indicate that two distinct HIV subtype C strains were introduced in Ethiopia at the beginning of the 1970s and 1980s, followed by rapid epidemic growth until it started to decline in the mid-1990s, a decade before ART roll-out in Ethiopia. The sharp decline coincided with several behavioral prevention interventions and awareness campaigns. Our finding highlights the significance of scaling up behavioral and risk reduction interventions in addition to ART scale-up in the HIV/AIDS control strategy.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

## AUTHOR CONTRIBUTIONS

DA, PB, and PM conceived and designed the study. DA, YK, PM, and TB coordinated the laboratory tests. DA, LE-G, SS, DK, and PM conducted the phylogenetic and phylodynamic analysis and interpreted the results. DA and PM wrote the manuscript. All authors reviewed the draft and contributed important intellectual content to the final version. All authors agreed and approved to the published version of the manuscript.

## FUNDING

PM was supported by the Swedish Research Council (grant numbers 2019-05235 and 2020-02344), Österlund Foundation, and a donation to PM and PB through the Medical Faculty, Lund University, supported the research. LRE-G and DK were supported by the Max Planck Society. The computations were enabled by resources in project SNIC 2021-5-69 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

## ACKNOWLEDGMENTS

We express our deep appreciation to all patients from whom data were collected and analyzed and all staff involved in the HIV-1 drug resistance survey at the St. Paul General Specialized Hospital in Addis Ababa, Ethiopia. We also acknowledge the support given by the EPHI research laboratory team and St. Paul General Specialized Hospital and antiretroviral clinics staff.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.821006/full#supplementary-material

**Supplementary Table S1 |** The GenBank accession numbers and the year of sampling of sequences used for this study.

## REFERENCES

Abebe, A., Lukashov, V. V., Pollakis, G., Kliphuis, A., Fontanet, A. L., Goudsmit, J., et al. (2001a). Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. *AIDS* 15, 1555–1561. doi: 10.1097/00002030-200108170-00013

Abebe, A., Lukashov, V. V., Rinke De Wit, T. F., Fisseha, B., Tegbaru, B., Kliphuis, A., et al. (2001b). Timing of the introduction into Ethiopia of subcluster C' of HIV type 1 subtype C. *AIDS Res. Hum. Retrovir.* 17, 657–661. doi: 10.1089/088922201300119770

Abebe, A., Pollakis, G., Fontanet, A. L., Fisseha, B., Tegbaru, B., Kliphuis, A., et al. (2000). Identification of a genetic subcluster of HIV type 1 subtype C (C') widespread in Ethiopia. *AIDS Res. Hum. Retrovir.* 16, 1909–1914. doi: 10.1089/08892220050195865

Aldous, J. L., Pond, S. K., Poon, A., Jain, S., Qin, H., Kahn, J. S., et al. (2012). Characterizing HIV transmission networks across the United States. *Clin. Infect. Dis.* 55, 1135–1143. doi: 10.1093/cid/cis612

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Arimide, D. A., Abebe, A., Kebede, Y., Adugna, F., Tilahun, T., Kassa, D., et al. (2018). HIV-genetic diversity and drug resistance transmission clusters in Gondar, northern Ethiopia, 2003-2013. *PLoS One* 13:e0205446. doi: 10.1371/journal.pone.0205446

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N., et al. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *N. Engl. J. Med.* 365, 493–505. doi: 10.1056/NEJMoa1105243

Dalai, S. C., de Oliveira, T., Harkins, G. W., Kassaye, S. G., Lint, J., Manasa, J., et al. (2009). Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. *AIDS* 23, 2523–2532. doi: 10.1097/QAD.0b013e3283320ef3

Delatorre, E. O., and Bello, G. (2012). Phylodynamics of HIV-1 subtype C epidemic in East Africa. *PLoS One* 7:e41904. doi: 10.1371/journal.pone.0041904

Deribew, A., Biadgilign, S., Deribe, K., Dejene, T., Tessema, G. A., Melaku, Y. A., et al. (2019). The burden of HIV/AIDS in Ethiopia from 1990 to 2016: evidence from the global burden of diseases 2016 study. *Ethiop. J. Health Sci.* 29, 859–868. doi: 10.4314/ejhs.v29i1.7

Desta, S., Feleke, W., Yusuf, M., Mehiret, M., Geyid, A., Ghidinelli, M., et al. (1990). Prevalence of STD and STD related risk factors in sex workers of Addis Ababa. *Ethiop. J. Heal. Dev.* 4, 149–153.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973. doi: 10.1093/molbev/mss075

EPHI (2014). Ethiopian National Key Population HIV Bio-Behavioral Surveillance Round I, 2013 Report: EPHI; 2014.

Esbjörnsson, J., Mild, M., Audelin, A., Fonager, J., Skar, H., Bruun Jørgensen, L., et al. (2016). HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic countries. *Virus Evol.* 2:vew010. doi: 10.1093/ve/vew010

Esbjornsson, J., Mild, M., Mansson, F., Norrgren, H., and Medstrand, P. (2011). HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: origin, demography and migrations. *PLoS One* 6:e17025. doi: 10.1371/journal.pone.0017025

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346, 56–61. doi: 10.1126/science.1256739

Faria, N. R., Vidal, N., Lourenco, J., Raghwani, J., Sigaloff, K. C. E., Tatem, A. J., et al. (2019). Distinct rates and patterns of spread of the major HIV-1 subtypes in central and East Africa. *PLoS Pathog.* 15:e1007976. doi: 10.1371/journal.ppat.1007976

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* 30, 713–724. doi: 10.1093/molbev/mss265

Giovanetti, M., Ciccozzi, M., Parolin, C., and Borsetti, A. (2020). Molecular epidemiology of HIV-1 in African countries: a comprehensive overview. *Pathogens* 9:1072. doi: 10.3390/pathogens9121072

Gray, R. R., Tatem, A. J., Lamers, S., Hou, W., Laeyendecker, O., Serwadda, D., et al. (2009). Spatial phylodynamics of HIV-1 epidemic emergence in East Africa. *AIDS* 23, F9–F17. doi: 10.1097/QAD.0b013e32832faf61

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hadgu, T. G., Egziabher, E., Gizaw, G., Yilma, A., Khodakevich, L., Zewdie, D., et al. (1990). Intersectoral collaboration in AIDS control in Ethiopia. *Ethiop. J. Heal. Dev.* 4, 7–99.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Halperin, D. T., Mugurungi, O., Hallett, T. B., Muchini, B., Campbell, B., Magure, T., et al. (2011). A surprising prevention success: why did the HIV epidemic decline in Zimbabwe? *PLoS Med.* 8:e1000414. doi: 10.1371/journal.pmed.1000414

Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjörnsson, J. (2017). Defining HIV-1 transmission clusters based on sequence data. *AIDS* 31, 1211–1222. doi: 10.1097/QAD.0000000000001470

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Hessol, N. A., Lifson, A. R., O'Malley, P. M., Doll, L. S., Jaffe, H. W., and Rutherford, G. W. (1989). Prevalence, incidence, and progression of human immunodeficiency virus infection in homosexual and bisexual men in hepatitis B vaccine trials, 1978-1988. *Am. J. Epidemiol.* 130, 1167–1175. doi: 10.1093/oxfordjournals.aje.a115445

Hill, V., and Baele, G. (2019). Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. *Mol. Biol. Evol.* 36, 2620–2628. doi: 10.1093/molbev/msz172

Hladik, W., Shabbir, I., Jelaludin, A., Woldu, A., Tsehaynesh, M., and Tadesse, W. (2006). HIV/AIDS in Ethiopia: where is the epidemic heading? *Sex. Transm. Infect.* 82(Suppl. 1), i32–i35. doi: 10.1136/sti.2005.016592

Hué, S., Clewley, J. P., Cane, P. A., and Pillay, D. (2004). HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18, 719–728. doi: 10.1097/00002030-200403260-00002

Hué, S., Pillay, D., Clewley, J. P., and Pybus, O. G. (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4425–4429. doi: 10.1073/pnas.0407534102

Jordan, M. R., Bennett, D. E., Bertagnolio, S., Gilks, C. F., and Sutherland, D. (2008). World Health Organization surveys to monitor HIV drug resistance prevention and associated factors in sentinel antiretroviral treatment sites. *Antivir. Ther.* 13(Suppl. 2), 15–23. doi: 10.1186/s13104-016-2101-8

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kebede, D., Aklilu, M., and Sanders, E. (2000). The HIV epidemic and the state of its surveillance in Ethiopia. *Ethiop. Med. J.* 38, 283–302.

Khodakevich, L., Mehret, M., Negassa, H., and Shanko, B. (1990). Progression of human immunodeficeiencyy virus epidemic in Ethiopia. *Ethiop. J. Heal. Dev.* 4, 183–187.

Kloos, H., and Mariam, D. H. (2000). HIV/AIDS in Ethiopia: an overview. *Northeast. Afr. Stud.* 7, 13–40. doi: 10.1353/nas.2004.0006

Lester, F. T., Ayehunie, S., and Zewdie, D. (1988). Acquired immunodeficiency syndrome: seven cases in an Addis Ababa hospital. *Ethiop. Med. J.* 26, 139–145.

Martin, J. L. (1987). The impact of AIDS on gay male sexual behavior patterns in New York City. *Am. J. Public Health* 77, 578–581. doi: 10.2105/AJPH.77.5.578

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26, 2462–2463. doi: 10.1093/bioinformatics/btq467

Mebret, M., Kbodakcvicb, L., Zewdie, D., Ayebunic, S., Gizaw, G., Shanko, B., et al. (1990). HIV-1 infection and related risk factors among female sex workers in urban areas of Ethiopia. *Ethiop. J. Heal. Dev.* 4, 163–170.

Mehret, M. (1990). HIV-1 infection and related risk factors among female sex workers in urban areas in Ethiopia. *Ethiop. J. Health Dev.* 4(Suppl. 2), 163–170.

Mehret, M., Khodakevich, L., Zewdie, D., Gizaw, G., Ayehune, S., Shanko, B., et al. (1990a). HIV-1 infection among employees of the Ethiopian Freightt transport corporation. *Ethiop. J. Heal. Dev.* 4, 177–182.

Mehret, M., Mertens, T. E., Caraël, M., Negassa, H., Feleke, W., Yitbarek, N., et al. (1996). Baseline for the evaluation of an AIDS programme using prevention indicators: a case study in Ethiopia. *Bull. World Health Organ.* 74, 509–516.

Mehret, M. K. L., Shanko, B., and Belete, F. (1990b). Sexual behaviours and some social features off female sex workers in the city of Addis Ababa. *Ethiop. J. Health Dev.* 4, 133–113.

Mehret, M. K. L., Zewdie, D., Ayehunie, S., Shanko, B., Gizaw, G., et al. (1990c). HIV-1 infection and some related risk factors among female sex workers in Addis Ababa. *Ethiop. J. Health Dev.* 4, 171–176.

Mekonnen, Y., Sanders, E., Aklilu, M., Tsegaye, A., Rinke de Wit, T. F., Schaap, A., et al. (2003). Evidence of changes in sexual behaviours among male factory workers in Ethiopia. *AIDS* 17, 223–231. doi: 10.1097/00002030-200301240-00013

Mir, D., Graf, T., de Matos, E., Almeida, S., Pinto, A. R., Delatorre, E., et al. (2018). Inferring population dynamics of HIV-1 subtype C epidemics in eastern Africa and southern Brazil applying different Bayesian phylodynamics approaches. *Sci. Rep.* 8:8778. doi: 10.1038/s41598-018-26824-4

Mount, D. W. (2007). Using the basic local alignment search tool (BLAST). *Cold Spring Harb. Protoc.* 2007:pdb.top17. doi: 10.1101/pdb.top17

Negassa, H., Kefene, H., Khodakevich, L., Zewdie, D., and Shanko, B. (1990). Profile of AIDS case in Ethiopia. *Ethiop. J. Heal. Dev.* 4, 213–217.

Nelson, K. E., Celentano, D. D., Eiumtrakol, S., Hoover, D. R., Beyrer, C., Suprasert, S., et al. (1996). Changes in sexual behavior and a decline in HIV infection among young men in Thailand. *N. Engl. J. Med.* 335, 297–303. doi: 10.1056/NEJM199608013350501

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Okubagzhi, G., and Singh, S. (2002). Establishing an HIV/AIDS programme in developing countries: the Ethiopian experience. *AIDS* 16, 1575–1586. doi: 10.1097/00002030-200208160-00002

Parker, J., Rambaut, A., and Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* 8, 239–246. doi: 10.1016/j.meegid.2007.08.001

Pineda-Peña, A. C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., et al. (2013). Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* 19, 337–348. doi: 10.1016/j.meegid.2013.04.032

Pollakis, G., Abebe, A., Kliphuis, A., De Wit, T. F., Fisseha, B., Tegbaru, B., et al. (2003). Recombination of HIV type 1C (C′/C″) in Ethiopia: possible link of EthHIV-1C′ to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. *AIDS Res. Hum. Retrovir.* 19, 999–1008. doi: 10.1089/088922203322588350

Pybus, O. G., Charleston, M. A., Gupta, S., Rambaut, A., Holmes, E. C., and Harvey, P. H. (2001). The epidemic behavior of the hepatitis C virus. *Science* 292, 2323–2325. doi: 10.1126/science.1058321

Rambaut, A. (2016). FigTree v1.4.3: Tree Figure Drawing Tool. Available at: http://tree.bio.ed.ac.uk/software/figtree/ (Accessed October 4, 2019).

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian Phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly path-O-gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

Sallam, M., Esbjornsson, J., Baldvinsdottir, G., Indriethason, H., Bjornsdottir, T. B., Widell, A., et al. (2017). Molecular epidemiology of HIV-1 in Iceland: early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infect. Genet. Evol.* 49, 157–163. doi: 10.1016/j.meegid.2017.01.004

Schultz, A. K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., et al. (2009). jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res.* 37, W647–W651. doi: 10.1093/nar/gkp371

Sharp, P. M., and Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harb. Perspect. Med.* 1:a006841. doi: 10.1101/cshperspect.a006841

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Boni, J., Burgisser, P., et al. (2012). Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* 29, 347–357. doi: 10.1093/molbev/msr217

Stadler, T., Kuhnert, D., Bonhoeffer, S., and Drummond, A. J. (2013). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci.* 110, 228–233. doi: 10.1073/pnas.1207965110

Stoneburner, R. L., and Low-Beer, D. (2004). Population-level HIV declines and behavioral risk avoidance in Uganda. *Science* 304, 714–718. doi: 10.1126/science.1093166

Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C., and Bercoff, D. P. (2014). COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res.* 42:e144. doi: 10.1093/nar/gku739

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

Taaffe, J., Fraser-Hurt, N., Gorgens, M., and Harimurti, P. (2014). A Comprehensive Review of Empirical and Modeled HIV Incidence Trends (1990–2012). Policy Research Working Paper; No. 7042. World Bank. Available at: http://hdl.handle.net/10986/20378 (Accessed March 10, 2021).

Thomson, M. M., and Fernandez-Garcia, A. (2011). Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning. *Virology* 415, 30–38. doi: 10.1016/j.virol.2011.03.021

Tsega, E., Mengesha, B., Nordenfelt, E., Hansson, B. G., and Lindberg, J. (1988). Serological survey of human immunodeficiency virus infection in Ethiopia. *Ethiop. Med. J.* 26, 179–184.

Tsegaye, A., Rinke De Wit, T. F., Mekonnen, Y., Beyene, A., Aklilu, M., Messele, T., et al. (2002). Decline in prevalence of HIV-1 infection and syphilis among young women attending antenatal care clinics in Addis Ababa, Ethiopia: results from sentinel surveillance, 1995-2001. *J. Acquir. Immune Defic. Syndr.* 30, 359–362. doi: 10.1097/00126334-200207010-00013

Tully, D. C., and Wood, C. (2010). Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. *AIDS* 24, 1577–1582. doi: 10.1097/QAD.0b013e32833999e1

UNAIDS (2020). Global HIV & AIDS statistics—Fact Sheet. Available at: https://www.unaids.org/en/resources/fact-sheet (Accessed March 10, 2021).

Vasylyeva, T. I., du Plessis, L., Pineda-Pena, A. C., Kuhnert, D., Lemey, P., Vandamme, A. M., et al. (2019). Tracing the impact of public health interventions on HIV-1 transmission in Portugal using molecular epidemiology. *J. Infect. Dis.* 220, 233–243. doi: 10.1093/infdis/jiz085

Wensing, A. M., Calvez, V., Günthard, H. F., Johnson, V. A., Paredes, R., Pillay, D., et al. (2016). 2017 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 24, 132–133.

Wilkinson, E., Engelbrecht, S., and de Oliveira, T. (2015). History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. *Sci. Rep.* 5:16897. doi: 10.1038/srep16897

Wilkinson, E., Rasmussen, D., Ratmann, O., Stadler, T., Engelbrecht, S., and de Oliveira, T. (2016). Origin, imports and exports of HIV-1 subtype C in South Africa: a historical perspective. *Infect. Genet. Evol.* 46, 200–208. doi: 10.1016/j.meegid.2016.07.008

Wolday, D., Meles, H., Hailu, E., Messele, T., Mengistu, Y., Fekadu, M., et al. (2007). Temporal trends in the incidence of HIV infection in antenatal clinic attendees in Addis Ababa, Ethiopia, 1995-2003. *J. Intern. Med.* 261, 132–137. doi: 10.1111/j.1365-2796.2006.01740.x

Woods, C. K., Brumme, C. J., Liu, T. F., Chui, C. K., Chu, A. L., Wynhoven, B., et al. (2012). Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. *J. Clin. Microbiol.* 50, 1936–1942. doi: 10.1128/JCM.06689-11

Yusim, K., Peeters, M., Pybus, O. G., Bhattacharya, T., Delaporte, E., Mulanga, C., et al. (2001). Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 356, 855–866. doi: 10.1098/rstb.2001.0859

Zewdie, D., Gizaw, G., Khodakevich, L., Degifie, G., and Wemeue, M. (1990). Development and management of the AIDS control programme in Ethiopia. *Ethiop. J. Heal. Dev.* 4, 87–96.

**Author Disclaimer:** The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the funding agencies.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 07 March 2022
doi: 10.3389/fmicb.2022.835443

# Dissemination Dynamics of HIV-1 Subtype B Pandemic and Non-pandemic Lineages Circulating in Amazonas, Brazil

*Ighor Arantes[1], Tiago Gräf[2], Paula Andrade[1], Yury Oliveira Chaves[3], Monick Lindenmeyer Guimarães[1] and Gonzalo Bello[1]**

*[1] Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, Brazil, [2] Instituto Gonçalo Moniz, Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Brazil, [3] Laboratório de Diagnóstico e Controle de Doenças Infecciosas na Amazônia, Instituto Leônidas e Maria Deane, Fundação Oswaldo Cruz (FIOCRUZ), Manaus, Brazil*

OPEN ACCESS

**Edited by:**
*Michael M. Thomson,*
*Instituto de Salud Carlos III (ISCIII),*
*Spain*

**Reviewed by:**
*Juan Angel Patino Galindo,*
*Columbia University, United States*
*Adriano de Bernardi Schneider,*
*University of California, Santa Cruz,*
*United States*

***Correspondence:**
*Gonzalo Bello*
*gbello@ioc.fiocruz.br*

**Specialty section:**
*This article was submitted to*
*Virology,*
*a section of the journal*
*Frontiers in Microbiology*

**Received:** *14 December 2021*
**Accepted:** *14 February 2022*
**Published:** *07 March 2022*

**Citation:**
*Arantes I, Gräf T, Andrade P, Oliveira Chaves Y, Guimarães ML and Bello G (2022) Dissemination Dynamics of HIV-1 Subtype B Pandemic and Non-pandemic Lineages Circulating in Amazonas, Brazil. Front. Microbiol. 13:835443. doi: 10.3389/fmicb.2022.835443*

The HIV-1 epidemic in the Amazonas state, as in most of Brazil, is dominated by subtype B. The state, nonetheless, is singular for its significant co-circulation of the variants $B_{CAR}$, which can mostly be found in the Caribbean region, and $B_{PAN}$, a clade that emerged in the United States and aggregates almost the totality of subtype B infections world-wide. The Amazonian HIV-1 epidemic provides a unique scenario to compare the epidemic potential of $B_{PAN}$ and $B_{CAR}$ clades spreading in the same population. To reconstruct the spatiotemporal dynamic and demographic history of both subtype B lineages circulating in Amazonas, we analyzed 1,272 HIV-1 *pol* sequences sampled in that state between 2009 and 2018. Our phylogeographic analyses revealed that while most $B_{CAR}$ infections resulted from a single successful founder event that took place in the Amazonas state around the late 1970s, most $B_{PAN}$ infections resulted from the expansion of multiple clusters seeded in the state since the late 1980s. Our data support the existence of at least four large clusters of the pandemic form in Amazonas, two of them nested in Brazil's largest known subtype B cluster ($B_{BR-I}$), and two others resulting from new introductions detected here. The reconstruction of the demographic history of the most prevalent $B_{PAN}$ ($n$ = 4) and $B_{CAR}$ ($n$ = 1) clades identified in Amazonas revealed that all clades displayed a continuous expansion [effective reproductive number ($R_e$) > 1] until most recent times. During the period of co-circulation from the late 1990s onward, the $R_e$ of Amazonian $B_{PAN}$ and $B_{CAR}$ clusters behaved quite alike, fluctuating between 2.0 and 3.0. These findings support that the $B_{CAR}$ and $B_{PAN}$ variants circulating in the Brazilian state of Amazonas displayed different evolutionary histories, but similar epidemic trajectories and transmissibility over the last two decades, which is consistent with the notion that both subtype B variants display comparable epidemic potential. Our findings also revealed that despite significant advances in the treatment of HIV infections in the Amazonas state, $B_{CAR}$ and $B_{PAN}$ variants continue to expand and show no signs of the epidemic stabilization observed in other parts of the country.

**Keywords: HIV-1, subtype B, Brazil, Amazonas, phylodynamics**

## INTRODUCTION

The HIV-1 subtype B pandemic started when the ancestral virus arrived and first established itself in the Caribbean region during the mid-1960s (Gilbert et al., 2007). The subsequent subtype B spread generated a set of local clades, designated as $B_{CAR}$, that remained mostly confined to the Caribbean region (Gilbert et al., 2007; Cabello et al., 2014). One of those viruses, however, migrated from the Caribbean to the United States around the late 1960s and established a pandemic clade, called $B_{PAN}$, that was then disseminated worldwide (Worobey et al., 2016). The remarkable dissemination of the $B_{PAN}$ clade was probably shaped by ecological, rather than virological factors (Gilbert et al., 2007; Arantes et al., 2020); but there are no studies comparing the epidemic potential of $B_{CAR}$ and $B_{PAN}$ clades co-circulating outside the Caribbean region.

The HIV-1 subtype B epidemic in Brazil is mainly driven by the $B_{PAN}$ clade, with a few notable exceptions, like the Northern state of Amazonas, which is characterized by the co-circulation of the $B_{PAN}$ (75%) and $B_{CAR}$ (25%) variants at a high prevalence (Divino et al., 2016; Crispim et al., 2019; Gräf et al., 2021). Among regional HIV-1 epidemics in Brazil, the one in the state of Amazonas stands out as the second largest AIDS detection rate (34.8 cases per 100,000 inhabitants) in the country, well above the mean national rate (17.8 cases per 100,000 inhabitants) (Ministério da Saúde, 2020). Furthermore, despite important advances in HIV diagnosis and treatment, the Amazonas HIV/AIDS epidemic is not stabilized and has displayed a rising AIDS incidence over the last 10 years (Ministério da Saúde, 2020).

The Amazonian HIV-1 epidemic thus provides a great opportunity to compare the epidemic dynamics of the $B_{PAN}$ and $B_{CAR}$ clades spreading in the same population outside the Caribbean region. Previous studies identified four major $B_{PAN}$ ($B_{PAN-BR-I}$ to $B_{PAN-BR-IV}$) and four major $B_{CAR}$ ($B_{CAR-BR-I}$ to $B_{CAR-BR-IV}$) clusters circulating in Brazil (Mir et al., 2015; Divino et al., 2016). The lineage $B_{CAR-BR-I}$ aggregates the majority (51%) of non-pandemic subtype B sequences from Brazil (Divino et al., 2016) and its most recent common ancestor (MRCA) could be traced back to Amazonas in the late 1970s, from a viral migration probably originated in the French Guiana (Divino et al., 2016; Arantes et al., 2019). The cluster composition and the evolutionary history of the $B_{PAN}$ clade in the Amazonas state are currently unknown.

This work aims to characterize the spatiotemporal dynamics of the major $B_{PAN}$ clusters circulating in the state of Amazonas and to compare the evolutionary and demographic history of major pandemic and non-pandemic subtype B lineages that drive the expanding HIV-1 subtype B epidemic in this Northern Brazilian state.

## METHODOLOGY

### HIV-1 Subtype B *pol* Sequence Dataset From Amazonas State

In this study, we used a total of 1,272 HIV-1 subtype B *pol* sequences (nucleotides 2,253–3,275 of reference strain HXB2) from Amazonas state sampled between 2009 and 2018 that were either available at the Los Alamos HIV Database[1] by March 2021 or were recently published (Chaves et al., 2021; Gräf et al., 2021) and made available in GenBank[2]. Only one sequence per subject was selected. Sequences were aligned using the ClustalW program (Larkin et al., 2007) and all sites associated with major antiretroviral drug resistance in protease and reverse transcriptase were excluded.

### Phylogenetic Classification of Amazonian HIV-1 Subtype B Sequences

Amazonian HIV-1 subtype B sequences were initially classified as $B_{CAR}$ or $B_{PAN}$ and next within major $B_{CAR}$ or $B_{PAN}$ Brazilian clusters by using an evolutionary placement algorithm (EPA) available in RAxML v.8.0.0 (Stamatakis, 2014) for the rapid assignment of query sequences to edges of a reference phylogenetic tree under a maximum-likelihood (ML) model. This analysis allowed us to classify sequences within 10 different clades: $B_{CAR-BR-I}$ to $B_{CAR-BR-IV}$, $B_{PAN-BR-I}$ to $B_{PAN-BR-IV}$, other $B_{CAR}$ lineages, and other $B_{PAN}$ lineages.

### Selection of Brazilian and Global Reference HIV-1 Subtype B *Pol* Datasets

After the initial cluster assignment, the HIV-1 subtype B *pol* Amazonian sequences were aligned with different sub-sets of non-Amazonian subtype B *pol* reference sequences (covering nucleotides 2,253–3,260 relative to HXB2 genome). Amazonian subtype B sequences classified within major $B_{CAR}$ ($B_{CAR-BR-I}$ to $B_{CAR-BR-IV}$) or $B_{PAN}$ ($B_{PAN-BR-I}$ to $B_{PAN-BR-IV}$) Brazilian clusters were aligned with Brazilian sequences from other states that were previously classified within those major lineages (Mir et al., 2015; Divino et al., 2016). Amazonian subtype B sequences classified as "others $B_{CAR}$ lineages" were aligned with sequences representative of the $B_{CAR}$ diversity in the Caribbean region ($n$ = 228) that were also described previously (Cabello et al., 2014; Mendoza et al., 2014). Finally, Amazonian subtype B sequences classified as "others $B_{PAN}$ lineages" were aligned with: (i) one subset of closely related $B_{PAN}$ sequences from Brazil ($n$ = 687) selected from a large dataset of Brazilian sequences ($n$ = 88,441) described previously (Gräf et al., 2021) and (ii) one subset of closely related $B_{PAN}$ sequences from different countries ($n$ = 1,700) selected from a large dataset of worldwide sampled sequences ($n$ = 71,160) recovered from Los Alamos HIV Database. We used the basic local alignment search tool (BLAST)[3] to select the 10 subtype B reference sequences (Brazilian and worldwide) with the highest similarity score (>95%) to each Amazonian subtype B sequence.

### Detection of Major Amazonian HIV-1 Subtype B Clades

Amazonian and non-Amazonian subtype B sequences were subject to ML phylogeographic analyses to identify the $B_{CAR}$ and $B_{PAN}$ sub-clusters that probably originated in Amazonas.

[1] http://www.hiv.lanl.gov

[2] https://www.ncbi.nlm.nih.gov/genbank/

[3] www.ncbi.nlm.nih.gov/BLAST

ML phylogenetic trees were inferred with the PhyML program (Guindon et al., 2010) using an online web server (Guindon et al., 2005) under the GTR + I + Γ4 nucleotide substitution model, as selected by the jModelTest program (Posada, 2008), and the SPR branch-swapping algorithm of heuristic tree search. The reliability of the obtained tree topology was estimated with the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006) based on the Shimodaira-Hasegawa-like procedure. Trees were rooted using subtype D sequences and visualized using the FigTree v1.4.0 software (Rambaut, 2009). The ML trees were employed for the ancestral character state reconstruction (ACR) of epidemic locations with PastML (Ishikawa et al., 2019), using the maximum likelihood Joint (Pupko et al., 2000) and marginal posterior probabilities approximation (MPPA) methods with an F81-like model. Beyond Amazonas, remaining Brazilian states were aggregated in five discrete locations according to their geographic regions and sequences from other countries were aggregated in a single "non-Brazilian" location. Amazonian $B_{CAR}$ and $B_{PAN}$ clades were defined as those monophyletic clusters with high support (aLRT $\geq$ 0.85) that were mostly composed of sequences from Amazonas (>85%) and displayed Amazonas as the most probable ($P \geq 0.85$) state location of its MRCA. Amazonian clades were further subdivided according to size into large ($n > 30$), medium ($n$ = 10–30), and small ($n$ = 2–9) clades.

## Phylodynamic Analysis

The study of epidemiological and evolutionary parameters of major $B_{CAR}$ and $B_{PANDEMIC}$ clusters from Amazonas was done by Bayesian inference using coalescent and birth-death tree priors implemented, respectively, in BEAST v1.10 (Drummond et al., 2002; Suchard et al., 2018) with BEAGLE (Suchard and Rambaut, 2009) to improve run-time, and in BEAST 2.6 (Bouckaert et al., 2019) software packages. Dated phylogenies were inferred with the flexible Bayesian skyline coalescent model (Drummond, 2005). Changes across time in their effective sample size ($N_e$) were estimated using the coalescent Bayesian Skygrid (BSKG) model (Gill et al., 2012), and in their effective reproductive number ($R_e$), using the Birth-death Skyline (BDSKY) model (Stadler et al., 2012). For BDSKY, the sampling rate (δ) was set to zero for the period before the oldest sample and estimated afterward. The $R_e$ was modeled in a piecewise manner in equidistant intervals from the most recent sample up to the root of the tree with a lognormal prior (mean = 0; standard deviation = 1), and the becoming non-infectious rate with a lognormal prior (mean = 0.25; standard deviation = 0.5). All Bayesian MCMC analyses were performed using the GTR + I + Γ4 nucleotide substitution model, and a relaxed uncorrelated lognormal molecular clock model (Drummond et al., 2006) with a uniform prior distribution on the substitution rate that encompasses mean values previously estimated for the subtype B *pol* gene (2.0–3.0 × $10^{-3}$ subst./site/year) (Hue et al., 2005; Zehender et al., 2010; Chen et al., 2011; Mendoza et al., 2014; Bello et al., 2018). MCMC chains were run for 50–100 × $10^6$ generations and convergence and uncertainty of parameter estimates were assessed by calculating the effective sample size (ESS) and 95% highest probability density (HPD) values, respectively, after excluding the initial 10% of each run with Tracer v1.7.1 (Rambaut et al., 2018). The convergence of parameters was considered when ESS $\geq$ 200.

## Statistical Analysis

Demographic information of age group and gender of individuals with samples included in the present study were compared using Pearson's chi-squared test as implemented in R version 3.6.3 (R Core Team, 2018), with 10,000 replicates. The false discovery rate (FDR) method was used to correct for multiple hypothesis testing and to reduce false positives. Statistical significance was defined as $p$-values <0.05.

# RESULTS

The 1,272 HIV-1 subtype B *pol* sequences from Amazonas were assigned to either $B_{CAR}$ (23%) or $B_{PAN}$ (77%) lineages (**Table 1**). The sub-lineage assignment reveals that most $B_{CAR}$ Amazonian sequences belong to the major Brazilian clade $B_{CAR-BR-I}$ (89%), while the remaining sequences were classified within clades $B_{CAR-BR-II}$ (1%), $B_{CAR-BR-III}$ (1%), or branched outside known Brazilian clades (8%) (**Table 1**). Although a high proportion of $B_{PAN}$ Amazonian sequences also branched within major Brazilian $B_{PAN}$ clades (45%), particularly the $B_{PAN-BR-I}$ (37%), most pandemic sequences (55%) branched outside known countrywide Brazilian clades (**Table 1**). To identify the major subtype clades circulating in Amazonas, we next conducted independent ML phylogeographic analyses for: (i) sequences that branched within major Brazilian $B_{CAR}$ or $B_{PAN}$ clades, by combining Brazilian sequences sampled in Amazonas and other states, and (ii) sequences that branched outside major Brazilian clades, by combining Amazonian sequences with either $B_{CAR}$ sequences of Caribbean origin or $B_{PAN}$ sequences sampled in

**TABLE 1** | Lineage classification of HIV-1 subtype B *pol* sequences from Amazonas state.

| Lineage | Sub-lineage | *N* (%) | Sampling range |
|---|---|---|---|
| $B_{PAN}$ | $B_{PAN-BR-I}$ | 360 (37%) | 2009–2018 |
| | $B_{PAN-BR-II}$ | 38 (4%) | |
| | $B_{PAN-BR-III}$ | 23 (2%) | |
| | $B_{PAN-BR-IV}$ | 23 (2%) | |
| | Others | 530 (55%) | |
| | **Total** | **974 (100%)** | |
| $B_{CAR}$ | $B_{CAR-BR-I}$ | 267 (90%) | 2009–2018 |
| | $B_{CAR-BR-II}$ | 2 (1%) | 2016–2017 |
| | $B_{CAR-BR-IV}$ | 4 (1%) | 2015–2017 |
| | Others | 25 (8%) | 2009–2018 |
| | **Total** | **298 (100%)** | |

*The table details the distribution of 1,272 HIV-1 subtype B pol sequences (nucleotides 2,253–3,275 of reference strain HXB2) from Brazil's Amazonas state across the known clusterization profile of the pandemic ($B_{PAN}$, n = 974, 77%) and non-pandemic ($B_{CAR}$, n = 298, 23%) forms. For each cluster, its absolute and relative frequency of samples from Amazonas state, as well as their distribution in time, are indicated. The class "Other" aggregates sequences not clustered among established major Brazilian clusters.*

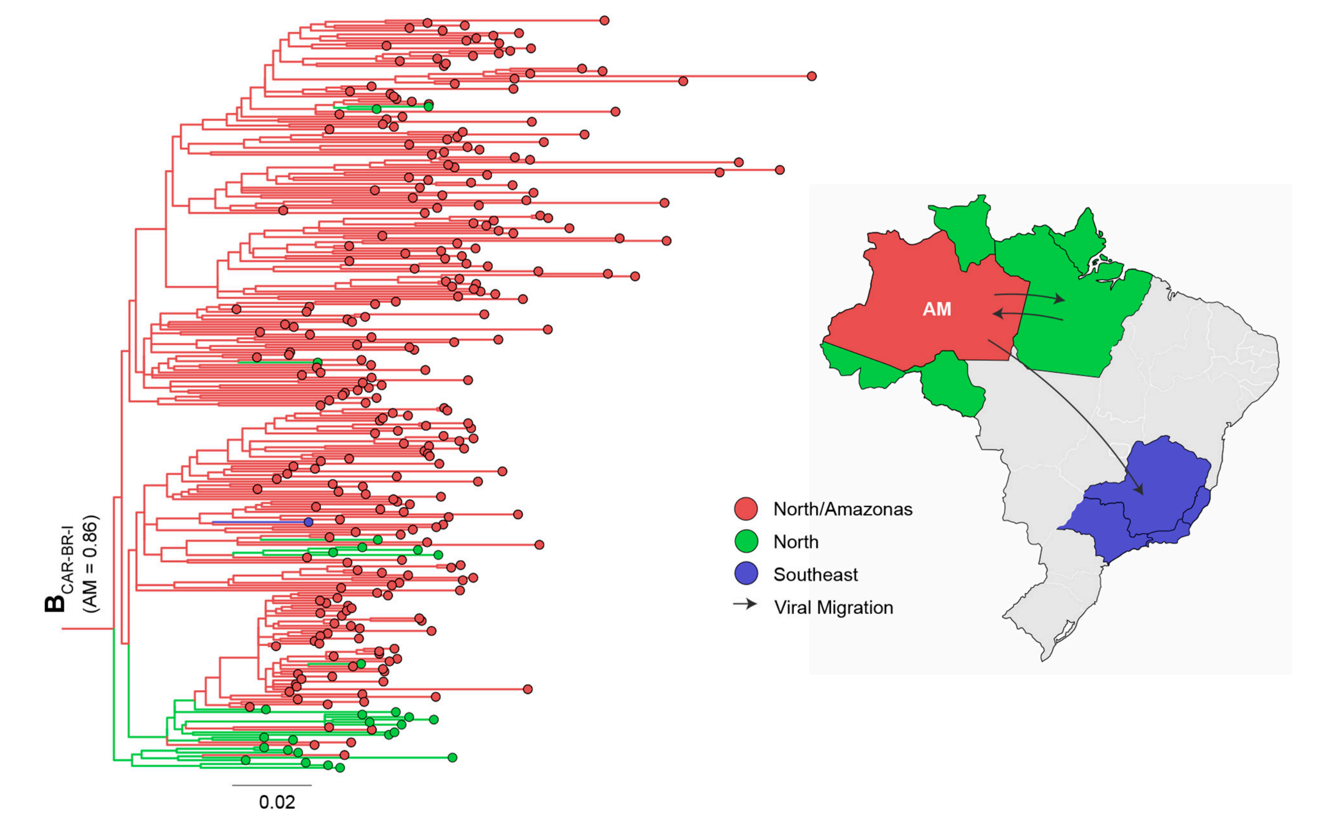


**FIGURE 1 |** ML phylogeographic analysis of major cluster $B_{CAR-BR-I}$ disseminated in Brazil. A total of 296 $B_{CAR-BR-I}$ Brazilian sequences sampled in Amazonas (*n* = 267) and other states (*n* = 29) were analyzed. The number in parenthesis indicates the inferred marginal probability that the clade ancestor was located in the Amazonas state. The location of taxonomic units at internal nodes across the ML tree was reconstructed and represented according to the color scheme shown in the map. Outside Amazonas state, other Brazilian units were aggregated according to the country region. AM, Amazonas; N, North; SE, Southeast. The tree was rooted using HIV-1 subtype D reference sequences (not shown). The branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site. Major migrations of viral clade $B_{CAR-BR-I}$ are represented in the map.

Brazil and worldwide. The ML phylogeographic analyses confirm multiple introductions of $B_{CAR}$ (*n* = 18) and $B_{PAN}$ (*n* = 291) variants into the Amazonas state (**Supplementary Table 1**). The major $B_{CAR}$ founder event resulted in the $B_{CAR-BR-I}$ clade (**Figure 1**) while the remaining $B_{CAR}$ sequences were distributed among a few local clusters of small size (2–9 sequences, 7%) or singletons (4%) (**Supplementary Figures 1A–C**). A few $B_{PAN}$ introductions (*n* = 4) originated highly supported (aLRT > 0.85) Amazonian $B_{PAN}$ clades of large size that were mostly composed by sequences from Amazonas (>85%) and most probably arose in Amazonas ($P > 0.90$). Two major clusters, $B_{PAN-BR-I-AM-I}$ (*n* = 39) and $B_{PAN-BR-I-AM-II}$ (*n* = 35), were nested within the large Brazilian clade $B_{PAN-BR-I}$ (**Figure 2**). The other two clusters, $B_{PAN-AM-I}$ (*n* = 86) and $B_{PAN-AM-II}$ (*n* = 60), branched outside the major $B_{PAN}$ Brazilian clades (**Figure 3**). Because the phylogenetic placement of some Amazonian basal sequences in clusters $B_{PAN-AM-I}$ and $B_{PAN-AM-II}$ changed according to the reference (Brazilian or worldwide) dataset, we define the final size of those clusters by the monophyletic groups supported by both analyses. The four major Amazonian $B_{PAN}$ clades, together, comprise 23% of all $B_{PAN}$ Amazonian sequences analyzed and the remaining sequences branched within Amazonian clades of medium (24%) or small (35%) size or appeared as singletons that branched with non-Amazonian sequences (18%) (**Figures 2**, **3** and **Supplementary Figures 1D–F**).

To study in more detail the evolutionary and demographic history of lineages $B_{CAR}$ and $B_{PAN}$ spreading in Amazonas, we selected the five major clades that display both local epidemic importance – as, combined, they comprise 38% of HIV-1 subtype B infections in the state – and adequate sample sizes to give reliable demographic estimates. Time-scaled trees were reconstructed using a Bayesian coalescent model with an informative clock rate prior due to the weak temporal structure of Amazonian subtype B *pol* datasets (**Supplementary Figure 2**). Posterior estimates, that were, as expected, largely influenced by the selected clock rate prior, traced the median $T_{MRCA}$ of major Amazonian clades to the late 1970s for $B_{CAR-BR-I}$, the late 1980s for $B_{PAN-AM-I}$, the mid-1990s for $B_{PAN-AM-II}$, and the late 1990s for $B_{PAN-BR-I-AM-I}$ and $B_{PAN-BR-I-AM-II}$ (**Table 2**). These findings support that the major $B_{CAR}$ clade was successfully spreading in Amazonas for about 10 years before the emergence of the major $B_{PAN}$ clades, which may explain the singular high prevalence of non-pandemic subtype B variants in Amazonas with respect to most other Brazilian states. The

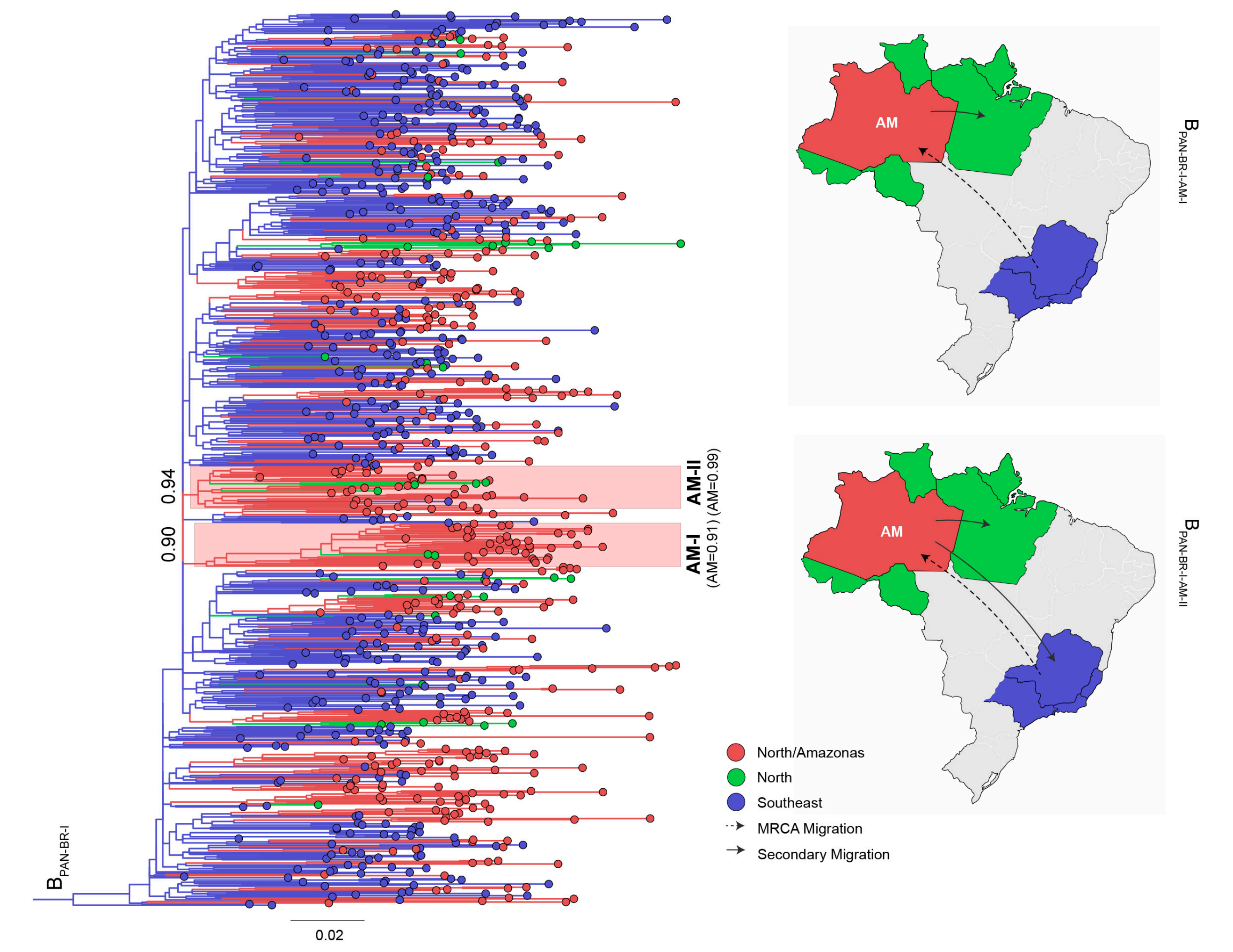


**FIGURE 2 |** ML phylogeographic analysis of major $B_{PAN-BR-I}$ cluster disseminated in Brazil. A total of 844 $B_{PAN-BR-I}$ Brazilian sequences sampled in Amazonas (*n* = 360) and other states from the Northern and Southeastern regions (*n* = 484) were analyzed together. We selected sequences from Northern and Southeastern states because were the regions more strongly connected with the Amazonas state. Major Amazonian clades $B_{PAN-BR-I-AM-I}$ and $B_{PAN-BR-I-AM-II}$ are indicated by colored shaded boxes along with cluster aLRT support (number at basal branch) and the inferred marginal probability that cluster ancestor was located in Amazonas state (in parentheses). The location of taxonomic units at internal nodes across the ML tree was reconstructed and represented according to the color scheme shown in the map. Sequences were grouped in three discrete locations: Amazonas state (AM), Northern region (N), and Southeastern region (SE). The tree was rooted using HIV-1 subtype D reference sequences (not shown). The branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site. Major migrations of viral clades $B_{PAN-BR-I-AM-I}$ and $B_{PAN-BR-I-AM-II}$ are represented in the map.

BSKG model supports that the $N_e$ of lineages $B_{CAR-BR-I}$ and $B_{PAN-BR-I-AM-I}$ steadily increased until recent years, while the $N_e$ of lineages $B_{PAN-AM-I}$, $B_{PAN-AM-II}$, and $B_{PAN-BR-I-AM-II}$ increased until the late-2000s, but then stabilized in more recent years (**Figures 4A–E**). The temporal trajectories of the $R_e$ estimated using the Bayesian BDSKY model, however, support that all major Amazonian HIV-1 subtype B clades continuously expanded (median $R_e > 1$) over all the studied period, with some temporal fluctuations in the rate of expansion (**Figures 4A–E**). The clade $B_{CAR-BR-I}$ reached the highest median $R_e$ (2.5–2.6) between the late 1970s and the early 1990s, while the $B_{PAN}$ clades reached the highest median $R_e$ (2.9–3.4) between the mid-1990s and the mid-2000s (**Table 2**). Despite those differences in the early phase of spread, all major Amazonian subtype B clades converge to the roughly similar median growth rate ($R_e$ = 1.6–2.3) at the most recent time period analyzed (2010–2018), with no evidence of recent epidemic stabilization ($R_e > 1$) (**Figure 4F**).

## DISCUSSION

Previous studies demonstrate that the expanding HIV-1 subtype B epidemic in the Northern Brazilian state of Amazonas was driven by both pandemic ($B_{PAN}$) and non-pandemic ($B_{CAR}$) viral variants (Cabello et al., 2015; Divino et al., 2016; Arantes et al., 2019; Crispim et al., 2019; Chaves et al., 2021; Gräf et al., 2021), thus creating a great opportunity to compare the epidemic dynamics of both subtype B forms spreading in

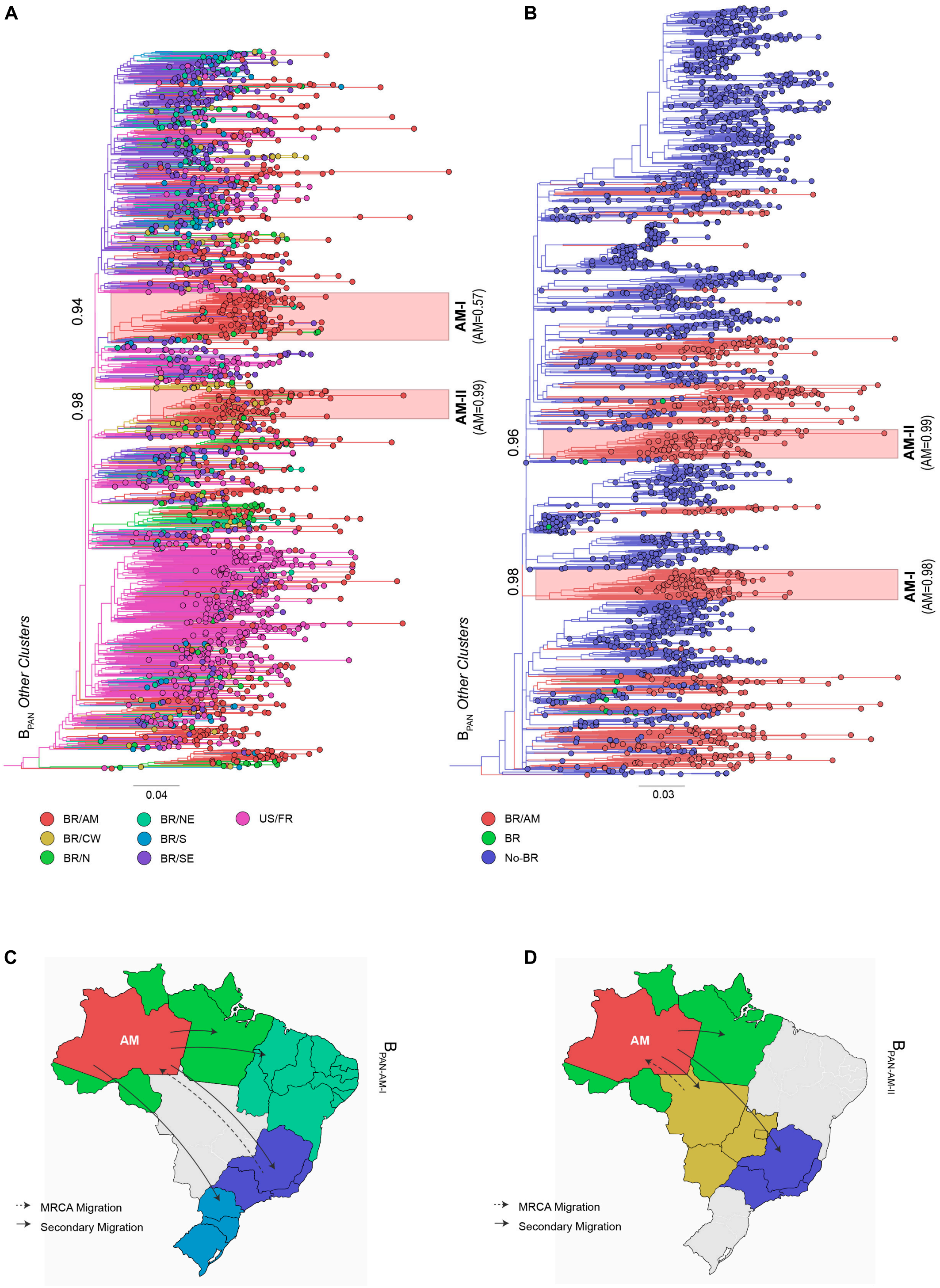


**FIGURE 3 |** ML phylogeographic analysis of HIV-1 $B_{PAN}$ Amazonian sequences that branched outside major Brazilian clades. A total of 530 $B_{PAN}$ sequences sampled in Amazonas were aligned with: **(A)** Closely related $B_{PAN}$ sequences sampled in other Brazilian states (*n* = 687) plus reference $B_{PAN}$ sequences from the United States and France (*n* = 498), and **(B)** closely related $B_{PAN}$ sequences sampled worldwide (*n* = 1,683). Major Amazonian clades $B_{PAN-AM-I}$ and $B_{PAN-AM-II}$ are indicated by colored shaded boxes along with cluster aLRT support (number at basal branch) and the inferred marginal probability that cluster ancestor was located in Amazonas state (in parenthesis). The location of taxonomic units at internal nodes across the ML tree was reconstructed and represented according to the color scheme shown in the map. Sequences were grouped in seven **(A)** and three **(B)** discrete locations: Amazonas state (AM), Central-Western region (CW), Northern region (N); Northeastern region (NE), Southeastern region (SE), Southern region (S), United States/France (US/FR), other Brazilian states (BR), and non-Brazilian (Non-BR). Tree was rooted using HIV-1 subtype D reference sequences (not shown). The branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site. Major migrations of viral clades **(C)** $B_{PAN-AM-I}$ and **(D)** $B_{PAN-AM-II}$ are represented in the maps.

**TABLE 2 |** Bayesian estimates of evolutionary and demographic parameters of major HIV-1 subtype B clades originated in the Amazonas state.

| Clade | *N* | $T_{MRCA}$<sup></sup>$^a$ (95% HPD) | $R_e$$^b$ (95% HPD) | |
|---|---|---|---|---|
| $B_{CAR-BR-I}$ | 267 | 1978 (1970–1987) | 1978–1985 | 2.6 (0.2–4.6) |
| | | | 1986–1993 | 2.5 (1.5–4.2) |
| | | | 1994–2001 | 1.9 (1.2–3.0) |
| | | | 2002–2009 | 1.9 (1.2–3.0) |
| | | | 2010–2018 | 2.3 (1.5–3.5) |
| $B_{PAN-AM-I}$ | 86 | 1988 (1981–1994) | 1988–1997 | 2.0 (0.5–4.1) |
| | | | 1998–2007 | 2.9 (1.6–4.8) |
| | | | 2008–2018 | 1.6 (1.1–2.6) |
| $B_{PAN-AM-II}$ | 60 | 1995 (1990–2000) | 1995–2006 | 2.9 (1.6–5.4) |
| | | | 2007–2018 | 2.2 (1.4–3.6) |
| $B_{PAN-BR-I-AM-I}$ | 39 | 1998 (1992–2002) | 1998–2007 | 3.1 (1.6–5.6) |
| | | | 2008–2018 | 2.1 (1.1–3.7) |
| $B_{PAN-BR-I-AM-II}$ | 35 | 1998 (1992–2003) | 1998–2007 | 3.4 (1.7–5.9) |
| | | | 2008–2018 | 1.9 (1.1–3.3) |

*The table details the evolutionary and epidemiological parameters of the major HIV-1 subtype B clusters of pandemic ($B_{PAN}$) and non-pandemic ($B_{CAR}$) forms circulating in Brazil's Amazonas state.*

[a] *Median value and 95% HPD interval of the time to the most recent common ancestor ($T_{MRCA}$).*

[b] *Median value and 95% HPD interval of the effective reproductive number ($R_e$) inferred in a bird-death statistical model.*

the same population. This study revealed that the $B_{PAN}$ and $B_{CAR}$ epidemics in Amazonas have been shaped by different evolutionary histories, but displayed very similar transmissibility and expansion dynamics at most recent times.

Our analysis confirmed that variants $B_{CAR}$ and $B_{PAN}$ were introduced multiple times in the Amazonas state, although the estimated number of $B_{PAN}$ introductions was 16 times higher than that of $B_{CAR}$. The founder event that originated the clade $B_{CAR-BR-I}$ occurred in the late 1970s (Divino et al., 2016; Arantes et al., 2019) and gave rise to 89% of $B_{CAR}$ and 19% of total subtype B infections in Amazonas. In sharp contrast, most $B_{PAN}$ sequences from Amazonas branched into multiple state-specific clusters of medium/small size (57%) or appeared as unclustered infections (19%). The four largest $B_{PAN}$ Amazonian clades identified probably arose between the late 1980s and late 1990s and, together, comprise 23% of the $B_{PAN}$ and 17% of all subtype B infections in the state. These findings support that the early introduction (late 1970s) of the $B_{CAR-BR-I}$ ancestor in Amazonas from neighboring Caribbean countries probably drove its successful establishment and wide dissemination in the state. Although the $B_{PAN}$ strains arrived in Amazonas later, they reached a high prevalence because they were introduced at much higher numbers and spread through more transmission networks than $B_{CAR}$ strains.

The AIDS detection rate increased ∼10% in the Amazonas state between 2009 and 2019 (Ministério da Saúde, 2020). This finding is consistent with our BDSKY analyses that support a continuous expansion ($R_e > 1$) of major $B_{CAR}$ and $B_{PAN}$ Amazonian clades over all the studied period. The BSKG model indicates a recent stabilization of some Amazonian $B_{PAN}$ clades since the late 2000s, and a previous study conducted by our group also indicated a recent epidemic stabilization of the clade $B_{CAR-BR-I}$ since the late 2000s (Arantes et al., 2019). Although the median estimated $R_e$ of the $B_{CAR}$ and $B_{PAN}$ Amazonian clades was somewhat lower between 2010 and 2018 (1.6–2.3) than during the previous decades (2.5–3.4), we found no solid evidence of epidemic stabilization or reduction in the BDSKY analyses. A previous study pointed out that the BSKG model requires strongly informative data to prevent erroneous estimates of $N_e$ stabilization (Volz and Didelot, 2018). Thus, we hypothesize that the much larger number of recent (2009–2018) $B_{CAR-BR-I}$ sequences used in the present study ($n$ = 267) compared to the previous one ($n$ = 45) allowed us to obtain a more accurate demographic reconstruction of the epidemic pattern in the last two decades.

It is interesting to note that the $B_{CAR}$ (2.5–2.6) and $B_{PAN}$ (2.9–3.4) Amazonian clades reached similar highest median $R_e$ values. Furthermore, the highest median $R_e$ estimated here using a birth-death approach was comparable to the previous ones estimated for the $B_{CAR-BR-I}$ clade (3.8) using a coalescent-based approach (Arantes et al., 2019), but lower than those estimated for major $B_{PAN}$ Brazilian lineages spreading in the Southeastern region (5.0–7.9) (Mir et al., 2015). These findings support that differences in the spreading dynamics of subtype B lineages may reflect discrepancies in the connectivity of underlying transmission networks across different Brazilian states/regions, rather than intrinsic differences in viral transmissibility. A preliminary analysis of the available demographic data (sex and age) of HIV-infected subjects from Amazonas revealed no significant differences between major $B_{CAR}$ and $B_{PAN}$ clades (**Supplementary Table 2**), supporting that both viral lineages are possibly spreading through networks with similar epidemiological properties. This observation is also consistent with a previous study that revealed comparable epidemic growth rates of $B_{CAR}$ and $B_{PAN}$ lineages circulating in the French Guiana (Bello et al., 2018).

Our study has some limitations. First, inferences about potential sources, total number of viral introductions, and local clade size in Amazonas were limited by both the incomplete sampling of local population and the limited number of non-Amazonian reference sequences included in each ML phylogenetic analysis as revealed by the variable phylogenetic and phylogeographic placement of some Amazonian sequences that branched basal to each local clade. The bulk of $B_{CAR}$ and $B_{PAN}$ sequences that compose each major Amazonian clade, however, remained constant across analyses, and our major phylogeographic conclusions were robust to sampling bias. Second, time-scale reconstructions were largely influenced by the selected clock rate prior due to the weak temporal structure of Amazonian HIV-1 subtype B *pol* datasets. Despite this, the $T_{MRCA}$ here obtained were fully consistent with the overall time-scale of dissemination of the HIV-1 $B_{CAR}$ and $B_{PAN}$ lineages in the Americas and Brazil described in previous studies (Gilbert et al., 2007; Cabello et al., 2014; Mir et al., 2015; Worobey et al., 2016; Arantes et al., 2019; Bello et al., 2019). Finally, the lack of epidemiological data regarding the mode of transmission of individuals analyzed reduced the power of our

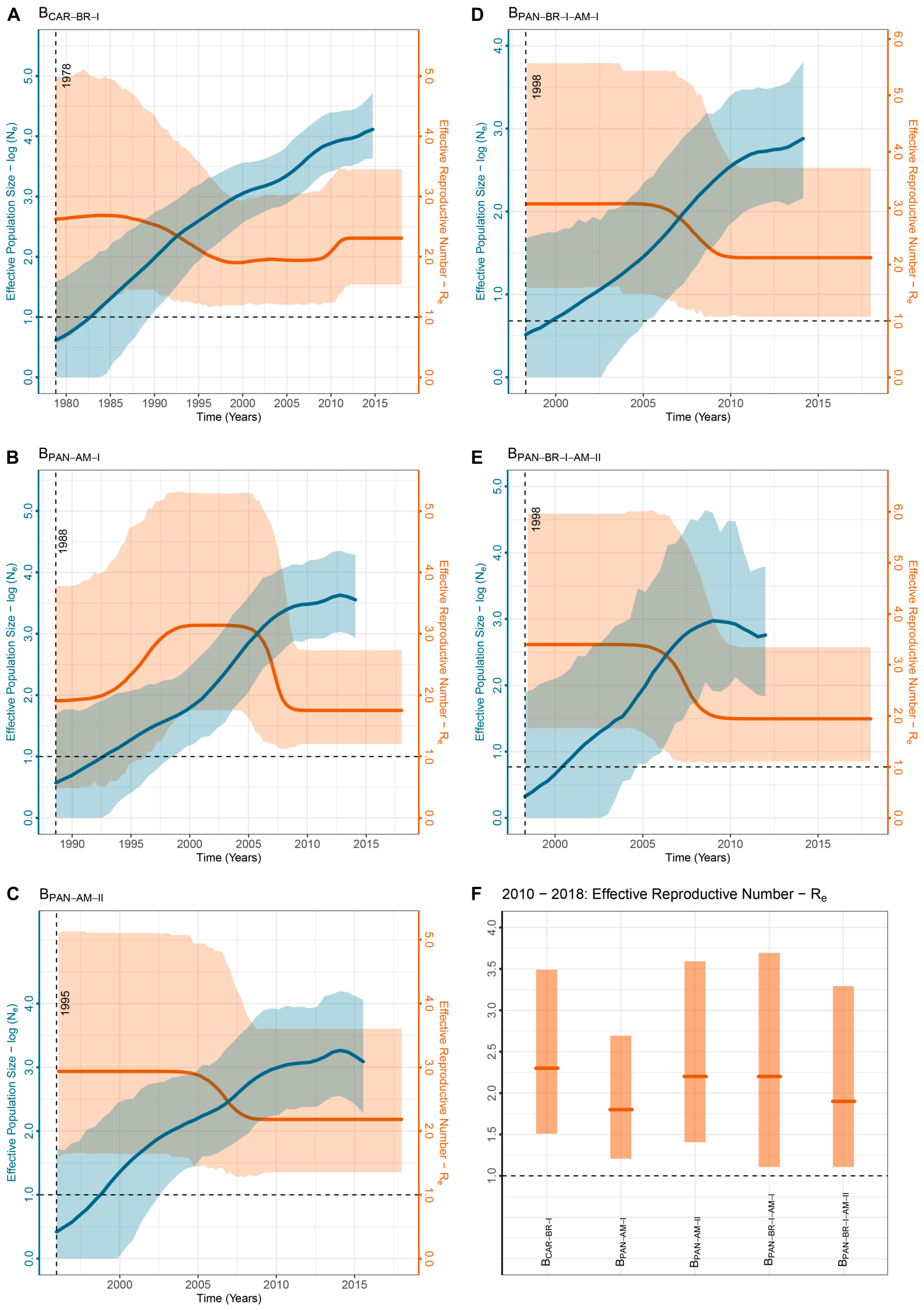


**FIGURE 4 |** Demographic history of subtype B cluster in Brazil's Amazonas state. Each plot **(A–E)** details the demographic history of one subtype B large cluster ($n \geq 30$) in Amazonas from the pandemic ($B_{PAN}$) and non-pandemic ($B_{CAR}$) forms. The graphs exhibit their effective number of HIV-1 infections under the Bayesian Skygrid (BSKG) model in blue ($N_e$, *y*-left-axis), and their effective reproductive number under the Birth-death Skyline (BDSKY) model in orange ($N_e$, *y*-right-axis). For both parameters are indicated their median (solid lines) and 95% HPD intervals (pale areas) estimates. A dashed vertical line indicates the TMRCA of the clades, accompanied by its median value. The last graph **(F)** compares the $R_e$ obtained for the five clusters in the last period of analysis (2010–2018). For each cluster, its median $R_e$ (solid line) and 95% HPD interval (pale area) inferred values are represented.

study to confirm any association between the inferred rate of viral spread and the ecological characteristics of local transmission networks in Amazonas.

In summary, this study highlights that the HIV-1 epidemic in the Amazonas state mostly results from the local expansion of one $B_{CAR}$ strain ($B_{CAR-BR-I}$) introduced around the late 1970s and of multiple $B_{PAN}$ viral strains introduced since the late 1980s. Albeit the earlier introduction of the $B_{CAR-BR-I}$ clade granted a much prolonged period of local spread than that of the $B_{PAN}$ strains, this was compensated by a much higher number of independent introductions and the concurrent establishment of multiple $B_{PAN}$ local transmission networks. Despite significant differences in the pattern of early establishment, major $B_{CAR}$ and $B_{PAN}$ clades circulating in Amazonas have been spreading at a quite similar rate over the last two decades, arguing against the hypothesis of significant differences in their intrinsic transmissibility. Our analyses also demonstrate that major Amazonian $B_{CAR}$ and $B_{PAN}$ clades continued to spread and showed no clear signs of recent epidemic stabilization, supporting the relevance of designing more effective strategies to prevent HIV transmission in the region.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GenBank accession numbers: KEXV01000001 to KEXV01046877; HQ127524 to HQ127607; KU762066 to KU762066; MH673055 to MH673280; and MW545333 to MW545424; Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov).

## ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. Written informed consent for participation was not required for this study in accordance with the National Legislation and the Institutional Requirements.

## AUTHOR CONTRIBUTIONS

GB conceived and designed the study and supervised the experiments. IA conducted the experiments and analyzed the data. YO, TG, and MG provided HIV-1 sequence data and intellectual input. IA and GB wrote the first draft of the manuscript. All authors assisted with the writing and approved the final manuscript.

## FUNDING

GB was supported by fundings from CNPq (Grant No. 304883/2020-4) and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ, Grant No. E-26/202.896/2018). IA was funded by a fellowship from FAPERJ and Fundação Oswaldo Cruz (FIOCRUZ).

## ACKNOWLEDGMENTS

This study was carried out with public data provided by the Department of Diseases of Chronic Condition and Sexually Transmitted Infections of the Secretariat for Health Surveillance of the Ministry of Health. We would like to thank the reviewers for their valuable and insightful comments toward improving our manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.835443/full#supplementary-material

## REFERENCES

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552. doi: 10.1080/10635150600755453

Arantes, I., Esashika Crispim, M. A., Nogueira da Guarda Reis, M., Martins Araújo Stefani, M., and Bello, G. (2019). Reconstructing the dissemination dynamics of the major HIV-1 subtype B non-pandemic lineage circulating in brazil. *Viruses* 11:909. doi: 10.3390/v11100909

Arantes, I., Ribeiro-Alves, M., de Azevedo, S. S. D., Delatorre, E., and Bello, G. (2020). Few amino acid signatures distinguish HIV-1 subtype B pandemic and non-pandemic strains. *PLoS One* 15:e0238995. doi: 10.1371/journal.pone.0238995

Bello, G., Arantes, I., Lacoste, V., Ouka, M., Boncy, J., Césaire, R., et al. (2019). Phylogeographic Analyses Reveal the Early Expansion and Frequent Bidirectional Cross-Border Transmissions of Non-pandemic HIV-1 Subtype B Strains in Hispaniola. *Front. Microbiol.* 10:1340. doi: 10.3389/fmicb.2019.01340

Bello, G., Nacher, M., Divino, F., Darcissac, E., Mir, D., and Lacoste, V. (2018). The HIV-1 subtype B epidemic in french guiana and suriname is driven by ongoing transmissions of pandemic and non-pandemic lineages. *Front. Microbiol.* 9:1738. doi: 10.3389/fmicb.2018.01738

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650. doi: 10.1371/journal.pcbi.1006650

Cabello, M., Junqueira, D. M., and Bello, G. (2015). Dissemination of nonpandemic Caribbean HIV-1 subtype B clades in Latin America. *AIDS* 29, 483–492. doi: 10.1097/qad.0000000000000552

Cabello, M., Mendoza, Y., and Bello, G. (2014). Spatiotemporal dynamics of dissemination of non-pandemic HIV-1 subtype B clades in the caribbean region. *PLoS One* 9:e106045. doi: 10.1371/journal.pone.0106045

Chaves, Y. O., Pereira, F. R., de Souza Pinheiro, R., Batista, D. R. L., da Silva Balieiro, A. A., de Lacerda, M. V. G., et al. (2021). High detection rate of HIV drug resistance mutations among patients who fail combined antiretroviral therapy in manaus, brazil. *BioMed. Res. Int.* 2021:5567332. doi: 10.1155/2021/5567332

Chen, J. H. K., Wong, K. H., Chan, K. C. W., To, S. W. C., Chen, Z., and Yam, W. C. (2011). Phylodynamics of HIV-1 subtype B among the men-having-sex-with-men (MSM) population in Hong Kong. *PLoS One* 6:e25286. doi: 10.1371/journal.pone.0025286

Crispim, M. A. E., Reis, M. N. D. G., Abrahim, C., Kiesslich, D., Fraiji, N., Bello, G., et al. (2019). Homogenous HIV-1 subtype B from the Brazilian Amazon with infrequent diverse BF1 recombinants, subtypes F1 and C among blood donors. *PLoS One* 14:e0221151. doi: 10.1371/journal.pone.0221151

Divino, F., de Lima Guerra Corado, A., Gomes Naveca, F., Stefani, M. M. A., and Bello, G. (2016). High prevalence and onward transmission of non-pandemic HIV-1 subtype B clades in northern and northeastern brazilian regions. *PLoS One* 11:e0162112. doi: 10.1371/journal.pone.0162112

Drummond, A. J. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192. doi: 10.1093/molbev/msi103

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88. doi: 10.1371/journal.pbio.0040088

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161, 1307–1320. doi: 10.1093/genetics/161.3.1307

Gilbert, M. T. P., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., and Worobey, M. (2007). The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl. Acad. Sci.* 104, 18566–18570. doi: 10.1073/pnas.0705329104

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2012). Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* 30, 713–724. doi: 10.1093/molbev/mss265

Gräf, T., Bello, G., Andrade, P., Arantes, I., Pereira, J. M., da Silva, A. B. P., et al. (2021). HIV-1 molecular diversity in Brazil unveiled by 10 years of sampling by the national genotyping network. *Sci. Rep.* 11:15842. doi: 10.1038/s41598-021-94542-5

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005). PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* 33, W557–W559. doi: 10.1093/nar/gki352

Hue, S., Pillay, D., Clewley, J. P., and Pybus, O. G. (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl. Acad. Sci.* 102, 4425–4429. doi: 10.1073/pnas.0407534102

Ishikawa, S. A., Zhukova, A., Iwasaki, W., and Gascuel, O. (2019). A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Mol. Biol. Evol.* 36, 2069–2085. doi: 10.1093/molbev/msz131

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Mendoza, Y., Martínez, A. A., Castillo Mewa, J., González, C., García-Morales, C., Avila-Ríos, S., et al. (2014). Human immunodeficiency virus type 1 (HIV-1) subtype B epidemic in Panama is mainly driven by dissemination of country-specific clades. *PLoS One* 9:e95360. doi: 10.1371/journal.pone.0095360

Ministério da Saúde (2020). *Boletim Epidemiológico HIV/AIDS 2020*. Available online at: http://www.aids.gov.br/system/tdf/pub/2016/67456/boletim_hiv_aids_2020_com_marcas_2.pdf?file=1&type=node&id=67456&force=1 (accessed December 14, 2021).

Mir, D., Cabello, M., Romero, H., and Bello, G. (2015). Phylodynamics of major HIV-1 subtype B pandemic clades circulating in Latin America. *AIDS* 29, 1863–1869. doi: 10.1097/qad.0000000000000770

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* 25, 1253–1256. doi: 10.1093/molbev/msn083

Pupko, T., Pe, I., Shamir, R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.* 17, 890–896. doi: 10.1093/oxfordjournals.molbev.a026369

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org

Rambaut, A. (2009). *FigTree v1.4: Tree Figure Drawing Tool*. Available online at: http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 25 Nov, 2018)

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Stadler, T., Kuhnert, D., Bonhoeffer, S., and Drummond, A. J. (2012). Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci.* 110, 228–233. doi: 10.1073/pnas.1207965110

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Vir. Evol.* 4:vey016. doi: 10.1093/ve/vey016

Suchard, M. A., and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25, 1370–1376. doi: 10.1093/bioinformatics/btp244

Volz, E. M., and Didelot, X. (2018). Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst. Biol.* 67, 719–728. doi: 10.1093/sysbio/syy007

Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., et al. (2016). 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 539, 98–101. doi: 10.1038/nature19827

Zehender, G., Ebranati, E., Lai, A., Santoro, M. M., Alteri, C., Giuliani, M., et al. (2010). Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. *JAIDS* 55, 156–160. doi: 10.1097/qai.0b013e3181eb3002

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 09 March 2022
doi: 10.3389/fmicb.2022.843330

# Phylogeographic Assessment Reveals Geographic Sources of HIV-1 Dissemination Among Men Who Have Sex With Men in Kenya

George M. Nduva[1,2]*, Frederick Otieno[3], Joshua Kimani[4,5], Lyle R. McKinnon[4,5,6], Francois Cholette[5,7], Paul Sandstrom[7], Susan M. Graham[2,8], Matt A. Price[9,10], Adrian D. Smith[11], Robert C. Bailey[3,12], Amin S. Hassan[1,2†], Joakim Esbjörnsson[1,11†] and Eduard J. Sanders[2,11†]

[1] Department of Translational Medicine, Lund University, Lund, Sweden, [2] Kenya Medical Research Institute-Wellcome Trust Research Programme, Kilifi, Kenya, [3] Nyanza Reproductive Health Society, Kisumu, Kenya, [4] Department of Medical Microbiology, University of Nairobi, Nairobi, Kenya, [5] Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada, [6] Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa, [7] National Microbiology Laboratory at the JC Wilt Infectious Diseases Research Centre, Public Health Agency of Canada, Winnipeg, MB, Canada, [8] Department of Epidemiology, University of Washington, Seattle, WA, United States, [9] IAVI, San Francisco, CA, United States, [10] Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States, [11] Nuffield Department of Medicine, The University of Oxford, Oxford, United Kingdom, [12] Division of Epidemiology and Biostatistics, University of Illinois Chicago, Chicago, IL, United States

OPEN ACCESS

**Edited by:**
Kok Keng Tee,
University of Malaya, Malaysia

**Reviewed by:**
Mariza Gonçalves Morgado,
Oswaldo Cruz Foundation (Fiocruz),
Brazil
Monick Lindenmeyer Guimarães,
Oswaldo Cruz Foundation (Fiocruz),
Brazil

***Correspondence:**
George M. Nduva
george.makau_nduva.7410@med.lu.se

†These authors have contributed equally to this work and share senior authorship

**Specialty section:**
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

**Received:** 25 December 2021
**Accepted:** 19 January 2022
**Published:** 09 March 2022

**Citation:**
Nduva GM, Otieno F, Kimani J, McKinnon LR, Cholette F, Sandstrom P, Graham SM, Price MA, Smith AD, Bailey RC, Hassan AS, Esbjörnsson J and Sanders EJ (2022) Phylogeographic Assessment Reveals Geographic Sources of HIV-1 Dissemination Among Men Who Have Sex With Men in Kenya. Front. Microbiol. 13:843330. doi: 10.3389/fmicb.2022.843330

HIV-1 transmission dynamics involving men who have sex with men (MSM) in Africa are not well understood. We investigated the rates of HIV-1 transmission between MSM across three regions in Kenya: Coast, Nairobi, and Nyanza. We analyzed 372 HIV-1 partial *pol* sequences sampled during 2006–2019 from MSM in Coast ($N = 178$, 47.9%), Nairobi ($N = 137$, 36.8%), and Nyanza ($N = 57$, 15.3%) provinces in Kenya. Maximum-likelihood (ML) phylogenetics and Bayesian inference were used to determine HIV-1 clusters, evolutionary dynamics, and virus migration rates between geographic regions. HIV-1 sub-subtype A1 (72.0%) was most common followed by subtype D (11.0%), unique recombinant forms (8.9%), subtype C (5.9%), CRF 21A2D (0.8%), subtype G (0.8%), CRF 16A2D (0.3%), and subtype B (0.3%). Forty-six clusters (size range 2–20 sequences) were found—half (50.0%) of which had evidence of extensive HIV-1 mixing among different provinces. Data revealed an exponential increase in infections among MSM during the early-to-mid 2000s and stable or decreasing transmission dynamics in recent years (2017–2019). Phylogeographic inference showed significant (Bayes factor, BF > 3) HIV-1 dissemination from Coast to Nairobi and Nyanza provinces, and from Nairobi to Nyanza province. Strengthening HIV-1 prevention programs to MSM in geographic locations with higher HIV-1 prevalence among MSM (such as Coast and Nairobi) may reduce HIV-1 incidence among MSM in Kenya.

**Keywords: HIV-1, molecular epidemiology, phylogeographic, MSM, Kenya**

## INTRODUCTION

In sub-Saharan Africa, the HIV-1 epidemic among men who have sex with men (MSM) has only recently received attention—and the role of MSM in HIV-1 transmission has been acknowledged (Beyrer et al., 2010; Sanders et al., 2015a; Nduva et al., 2021). In Kenya, the national HIV-1 prevalence is 4.9% in the adult population, but is threefold higher in MSM than in heterosexual men (Kenya National Aids Control Council, 2019; National AIDS and STI Control Programme [NASCOP], 2020). HIV-1 prevalence among MSM in Kenya varies between regions—and ranges from 17.8% in Kisumu (Western Kenya) (Kunzweiler et al., 2017) to 24.5% in Coastal Kenya (Sanders et al., 2007), and from 25.0 to 26.4% in Nairobi (Smith et al., 2021a,b). There is evidence of high mobility of MSM sex workers between regions, which could link HIV-1 transmissions in different regions (Geibel et al., 2008). The Ministry of Health in Kenya through the National AIDS Control Council (NACC) has made efforts to strengthen HIV healthcare services for MSM (Gruskin and Tarantola, 2008; van der Elst et al., 2020). Yet, stigma against male-same-sex practices and policies criminalizing consensual same-sex sexual practices have obstructed progress (Cohen et al., 2013; van der Elst et al., 2013, 2020). In the past, geographic mobility has been shown to play an important role in HIV-1 dispersal (Faria et al., 2014; Grabowski et al., 2020). Taken together, it is possible that spatial differences in the HIV-1 distribution in Kenya combined with geographically mobile MSM sex workers could impact HIV-1 spread among MSM throughout the country (Faria et al., 2014; Grabowski et al., 2020). However, clear data on HIV-1 transmission dynamics within and between MSM in different geographic regions are lacking in Kenya.

HIV-1 transmission dynamics can be assessed by linking sociodemographic, clinical, and behavioral data with HIV-1 sequence data through phylogenetics (Brenner et al., 2007; Volz et al., 2013; Bruhn et al., 2014; Frentz et al., 2014; Pybus et al., 2015; Esbjörnsson et al., 2016; Poon et al., 2016; Ratmann et al., 2016; Hassan et al., 2017; Sallam et al., 2017). While limited HIV-1 sequences have been obtained from blood plasma from MSM living with HIV in Kenya, phylogenetic determination of patterns of HIV-1 transmission among Kenyan MSM suggests extensive MSM HIV-1 clustering (and infrequent HIV-1 mixing between MSM and presumed heterosexuals in the general population) (Bezemer et al., 2014; Hassan et al., 2018; Nduva et al., 2020, 2021, in press). In addition, a phylogenetic study in 2013 reported frequent HIV-1 gene flow between MSM in Coastal Kenya and Nairobi—albeit with small sample size and limited geographic coverage (Bezemer et al., 2014). In the period 2005–2019, more MSM HIV-1 sequences have become available from diverse geographical locations in Kenya, allowing in-depth characterization of evolutionary dynamics in the MSM HIV-1 epidemic in Kenya. Here, we used HIV-1 *pol* data to phylodynamically infer HIV-1 transmission rates among MSM in three different geographic regions in Kenya.

## MATERIALS AND METHODS

### Study Population

New sequences were generated from blood plasma obtained through studies conducted through the MSM Health Research Consortium—a multisite collaboration between researchers affiliated with KEMRI-Wellcome Trust (KWTRP) in Coastal Kenya, Nyanza Reproductive Health Society (NRHS) in Nyanza, and Sex Workers Outreach Program (SWOP) clinics in Nairobi. These included samples from Coast derived from participants in a prospective observational cohort (2006–2019) (Sanders et al., 2013), samples from Nairobi from a respondent-driven sample survey (Transform, 2017; Smith et al., 2021b), and samples from Nyanza derived from the Anza Mapema cohort (2015–2017) (Kunzweiler et al., 2018).

### HIV-1 *Pol* Sequence Dataset

The HIV-1 *pol* sequences were comprised of 1,020 nucleotides, HXB2 [K03455] positions 2267–3287. HIV-1 RNA was purified from patient blood plasma using the RNeasy Lipid Tissue Mini Kit (QIAGEN) as previously described (Esbjörnsson et al., 2010). Reverse transcription and amplification of partial *pol* gene were performed using the One-Step Superscript III RT/Platinum Taq High Fidelity Enzyme Mix (Thermo Fisher Scientific$^{TM}$) with the *pol*-specific primer pair JA269 and JA272 (Hedskog et al., 2010). First-round PCR products were amplified in a nested PCR with DreamTaq Green DNA Polymerase (Thermo Fisher Scientific$^{TM}$) using *pol*-specific primers JA271 and JA270 (Hedskog et al., 2010). PCR products were sequenced in both directions with the nested PCR primers using the BigDye terminator kit v1.1 (Applied Biosystems), and the sequences were determined on an ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems).

Additional Kenyan HIV-1 *pol* sequences (referred to as published sequences, 2006–2019) were retrieved (October 11th 2021) from the Los Alamos HIV-1 sequence database (Los Alamos National Laboratory, 2019). The combined new and published sequences (referred to as the Kenyan dataset) were annotated with information on sampling dates and geographical area of residence during sampling (i.e., province; Coast, Nairobi, and Nyanza).

### HIV-1 Subtyping

The Kenyan dataset was aligned with the HIV-1 Group M (subtypes A-K + Recombinants) subtype reference dataset (available at the Los Alamos HIV database)[1] using the MAFFT algorithm in Geneious Prime 2019 (Larkin et al., 2007; Los Alamos National Laboratory, 2019). The resulting alignment was used to construct a maximum-likelihood phylogenetic tree in PhyML using the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR+Γ4+I) (Guindon et al., 2005). Branch support was assessed using the Shimodaira–Hasegawa like approximate likelihood ratio test (aLRT-SH) in PhyML, with aLRT-SH $\geq$ 0.90 considered as significant (Anisimova and Gascuel, 2006; Hassan et al., 2017). Subtypes were assigned based on the Subtype/CRF-resolved phylogeny visualized using FigTree v1.4.4[2]. Subtype assignment was further verified using the REGA HIV-1 Subtyping Tool (v.3.0), and unique recombinant forms (URFs) were detected using the jumping profile hidden

[1] http://www.hiv.lanl.gov

[2] https://github.com/rambaut/figtree/releases

Markov model (jpHMM) (Schultz et al., 2009; Pineda-Peña et al., 2013).

## HIV-1 Cluster Analysis

Sequences were grouped into subtype-specific datasets, and a search for related sequences was done for each subtype-specific (A1, C, and D) dataset using the NCBI GenBank BLAST tool, limiting results to the 10 most similar hits per sequence, and retaining the oldest sequence per individual (Kouyos et al., 2010; Esbjörnsson et al., 2016; Sallam et al., 2017). Kenyan sequences and reference sequences were combined and aligned using the MAFFT algorithm in Geneious Prime 2019 (Larkin et al., 2007). Subtype-specific alignments were edited to exclude codon positions associated with drug resistance, and maximum-likelihood phylogenies were reconstructed in PhyML. For each subtype, monophyletic clades with aLRT-SH support $\geq$0.9 and which were dominated ($\geq$80%) by Kenyan sequences (compared to reference sequences) were defined as Kenyan HIV-1 clusters (Hassan et al., 2017). Clusters were classified based on the number of sequences per cluster into dyads (2 sequences), networks (3–14 sequences), and large clusters (>14 sequences) (Esbjörnsson et al., 2016; Nduva et al., 2020; Abidi et al., 2021).

## Bayesian Phylodynamic and Discrete Phylogeographic Inference

To date clusters and to estimate the effective population size through time ($N_{e.T}$), Bayesian phylodynamic inference was performed in BEAST 1.10.4 using the Bayesian Skygrid model, an uncorrelated lognormal relaxed clock, and the general time-reversible substitution model with a gamma-distributed rate variation and proportion of invariant sites (GTR+Γ4+I) (Drummond et al., 2005; Baele et al., 2012; Gill et al., 2013; Suchard et al., 2018). Only sequences classified as pure A1, C, and D subtypes were analyzed. BEAST runs were computed with a chain length of 100–300 million generations for each dataset, sampling every 10,000th–30,000th iteration, and discarding the first 10% as burn-in. Convergence was determined in Tracer v.1.7.0 and defined as effective sample sizes (ESS) $\geq$100 (Suchard et al., 2018). Maximum clade credibility (MCC) trees were summarized using Tree-Annotator v1.8.2 (BEAST suite).

To infer the direction of virus movements between geographic locations from HIV-1 sequence data, a discrete phylogeographic inference was computed using specific locations as independent discrete states (Lemey et al., 2009; Edwards et al., 2011; Faria et al., 2014). Several sensitivity analyses were performed to test the robustness of our data. Firstly, the Kenyan dataset was grouped by subtype (A1, C, and D), and the phylogeographic inference was performed using all the sequences per subtype. Secondly, to reduce sampling bias arising from the unproportionable allocation of sequences per location, sequences in the subtype A1-specific dataset (the largest of the three subtypes) were randomized and subsampled into a dataset with an equal number of sequences per province using in-house Perl scripts (available upon request). Lastly, subtype A1 sequences from Coast province were subsampled uniformly and used to estimate virus migration between three geographically distinct regions in Coastal Kenya (i.e., Mombasa, South Coast, and North Coast).

In the phylogeographic inference, the asymmetric model was adopted (over the alternative symmetric model) as it relaxes the assumption of constant diffusion rates through time to realistically model the location-exchange processes (Edwards et al., 2011; Faria et al., 2014). In addition to estimating the direction of HIV-1 migration, the proportions of forward and reverse rates of migrations between geographic locations were quantified using a robust counting approach (Markov jumps) implemented in BEAST (Minin and Suchard, 2008). Maximum clade credibility (MCC) trees annotated with demographic and epidemiological data were summarized in Tree-Annotator v1.10.4 (BEAST suite) and visualized in FigTree (v1.4.4). Well-supported virus movements and Bayes factors (BFs) assessing statistical support were summarized using SPREAD v1.0.7, and BF $\geq$ 3 was considered significant (Lemey et al., 2009).

## Statistical Analysis

Continuous data were presented using medians and interquartile ranges (IQRs). Frequencies and percentages were used to describe categorical data. A multivariable logistic regression model was used to assess associations between individual sequence characteristics [e.g., subtype, location of sampling, year (range) of sampling, and source of sequence data—i.e., published or newly generated] and phylogenetic clustering. Statistics and summary plots were done using Stata 15 (StataCorp LLC, College Station, TX, United States) and RStudio (version 1.2.5001) with the packages *yarrr* and *ggplot2* (Wickham, 2016; Phillips, 2017).

## Ethical Consideration

Plasma samples used to generate the new sequences were obtained from ongoing or concluded studies that were also approved by Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU 3747, 3280, and 3520, and SSC 894). Since published sequences were obtained from an open-access public domain, informed consent was not retrospectively obtained. Instead, we sought approval through a study protocol that was reviewed by KEMRI/SERU (SERU 3547).

# RESULTS

## Study Population, Sequence Dataset, and Subtype Distribution

Among the 372 HIV-1 partial *pol* sequences analyzed, 213 (57.3%) were generated in this study, and 159 (42.7%) were previously published. The majority ($N$ = 178, 47.9%) of the sequences were from Coast province, 137 (36.8%) from Nairobi province, and 57 (15.3%) from Nyanza province (**Figure 1**, **Table 1**, **Supplementary Figure 1**, and **Supplementary Tables 1**, **2**). Sequences belonged to sub-subtype A1 ($N$ = 268, 72.0%), subtype D ($N$ = 41, 11.0%), subtype C ($N$ = 22, 5.9%), subtype G ($N$ = 3, 0.8%), CRF 21A2D ($N$ = 3, 0.8%), CRF 16A2D ($N$ = 1, 0.3%), and subtype B

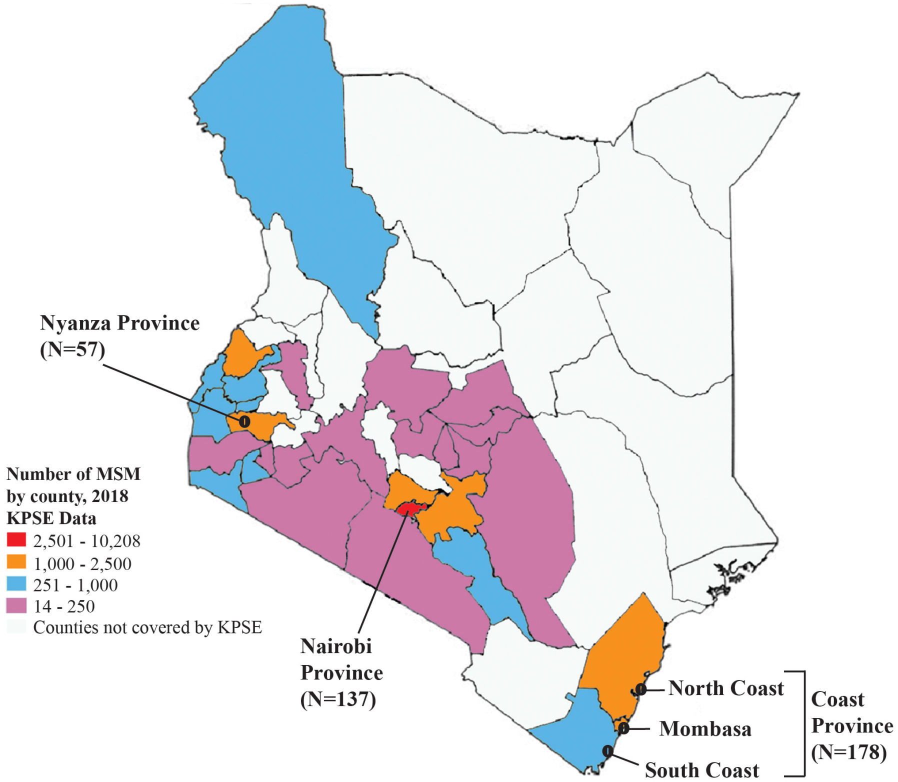


**FIGURE 1 |** Map of Kenya showing the distribution of sequences in this study. A map of Kenya showing the number of HIV-1 sequences from MSM analyzed in this study, and distribution by different geographic regions. The map is colored based on the estimated number of MSM as mapped at the county level during the 2018 key population size estimates national survey (National AIDS and STI Control Programme [NASCOP], 2019).

(*N* = 1, 0.3%). Unique recombinant forms (URFs) identified included A1D (*N* = 19, 5.1%), A1C (*N* = 7, 1.9%), D01AE (*N* = 5, 1.3%), A1B (*N* = 1, 0.3%), and DB (*N* = 1, 0.3%, **Figure 2**).

## Men Who Have Sex With Men HIV-1 Clusters

Clusters were determined from maximum-likelihood (ML) phylogenies reconstructed for the most prevalent HIV-1 subtypes in the population [subtypes A (A1), C, and D—cumulatively comprising 89.0% of the sequences in the Kenyan dataset]. Non-Kenyan HIV-1 reference sequences were obtained from GenBank based on similarity (where of 931 participant-unique sub-subtype A1 sequences remained after removal of redundancies; 488 for subtype C; and 350 for subtype D). Of 331 (A1, C, and D) sequences in the cluster analysis, 229 sequences (61.2%) formed 46 statistically supported clusters (size range: 2–20 sequences). Dyad/pairs were most common (*N* = 25, 54.4% of all clusters), followed by networks having 3–14 sequences (*N* = 18, 39.1%), and large clusters having more than 14 sequences (*N* = 3, 6.5%). The majority (*N* = 34, 73.9%) were sub-subtype A1 clusters, followed by subtype D (*N* = 8, 17.4%) and subtype C (*N* = 4, 8.7%, **Table 2** and **Supplementary Figure 2**).

## Geographic Stratification of Clustering Patterns

Stratification of clusters by geographic regions showed two distinct clustering patterns. First, some clusters (*N* = 23, 50.0%) had sequences belonging exclusively to one specific province including Coast (*N* = 14, 30.4%), Nairobi (*N* = 6, 13.0%), and Nyanza (*N* = 3, 6.5%) province-exclusive clusters. The remaining clusters (*N* = 23, 50.0%) were mixed between different provinces where HIV-1 mixing between Coast and Nairobi was most common (*N* = 13, 28.3% clusters), followed by mixing between Nyanza, Nairobi, and Coast (*N* = 5, 10.9%), Nyanza and Nairobi (*N* = 3, 6.5%), and Nyanza and Coast (*N* = 2, 4.4%, **Table 2** and **Supplementary Figure 2**). Sequences from Nairobi province were more likely to cluster compared to sequences from Coast province [adjusted odds ratio (aOR) 3.5, 95% confidence interval (CI) 1.2–10.4, *P* = 0.022, **Table 3**].

## Estimating Effective Population Size Through Time and Dating Clusters

In-depth phylodynamic analysis indicated that the number of MSM contributing to new HIV-1 A1 infections over time increased exponentially during the early 2000s, followed by a period with some fluctuation (but largely

**TABLE 1 |** Distribution of newly generated and published HIV-1 *pol* sequences (*N* = 372) from Kenyan MSM, overall, and by geographic location.

| **Category** | **Number of sequences (*N*, %)** | | | |
|---|---|---|---|---|
| **Geographic region** | **Coast** | **Nairobi** | **Nyanza** | **Total** |
| **Year (range)** | | | | |
| 2006–2010 | 117 (65.7%) | 1 (0.7%) | 0 (0.0%) | 118 (31.7%) |
| 2011–2015 | 32 (18.0%) | 1 (0.7%) | 19 (33.3%) | 52 (14.0%) |
| 2016–2019 | 29 (16.3%) | 135 (98.5%) | 38 (66.7%) | 202 (54.3%) |
| **Sequences** | | | | |
| New | 21 (11.8%) | 135 (98.5%) | 57 (100%) | 213 (57.3%) |
| Published | 157 (88.2%) | 2 (1.5%) | 0 (0.0%) | 159 (42.7%) |
| **Subtype** | | | | |
| A1 | 121 (68%) | 102 (74.5%) | 45 (79%) | 268 (72%) |
| D | 22 (12.4%) | 13 (9.5%) | 6 (10.5%) | 41 (11%) |
| URF | 16 (9%) | 14 (10.2%) | 3 (5.3%) | 33 (8.9%) |
| C | 14 (7.9%) | 5 (3.7%) | 3 (5.3%) | 22 (5.9%) |
| 21A2D | 0 (0%) | 3 (2.2%) | 0 (0%) | 3 (0.8%) |
| G | 3 (1.7%) | 0 (0%) | 0 (0%) | 3 (0.8%) |
| 16A2D | 1 (0.6%) | 0 (0%) | 0 (0%) | 1 (0.3%) |
| B | 1 (0.6%) | 0 (0%) | 0 (0%) | 1 (0.3%) |
| Total | 178 (47.9%) | 137 (36.8%) | 57 (15.3%) | 372 (100%) |

*MSM, men who have sex with men; URF, unique recombinant form; CRF, circulating recombinant form.*

steady) between 2000 and 2017, and mostly decreasing dynamics during recent years (2017–2019, **Figure 3A**). Likewise, for both subtype C and D lineages, the effective population size increased exponentially during 2007–2008 and has stabilized in recent years (2016–2019; **Figures 3B–D**).

Estimating dates of origins of all clusters indicated that the majority (65%) of transmissions within clusters took place between 2000 and 2014. The oldest sub-subtype A1 cluster had nine MSM from Nyanza, Nairobi, and Coast and had originated during 1987, while the youngest cluster was dated to 2014 among MSM in Nyanza (**Figure 4A**, **Supplementary Figure 3**, and **Supplementary Table 3**). The largest A1 cluster (*N* = 20, 2008–2017) had remained active over 20 years since the estimated time to the most recent common ancestor (tMRCA) in 1997 and was geographically spread out to Nyanza, Nairobi, and Coast provinces. The second-largest A1 cluster (*N* = 19, 2008–2017) originated in 1996 and had sequences from Nyanza, Nairobi, and Coast provinces. The four subtype C clusters originated during 1988, 1998, 2009, and 2014, respectively, while the earliest subtype D cluster originated during 1976 and the youngest during 2014 (**Figures 4B,C** and **Supplementary Table 3**). Overall, there was evidence of onward HIV-1 transmission among MSM, within longstanding and geographically diverse HIV-1 networks.

## HIV-1 Migration Between Provinces in Kenya

Ancestral locations and rates in historical virus jumps were first estimated based on all subtype-specific sequences in the Kenyan dataset (i.e., 268 sub-subtype A1, 41 subtype D, and 22 subtype C sequences). Phylogeographic analysis indicated significant support (Bayes Factor, BF ≥ 3) for virus migration from Coast to Nairobi (BF = 3716; subtype A1, BF = 268; subtype C; and BF = 16; subtype D) and from Nairobi to Nyanza (BF = 3716; subtype A1, BF = 43; subtype D, **Supplementary Table 4**). Exploring temporal trends in virus transitions between geographic provinces summarized from trait-annotated maximum clade credibility trees indicated that the proportion of virus export from Coast to Nairobi increased from 4.2% before 2000 to 14.2% during 2001–2010 and declined to 4.9% during 2011–2020. Likewise, virus export from Nairobi to Nyanza increased from 2.4% in 2000–2010 to 10.8% in 2011–2020, while reverse transitions were rare and occurred only from Nyanza to Nairobi (**Supplementary Figures 4**, **5** and **Supplementary Table 5**).

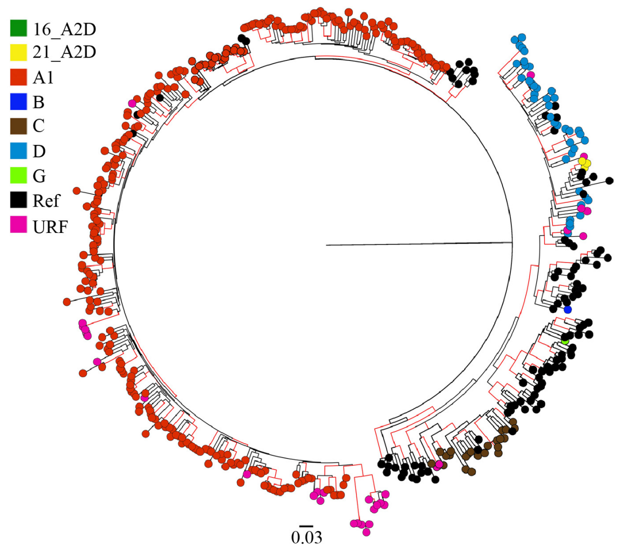


**FIGURE 2 |** HIV-1 genotypes among 372 MSM sequences from Kenya. Maximum-likelihood phylogenetic tree of 372 HIV-1 *pol* sequences from MSM living with HIV-1 in Kenya (and 194 HIV-1 Group M subtype reference sequences from the Los Alamos HIV database). Branch tips colors correspond to the respective HIV-1 subtype, sub-subtype, or recombinant form as shown in the legend. Branches with aLRT-SH support of more than ≥0.9 are colored red. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

A sensitivity analysis with uniform sampling per province was performed to confirm the robustness of the initial phylogeographic inference. The uniformly subsampled dataset comprised 135 HIV-1 sub-subtype A1 sequences (45 sequences each from Nairobi, Mombasa, and Nyanza province). Based on this analysis, there was significant support for HIV-1 migration from Coast to Nairobi (BF = 7766), Nairobi to Nyanza (BF = 1293), and Coast to Nyanza (BF = 336, **Table 4**). Furthermore, Markov jumps estimates with uniform sampling indicated that the majority (80.3%) of HIV-1 jumps between provinces occurred from Coast to other provinces including jumps from Coast to Nyanza (*N* = 26, 42.6% of all virus jumps between provinces) and from Coast to Nairobi

**TABLE 2 |** The number of Kenyan MSM HIV-1 clusters by cluster size and geographic region.

| | Dyads (2 sequences) | Networks (3–14) | Large clusters (≥14) | Total clusters |
|---|---|---|---|---|
| **Subtype** | | | | |
| A1 | 12 (66.7%) | 19 (76.0%) | 3 (100%) | 34 (73.9%) |
| C | 2 (11.1%) | 2 (8.0%) | 0 (0.0%) | 4 (8.7%) |
| D | 4 (22.2%) | 4 (16.0%) | 0 (0.0%) | 8 (17.4%) |
| **Geographic region** | | | | |
| Coast | 6 (24.0%) | 8 (44.4%) | 0 (0.0%) | 14 (30.4%) |
| Coast/Nairobi | 11 (44.0%) | 2 (11.1%) | 0 (0.0%) | 13 (28.3%) |
| Nairobi | 2 (8.0%) | 4 (22.2%) | 0 (0.0%) | 6 (13.0%) |
| Nyanza/Nairobi/Coast | 2 (8.0%) | 0 (0.0%) | 3 (100%) | 5 (10.9%) |
| Nyanza | 0 (0.0%) | 3 (16.67%) | 0 (0.0%) | 3 (6.5%) |
| Nyanza/Nairobi | 3 (12.0%) | 0 (0.0%) | 0 (0.0%) | 3 (6.5%) |
| Nyanza/Coast | 1 (4.0%) | 1 (5.56%) | 0 (0.0%) | 2 (4.4%) |
| Total | 25 (54.4%) | 18 (39.1%) | 3 (6.5%) | 46 (100%) |

*MSM, men who have sex with men. Clusters were classified based on the number of sequences per cluster into dyads (2 sequences), networks (3–14 sequences), and large clusters (>14 sequences).*

**TABLE 3 |** Factors associated with HIV-1 clustering among MSM with HIV-1 in Kenya.

| Characteristics | | Multivariate analysis |
|---|---|---|
| | | *aOR, (95% CI), *p*-value |
| Year (range) | 2006–2010 | Reference |
| | 2011–2015 | 1.0 (0.4–2.2), 0.937 |
| | 2016–2020 | 1.1 (0.3–3.4), 0.932 |
| Subtype | A1 | Reference |
| | C | 0.6 (0.2–1.5), 0.258 |
| | D | 1.0 (0.5–2.0), 0.884 |
| Province | Coast | Reference |
| | Nairobi | 3.5 (1.2–10.4), 0.022 |
| | Nyanza | 1.8 (0.5–5.9), 0.34 |
| Sequence | Published | Reference |
| | Newly generated | 2.5 (1.7–4.0), <0.001 |

*MSM, men who have sex with men; *aOR, adjusted odds ratio.*

($N$ = 23, 37.7%, **Figure 5** and **Table 5**). There was also some ($N$ = 10, 16.4%) virus exchange between Nairobi and Nyanza, such that virus jump Nairobi to Nyanza ($N$ = 7, 11.5%) was twofold higher than from Nyanza to Nairobi ($N$ = 3, 4.9%, **Table 5**).

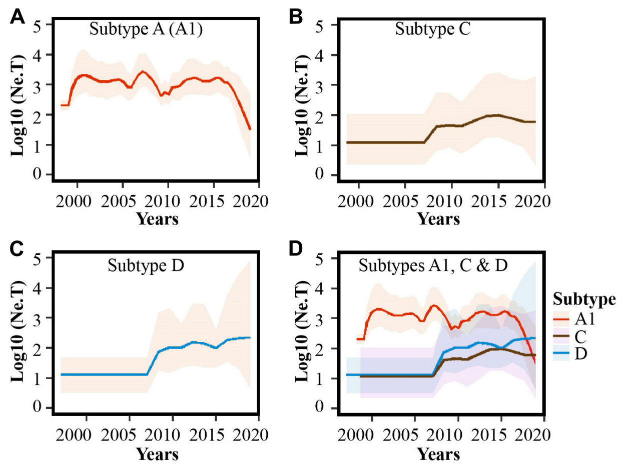


**FIGURE 3 |** Population dynamics of HIV-1 sub-subtype A1, subtype D, and subtype C lineages among MSM in Kenya. Bayesian Skygrid plots showing population dynamics of the **(A)** HIV-1 sub-subtype A1, **(B)** HIV-1 subtype C, **(C)** HIV-1 subtype D lineages, and **(D)** combined plots for HIV-1 A1, C and D lineages in Kenyan MSM. Median estimates of the number of MSM contributing to new infections are shown as a continuous line in each plot (colored red for sub-subtype A1, brown for subtype C, and blue for subtype D). The shaded area represents the 95% higher posterior density intervals of the inferred effective population size for each lineage.

## DISCUSSION

We found high rates of HIV-1 geographic mixing and a high proportion of HIV-1 sequences exported from the Coast and Nairobi to Nyanza province—implying that the Coast and Nairobi provinces could be major geographic sources of HIV-1 transmission among Kenyan MSM. Of all provinces in Kenya, the Coast and Nairobi provinces have the highest prevalence of HIV-1 among MSM (Kenya National Aids Control Council, 2009). In addition, MSM in Coastal Kenya are known to be highly mobile, and some engage in sex work in different locations across the country (Geibel et al., 2008). Taken together, our findings suggest that regions with the highest HIV-1 prevalence among MSM (such as Coast and Nairobi) may also have disseminated HIV-1 disproportionately to regions with lower HIV-1 prevalence among MSM (such as Nyanza province) in Kenya. s

There are a few presumed mechanisms by which Coastal Kenya may serve as an important source of infections among MSM. One plausible explanation might be that as a very well recognized destination for domestic tourism and sex tourism, MSM (or non-disclosing HET) visit the area for sex tourism, effectively disseminating the virus upon returning from

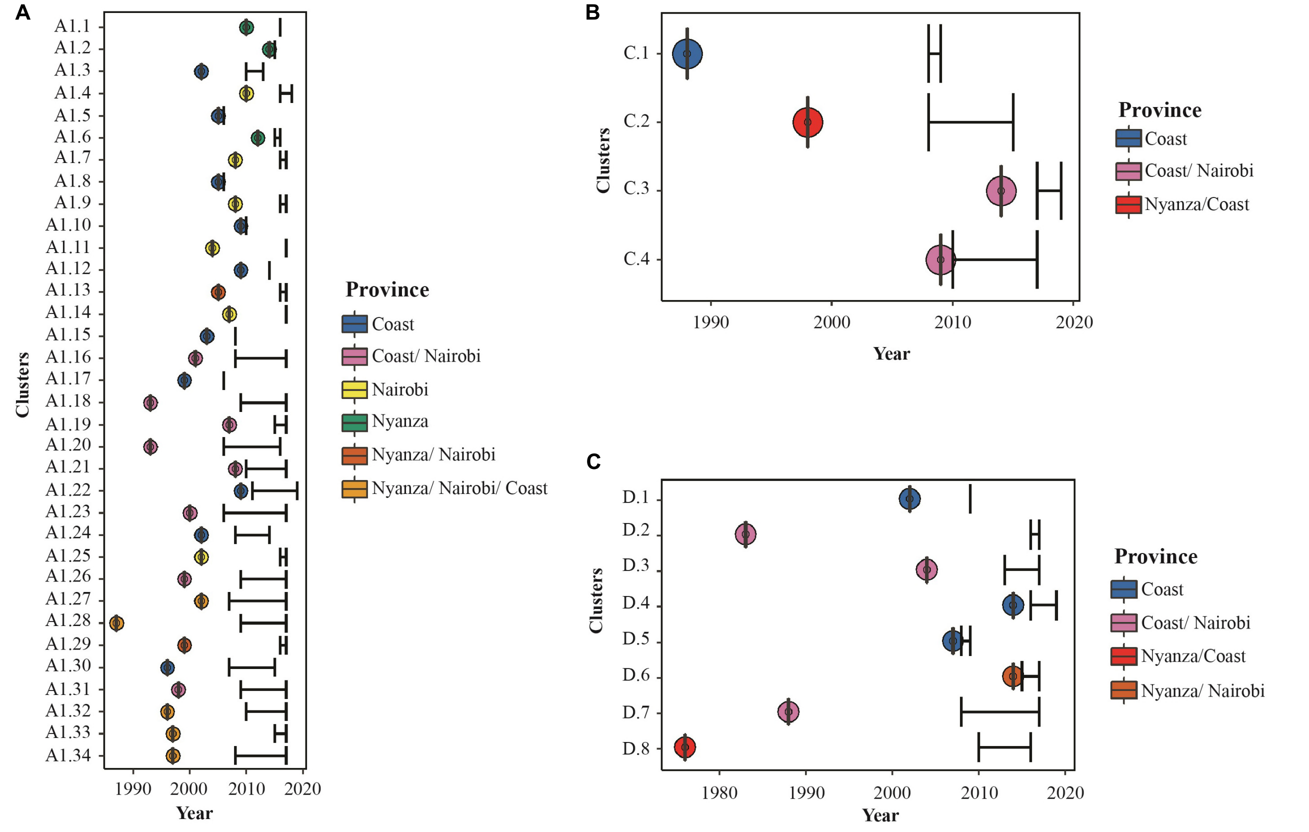


**FIGURE 4 |** Characteristics and posterior distribution of time to most recent common ancestors estimated for all Kenya clusters. Bayesian tMRCA estimates for **(A)** HIV-1 sub-subtype A1, **(B)** HIV-1 subtype C, and **(C)** HIV-1 subtype D lineages in Kenyan MSM HIV-1 clusters. Dots represent the estimated tMRCA and are colored as per the provinces represented by sequences in each cluster as shown in the legend. Black error bars represent sampling time (with lower interval representing the oldest sampling time per cluster and upper interval representing the most recent sampling time per cluster).

**TABLE 4 |** HIV-1 migration rates (Bayes factor, BF ≥ 3) between geographic locations in Kenya.

| The direction of migration events (from, to) | Bayes factor (BF) | Posterior probability |
|---|---|---|
| **Migration between provinces** | | |
| Coast-to-Nairobi | 7766 | 1 |
| Nairobi-to-Nyanza | 1293 | 1 |
| Coast-to-Nyanza | 336 | 1 |
| Nyanza-to-Nairobi | 3 | 0.7 |
| Nyanza-to-Coast | 3 | 0.7 |

Coast. A second potential determinant could be connected to geographically mobile MSM sex workers—hypothetically, HIV-1 may first be acquired and/or amplified in the Coast, and then exported to other provinces. Thus, the regional difference observed could potentially reflect amplification behavior within Coastal Kenya—and onward spread to other provinces linked to an MSM migration gradient. Data on migration were not available during the current analysis, but future studies may investigate this in detail. Future studies may also potentially investigate potential underlying demographic transitions—speculatively, young MSM sex workers may be drawn to Coast province while older or socially privileged MSM or MSM sex workers may leave the region for other provinces. Overall, implementing HIV-1 prevention and care directed to MSM in Kenya (and considering areas with higher rates of HIV-1 dissemination such as Coast and Nairobi) might reduce ongoing HIV-1 transmission at a countrywide scale, as has been shown in other settings (Bailey et al., 2007; Anderson et al., 2014; Gerberry et al., 2014; McGillen et al., 2016).

The majority (61.2%) of sequences analyzed in this study formed phylogenetically linked HIV-1 clusters, consistent with multiple introductions and ongoing infections among MSM within close networks in Kenya (Skar et al., 2011; Esbjörnsson et al., 2016; Sallam et al., 2017). Half of the clusters comprised sequences collected from MSM from different geographic regions—indicating geographically extensive HIV-1 linkages. High rates of clustering involving HIV-1 in MSM have been reported both in our setting and other higher-income settings and could be linked to an increased risk of infection among MSM within close networks, involving geographically mobile individuals (Geibel et al., 2008; Bezemer et al., 2014; Esbjörnsson et al., 2016; Sallam et al., 2017;

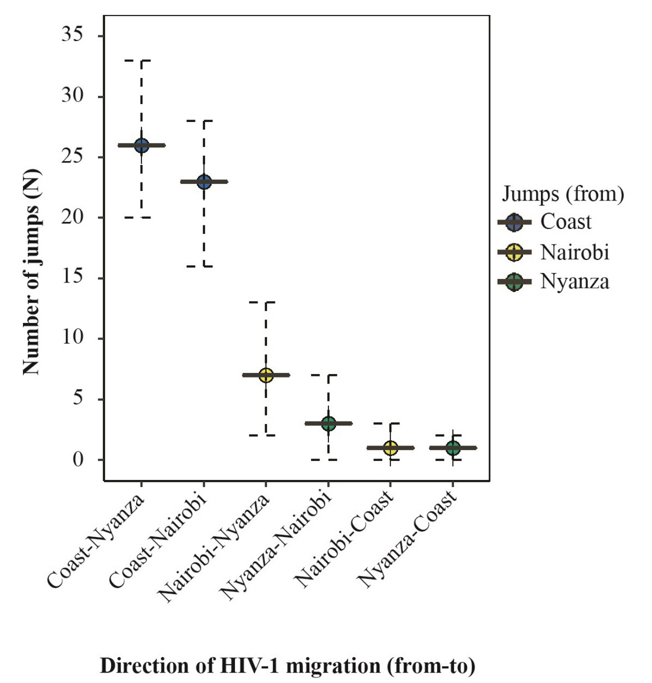


**FIGURE 5 |** Summary of the expected number of HIV-1 migration between geographic regions in Kenya. Summary of the median number (and 95% HPD interval) of Markov jumps inferred with a uniform sampling of geographic regions. Plots represent HIV-1 exchange between provinces. Plots are colored by the "source" location as shown in the legend. Only statistically significant transitions [Bayes Factor (BF) ≥ 3] are plotted.

**TABLE 5 |** The number of expected (Markov) jumps inferred for HIV-1 A1 migration between geographic locations.

| The direction of migration events (from, to) | Number of HIV-1 jumps (*N*, %) |
|---|---|
| **Between provinces** | **61 (100%)** |
| Coast–Nyanza | 26 (42.6%) |
| Coast–Nairobi | 23 (37.7%) |
| Nairobi–Nyanza | 7 (11.5%) |
| Nyanza–Nairobi | 3 (4.9%) |
| Nairobi–Coast | 1 (1.6%) |
| Nyanza–Coast | 1 (1.6%) |

Hassan et al., 2018). We estimated that a high proportion (65%) of HIV-1 transmissions occurred between 2000 and 2014 and that several clusters extended over multiple years, suggesting onward HIV-1 transmission among MSM within geographically diverse HIV-1 networks. HIV-1 sequences in this study were not closely related to reference sequences from the global epidemic, implying that the HIV-1 epidemic among MSM in Kenya is sustained locally.

In a broader context, several phylogenetic studies have revealed that the HIV-1 epidemic in Kenya is compartmentalized—where the majority of HIV-1 transmission occurs within risk groups (Bezemer et al., 2014; Nduva et al., 2020, in press). Our recent work at a countrywide scale has demonstrated a minor (8%) proportion of HIV-1 MSM and heterosexual clustering (Nduva et al., in press). Taken together, these studies indicate that ongoing transmission among MSM rarely impacts the general heterosexual HIV-1 epidemic in Kenya. MSM in Kenya have a high burden of HIV risk—to reduce overall HIV-1 incidence in Kenya, there is a need to implement directed HIV-1 prevention and treatment to MSM in Kenya.

The phylodynamic analysis investigating the evolutionary dynamics of the HIV-1 MSM sub-epidemic revealed an exponential increase in the number of infections during the early-to-mid 2000s (for HIV-1 A1, C, and D lineages)—indicative of multiple HIV-1 outbreaks among Kenyan MSM (Skar et al., 2011; Esbjörnsson et al., 2016; Sallam et al., 2017). Interestingly, the effective population size did not decrease following the nationwide introduction and scale-up of combination antiretroviral therapy (ART) in 2004. One potential reason for this is suboptimal access to HIV-1 treatment and prevention services by MSM in Kenya due to fear of legal and social stigma and discrimination (Micheni et al., 2015; Kenya National Aids Control Council, 2019; Stannah et al., 2019). Nevertheless, the effective population size for the dominant strain (HIV-1 A1) showed fewer new infections in recent years (2017–2019)—possibly reflecting earlier ART initiation due to changes in treatment recommendations (National AIDS and STDS Control Programme Ministry of Health, 2016) as well as some impact of risk reduction counseling, adherence support interventions (Möller et al., 2015; Graham et al., 2020), early recognition of acute HIV-1 infections, especially on the Kenyan Coast (Sanders et al., 2014, 2015b; Mugo et al., 2016), and some uptake of pre-exposure prophylaxis targeting MSM in recent years (Graham et al., 2015; van der Elst et al., 2015; Wahome et al., 2018; Kimani et al., 2019). Overall, increasing access to treatment, as well as destigmatization and diversification of providers, may further reduce HIV-1 incidence among MSM (Smith et al., 2021b).

The major strength of our study is the use of HIV-1 sequences from well-characterized acute and early infected MSM cohorts sampled over 14 years in a sub-Saharan African setting. A limitation is that the study had a small sample size, which limited the identification of HIV-1 links in the entire MSM HIV-1 epidemic in Kenya. Incomplete sampling likely resulted in missing links and reduced clustering of HIV-1 sequences (Novitsky et al., 2014). However, our sensitivity analyses before and after controlling for sampling bias indicated more jumps from Coastal Kenya to other provinces (and from Nairobi to Nyanza) than vice versa, indicating the robustness of the analyzed HIV-1 sequence dataset. Another limitation is skewed spatiotemporal sampling and variations in sampling methods between studies, which may have resulted in overrepresentation of some types of location-specific and/or subtype-specific clusters. Indeed, the HIV-1 C and HIV D lineages did not have a decreasing trend in recent years (2017–2019, compared to HIV-1 A1)—the reason for this could be related to skewed sampling over time in various geographic locations in this study. In addition, although the conflation of MSM and transgender people may have relevance for the distinction between sexual network types, we did not have data on gender identity—thus, some transgender people may have been conflated for MSM.

## CONCLUSION

We demonstrated extensive HIV-1 mixing among MSM in different regions in Kenya, where Coast and Nairobi provinces appear to have been a major source of virus dissemination. We hypothesize that MSM in these provinces may have disseminated HIV-1 disproportionately to MSM in other regions in the country. Increasing PrEP uptake and access to ART among MSM (and destigmatization and diversification of providers) is necessary to reduce ongoing HIV-1 transmission among MSM in Kenya.

## DATA AVAILABILITY STATEMENT

The data presented in the study are deposited in GenBank, accession numbers OM109723-OM109725, OM109756-OM109766, OM109772-OM109799, OM109814-OM109862, OM109879-OM109949, OM110011-OM110019, OM110126-OM110127, OM110136-OM110149, OM110169-OM110170, OM110171, OM110174, OM110178-OM110181, OM110193-OM110194, OM110212-OM110218, OM110229-OM110240, OM110245-OM110246, and OM110272-OM110282.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the plasma samples used to generate the new sequences were obtained from ongoing or concluded studies that were also approved by the Kenya Medical Research Institute (KEMRI) Scientific and Ethics Review Unit (SERU 3747, 3280, and 3520, and SSC 894). Since published sequences were obtained from an open-access public domain, informed consent was not retrospectively obtained. Instead, we sought approval through a study protocol that was reviewed by KEMRI/SERU (SERU 3547). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

GN, AH, ES, and JE conceptualized and designed the study, and provided funding for the study. FO, JK, LM, PS, SG, MP, AS, RB, and ES provided samples from which new sequences used in the study were generated. GN performed lab work and inferential analyses, produced all figures and tables, and wrote the manuscript. FC helped with virus sequencing. All authors reviewed, edited, and approved the manuscript for submission.

## FUNDING

This work was supported through the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative (grant # DEL-15-006). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (grant # 107752/Z/15/Z) and the United Kingdom Government. This work was also supported by funding from the Swedish Research Council (grant # 2016–01417) and Swedish Society for Medical Research (grant # SA-2016). This work was funded in part by IAVI and made possible by the support of many donors, including the United States Agency for International Development (USAID). The full list of IAVI donors is available at http://www.iavi.org.

## ACKNOWLEDGMENTS

We thank the staff affiliated with the MSM Health Research Consortium (MHRC) and IAVI for supporting studies involving key populations in Kenya. This manuscript was submitted for publication with permission from the Director of the Kenya Medical Research Institute (KEMRI).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.843330/full#supplementary-material

## REFERENCES

Abidi, S. H., Nduva, G. M., Siddiqui, D., Rafaqat, W., Mahmood, S. F., Amna, R., et al. (2021). Phylogenetic and drug-resistance analysis of HIV-1 sequences from an extensive paediatric HIV-1 outbreak in Larkana, Pakistan. *Front. Microbiol.* 12:658186. doi: 10.3389/fmicb.2021.658186

Anderson, S.-J., Cherutich, P., Kilonzo, N., Cremin, I., Fecht, D., Kimanga, D., et al. (2014). Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *Lancet* 384, 249–256. doi: 10.1016/S0140-6736(14)61053-9

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552. doi: 10.1080/10635150600755453

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., and Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* 29, 2157–2167. doi: 10.1093/molbev/mss084

Bailey, R. C., Moses, S., Parker, C. B., Agot, K., Maclean, I., Krieger, J. N., et al. (2007). Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* 369, 643–656.

Beyrer, C., Baral, S. D., Walker, D., Wirtz, A. L., Johns, B., and Sifakis, F. (2010). The expanding epidemics of HIV type 1 among men who have sex with men in low-and middle-income countries: diversity and consistency. *Epidemiol. Rev.* 32, 137–151. doi: 10.1093/epirev/mxq011

Bezemer, D., Faria, N. R., Hassan, A., Hamers, R. L., Mutua, G., Anzala, O., et al. (2014). HIV Type 1 transmission networks among men having sex with men and heterosexuals in Kenya. *AIDS Res. Hum. Retroviruses* 30, 118–126. doi: 10.1089/aid.2013.0171

Brenner, B. G., Roger, M., Routy, J. P., Moisi, D., Ntemgwa, M., Matte, C., et al. (2007). High rates of forward transmission events after acute/early HIV-1 infection. *J. Infect. Dis.* 195, 951–959. doi: 10.1086/512088

Bruhn, C. A., Audelin, A. M., Helleberg, M., Bjorn-Mortensen, K., Obel, N., Gerstoft, J., et al. (2014). The origin and emergence of an HIV-1 epidemic: from introduction to endemicity. *AIDS* 28, 1031–1040. doi: 10.1097/QAD.0000000000000198

Cohen, M. S., Smith, M. K., Muessig, K. E., Hallett, T. B., Powers, K. A., and Kashuba, A. D. (2013). Antiretroviral treatment of HIV-1 prevents transmission of HIV-1: where do we go from here? *Lancet* 382, 1515–1524. doi: 10.1016/S0140-6736(13)61998-4

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192. doi: 10.1093/molbev/msi103

Edwards, C. J., Suchard, M. A., Lemey, P., Welch, J. J., Barnes, I., Fulton, T. L., et al. (2011). Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr. Biol.* 21, 1251–1258. doi: 10.1016/j.cub.2011.05.058

Esbjörnsson, J., Månsson, F., Martínez-Arias, W., Vincic, E., Biague, A. J., da Silva, Z. J., et al. (2010). Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease-indication of an evolving epidemic in West Africa. *Retrovirology* 7:23. doi: 10.1186/1742-4690-7-23

Esbjörnsson, J., Mild, M., Audelin, A., Fonager, J., Skar, H., Bruun Jørgensen, L., et al. (2016). HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. *Virus Evol.* 2:vew010. doi: 10.1093/ve/vew010

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346, 56–61. doi: 10.1126/science.1256739

Frentz, D., van de Vijver, D., Abecasis, A., Albert, J., Hamouda, O., Jørgensen, L., et al. (2014). Patterns of transmitted HIV drug resistance in Europe vary by risk group. *PLoS One* 9:e94495. doi: 10.1371/journal.pone.0094495

Geibel, S., Luchters, S., King'Ola, N., Esu-Williams, E., Rinyiru, A., and Tun, W. (2008). Factors associated with self-reported unprotected anal sex among male sex workers in Mombasa, Kenya. *Sex. Transmitted Dis.* 35, 746–752. doi: 10.1097/OLQ.0b013e318170589d

Gerberry, D. J., Wagner, B. G., Garcia-Lerma, J. G., Heneine, W., and Blower, S. (2014). Using geospatial modelling to optimize the rollout of antiretroviral-based pre-exposure HIV interventions in Sub-Saharan Africa. *Nat. Commun.* 5, 1–15. doi: 10.1038/ncomms6454

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* 30, 713–724. doi: 10.1093/molbev/mss265

Grabowski, M. K., Lessler, J., Bazaale, J., Nabukalu, D., Nankinga, J., Nantume, B., et al. (2020). Migration, hotspots, and dispersal of HIV infection in Rakai, Uganda. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-020-14636-y

Graham, S. M., Micheni, M., Chirro, O., Nzioka, J., Secor, A. M., Mugo, P. M., et al. (2020). A randomized controlled trial of the Shikamana intervention to promote antiretroviral therapy adherence among gay, bisexual, and other men who have sex with men in Kenya: feasibility, acceptability, safety and initial effect size. *AIDS Behav.* 24, 2206–2219. doi: 10.1007/s10461-020-02786-5

Graham, S. M., Micheni, M., Kombo, B., Van Der Elst, E. M., Mugo, P. M., Kivaya, E., et al. (2015). Development and pilot testing of an intervention to promote care engagement and adherence among HIV-positive Kenyan MSM. *AIDS (London, England)* 29:S241. doi: 10.1097/QAD.0000000000000897

Gruskin, S., and Tarantola, D. (2008). Universal access to HIV prevention, treatment and care: assessing the inclusion of human rights in international and national strategic plans. *AIDS (London, England)* 22:S123. doi: 10.1097/01.aids.0000327444.51408.21

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005). PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* 33, W557–W559. doi: 10.1093/nar/gki352

Hassan, A. S., Esbjornsson, J., Wahome, E., Thiong'o, A., Makau, G. N., Price, M. A., et al. (2018). HIV-1 subtype diversity, transmission networks and transmitted drug resistance amongst acute and early infected MSM populations from Coastal Kenya. *PLoS One* 13:e0206177. doi: 10.1371/journal.pone.0206177

Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjörnsson, J. (2017). Defining HIV-1 transmission clusters based on sequence data. *AIDS (London, England)* 31:1211. doi: 10.1097/QAD.0000000000001470

Hedskog, C., Mild, M., Jernberg, J., Sherwood, E., Bratt, G., Leitner, T., et al. (2010). Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS One* 5:e11345. doi: 10.1371/journal.pone.0011345

Kenya National Aids Control Council (2009). *Kenya HIV Prevention Response and Modes of Transmission Analysis*. Nairobi: Kenya National AIDS Control Council.

Kenya National Aids Control Council (2019). *Kenya AIDS Strategic Framework 2014/2015–2018/2019, Nairobi*. Nairobi: Kenya National AIDS Control Council.

Kimani, M., van der Elst, E. M., Chiro, O., Oduor, C., Wahome, E., Kazungu, W., et al. (2019). Pr EP interest and HIV-1 incidence among MSM and transgender women in coastal Kenya. *J. Int. AIDS Soc.* 22:e25323. doi: 10.1002/jia2.25323

Kouyos, R. D., von Wyl, V., Yerly, S., Boni, J., Taffe, P., Shah, C., et al. (2010). Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* 201, 1488–1497. doi: 10.1086/651951

Kunzweiler, C. P., Bailey, R. C., Okall, D. O., Graham, S. M., Mehta, S. D., and Otieno, F. O. (2017). Factors associated with prevalent HIV infection among Kenyan MSM: the anza mapema study. *JAIDS J. Acquir. Immune Defic. Syndr.* 76, 241–249. doi: 10.1097/QAI.0000000000001512

Kunzweiler, C. P., Bailey, R. C., Okall, D. O., Graham, S. M., Mehta, S. D., and Otieno, F. O. (2018). Depressive symptoms, alcohol and drug use, and physical and sexual abuse among men who have sex with men in kisumu, kenya: the anza mapema study. *AIDS Behav.* 22, 1517–1529. doi: 10.1007/s10461-017-1941-0

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5:e1000520. doi: 10.1371/journal.pcbi.1000520

Los Alamos National Laboratory (2019). *HIV-1 Database at the Los Alamos National Laboratory*. Los Alamos, NM: Los Alamos National Laboratory.

McGillen, J. B., Anderson, S.-J., Dybul, M. R., and Hallett, T. B. (2016). Optimum resource allocation to reduce HIV incidence across sub-Saharan Africa: a mathematical modelling study. *Lancet HIV* 3, e441–e448. doi: 10.1016/S2352-3018(16)30051-0

Micheni, M., Rogers, S., Wahome, E., Darwinkel, M., van der Elst, E., Gichuru, E., et al. (2015). Risk of sexual, physical and verbal assaults on men who have sex with men and female sex workers in coastal Kenya. *AIDS (London, England)* 29, S231. doi: 10.1097/QAD.0000000000000912

Minin, V. N., and Suchard, M. A. (2008). Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* 56, 391–412. doi: 10.1007/s00285-007-0120-8

Möller, L. M. I, Stolte, G., Geskus, R. B., Okuku, H. S., Wahome, E., Price, M. A., et al. (2015). Changes in sexual risk behavior among MSM participating in a research cohort in coastal Kenya. *AIDS (London, England)* 29:S211. doi: 10.1097/QAD.0000000000000890

Mugo, P. M., Wahome, E. W., Gichuru, E. N., Mwashigadi, G. M., Thiong'o, A. N., Prins, H. A., et al. (2016). Effect of text message, phone call, and in-person appointment reminders on uptake of repeat HIV testing among outpatients screened for acute HIV infection in Kenya: a randomized controlled trial. *PLoS One* 11:e0153612. doi: 10.1371/journal.pone.0153612

National AIDS and STDS Control Programme Ministry of Health (2016). *Guidelines on Use of Antiretroviral Drugs for Treating and Preventing HIV Infections in Kenya 2016*. Nairobi: National AIDS and STDs Control Programme Ministry of Health.

National AIDS and STI Control Programme [NASCOP] (2019). *Key Population Mapping and Size Estimation in Selected Counties in Kenya: Phase 1*. Nairobi: National AIDS and STI Control Programme.

National AIDS and STI Control Programme [NASCOP] (2020). *Preliminary KENPHIA 2018 Report, Nairobi*. Nairobi: National AIDS and STI Control Programme.

Nduva, G. M., Hassan, A. S., Nazziwa, J., Graham, S. M., Esbjörnsson, J., and Sanders, E. J. (2020). HIV-1 transmission patterns within and between risk groups in coastal kenya. *Sci. Rep.* 10:6775. doi: 10.1038/s41598-020-63731-z

Nduva, G. M., Nazziwa, J., Hassan, A. S., Sanders, E. J., and Esbjörnsson, J. (2021). The role of phylogenetics in discerning HIV-1 mixing among vulnerable populations and geographic regions in sub-saharan africa: a systematic review. *Viruses* 13:1174. doi: 10.3390/v13061174

Nduva, G. M., Otieno, F., Kimani, J., Wahome, E., McKinnon, L. R., Cholette, F., et al. (in press). Quantifying rates of HIV-1 flow between risk groups and geographic locations in Kenya: a country-wide phylogenetic study. *Virus Evol.*

Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V., and Essex, M. (2014). Impact of sampling density on the extent of HIV clustering. *AIDS Res. Hum. Retroviruses* 30, 1226–1235. doi: 10.1089/aid.2014.0173

Phillips, N. D. (2017). Yarrr! The pirate's guide to R. *APS Observer.* 30:9.

Pineda-Peña, A.-C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., et al. (2013). Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* 19, 337–348. doi: 10.1016/j.meegid.2013.04.032

Poon, A. F., Gustafson, R., Daly, P., Zerr, L., Demlow, S. E., Wong, J., et al. (2016). Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *Lancet HIV* 3, e231–e238. doi: 10.1016/S2352-3018(16)00046-1

Pybus, O. G., Tatem, A. J., and Lemey, P. (2015). Virus evolution and transmission in an ever more connected world. *Proc. R. Soc. B* 282:20142878. doi: 10.1098/rspb.2014.2878

Ratmann, O., Van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S., Wensing, A., et al. (2016). Sources of HIV infection among men having sex with men and implications for prevention. *Sci. Transl. Med.* 8:320ra2. doi: 10.1126/scitranslmed.aad1863

Sallam, M., Esbjörnsson, J., Baldvinsdóttir, G., Indriðason, H., Björnsdóttir, T. B., Widell, A., et al. (2017). Molecular epidemiology of HIV-1 in iceland: early introductions, transmission dynamics and recent outbreaks among injection drug users. *Infect. Genet. Evol.* 49, 157–163. doi: 10.1016/j.meegid.2017.01.004

Sanders, E. J., Graham, S. M., Okuku, H. S., van der Elst, E. M., Muhaari, A., Davies, A., et al. (2007). HIV-1 infection in high risk men who have sex with men in Mombasa, Kenya. *AIDS* 21, 2513–2520. doi: 10.1097/QAD.0b013e3282f2704a

Sanders, E. J., Mugo, P., Prins, H. A., Wahome, E., Thiong'o, A. N., Mwashigadi, G., et al. (2014). Acute HIV-1 infection is as common as malaria in young febrile adults seeking care in coastal Kenya. *AIDS (London, England)* 28:1357. doi: 10.1097/QAD.0000000000000245

Sanders, E. J., Okuku, H. S., Smith, A. D., Mwangome, M., Wahome, E., Fegan, G., et al. (2013). High HIV-1 incidence, correlates of HIV-1 acquisition, and high viral loads following seroconversion among MSM. *AIDS* 27, 437–446. doi: 10.1097/QAD.0b013e32835b0f81

Sanders, E. J., Jaffe, H., Musyoki, H., Muraguri, N., and Graham, S. M. (2015a). Kenyan MSM: no longer a hidden population. *AIDS* 29, S195–S199. doi: 10.1097/QAD.0000000000000928

Sanders, E. J., Wahome, E., Powers, K. A., Werner, L., Fegan, G., Lavreys, L., et al. (2015b). Targeted screening of at-risk adults for acute HIV-1 infection in sub-Saharan Africa. *AIDS (London, England)* 29:S221. doi: 10.1097/QAD.0000000000000924

Schultz, A.-K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., et al. (2009). JPHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res.* 37, W647–W651. doi: 10.1093/nar/gkp371

Skar, H., Axelsson, M., Berggren, I., Thalme, A., Gyllensten, K., Liitsola, K., et al. (2011). Dynamics of two separate but linked HIV-1 CRF01_AE outbreaks among injection drug users in Stockholm, Sweden, and Helsinki, Finland. *J. Virol.* 85, 510–518. doi: 10.1128/JVI.01413-10

Smith, A. D., Fearon, E., Kabuti, R., Irungu, E., Kungu, M., Babu, H., et al. (2021a). Disparities in HIV/STI burden and care coverage among men and transgender persons who have sex with men in Nairobi, Kenya: a cross-sectional study. *BMJ Open* 11:e055783.

Smith, A. D., Kimani, J., Kabuti, R., Weatherburn, P., Fearon, E., and Bourne, A. (2021b). HIV burden and correlates of infection among transfeminine people and cisgender men who have sex with men in Nairobi, Kenya: an observational study. *Lancet HIV* 8, e274–e283. doi: 10.1016/S2352-3018(20)30310-6

Stannah, J., Dale, E., Elmes, J., Staunton, R., Beyrer, C., Mitchell, K., et al. (2019). HIV testing and engagement with the HIV treatment cascade among men who have sex with men in Africa: a systematic review and meta-analysis. *Lancet HIV* 6, e769–e787. doi: 10.1016/S2352-3018(19)30239-5

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

van der Elst, E. M., Gichuru, E., Muraguri, N., Musyoki, H., Micheni, M., Kombo, B., et al. (2015). Strengthening healthcare providers' skills to improve HIV services for MSM in Kenya. *AIDS (London, England)* 29:S237. doi: 10.1097/QAD.0000000000000882

van der Elst, E. M., Gichuru, E., Omar, A., Kanungi, J., Duby, Z., Midoun, M., et al. (2013). Experiences of Kenyan healthcare workers providing services to men who have sex with men: qualitative findings from a sensitivity training programme. *J. Int. AIDS Soc.* 16:18741. doi: 10.7448/IAS.16.4.18741

van der Elst, E. M., Mudza, R., Onguso, J. M., Kiirika, L., Kombo, B., Jahangir, N., et al. (2020). A more responsive, multi-pronged strategy is needed to strengthen HIV healthcare for men who have sex with men in a decentralized health system: qualitative insights of a case study in the Kenyan coast. *J. Int. AIDS Soc.* 23:e25597. doi: 10.1002/jia2.25597

Volz, E. M., Ionides, E., Romero-Severson, E. O., Brandt, M. G., Mokotoff, E., and Koopman, J. S. (2013). HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Med.* 10:e1001568; discussione1001568. doi: 10.1371/journal.pmed.1001568

Wahome, E., Thiong'o, A. N., Mwashigadi, G., Chirro, O., Mohamed, K., Gichuru, E., et al. (2018). An empiric risk score to guide PrEP targeting among MSM in coastal Kenya. *AIDS Behav.* 22, 35–44. doi: 10.1007/s10461-018-2141-2

Wickham, H. (2016). *ggplot2: Elegant Graphics For Data Analysis*. Berlin: Springer.

**Author Disclaimer:** The contents of this manuscript are the responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government, AAS, NEPAD Agency, Wellcome Trust, IAVI, Swedish Research Council, or the United Kingdom Government.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 31 March 2022
doi: 10.3389/fmicb.2022.782609

# Transmission Clusters, Predominantly Associated With Men Who Have Sex With Men, Play a Main Role in the Propagation of HIV-1 in Northern Spain (2013–2018)

*Horacio Gil[1*†], Elena Delgado[1†], Sonia Benito[1], Leonidas Georgalis[1], Vanessa Montero[1], Mónica Sánchez[1], Javier E. Cañada-García[1], Elena García-Bodas[1], Asunción Díaz[2], Michael M. Thomson[1*] and The Members of the Spanish Group for the Study of New HIV Diagnoses*

*[1] HIV Biology and Variability Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain, [2] HIV Surveillance and Behavioral Monitoring Unit, Centro Nacional de Epidemiología, Instituto de Salud Carlos III, Madrid, Spain*

Viruses of HIV-1-infected individuals whose transmission is related group phylogenetically in transmission clusters (TCs). The study of the phylogenetic relations of these viruses and the factors associated with these individuals is essential to analyze the HIV-1 epidemic. In this study, we examine the role of TCs in the epidemiology of HIV-1 infection in Galicia and the Basque County, two regions of northern Spain. A total of 1,158 HIV-1-infected patients from both regions with new diagnoses (NDs) in 2013–2018 were included in the study. Partial HIV-1 *pol* sequences were analyzed phylogenetically by approximately maximum-likelihood with FastTree 2. In this analysis, 10,687 additional sequences from samples from HIV-1-infected individuals collected in Spain in 1999–2019 were also included to assign TC membership and to determine TCs' sizes. TCs were defined as those which included viruses from ≥4 individuals, at least 50% of them Spaniards, and with ≥0.95 Shimodaira-Hasegawa-like node support in the phylogenetic tree. Factors associated to TCs were evaluated using odds ratios (OR) and their 95% CI. Fifty-one percent of NDs grouped in 162 TCs. Male patients (OR: 2.6; 95% CI: 1.5–4.7) and men having sex with men (MSM; OR: 2.1; 95% CI: 1.4–3.2) had higher odds of belonging to a TC compared to female and heterosexual patients, respectively. Individuals from Latin America (OR: 0.3; 95% CI: 0.2–0.4), North Africa (OR: 0.4; 95% CI: 0.2–1.0), and especially Sub-Saharan Africa (OR: 0.02; 95% CI: 0.003–0.2) were inversely associated to belonging to TCs compared to native Spaniards. Our results show that TCs are important components of the HIV-1 epidemics in the two Spanish regions studied, where transmission between MSM is predominant. The majority of migrants were infected with viruses not belonging to TCs that expand in Spain. Molecular epidemiology is essential to identify local peculiarities of HIV-1 propagation. The early detection of TCs and prevention of their expansion, implementing effective control measures, could reduce HIV-1 infections.

**Keywords: Spain, HIV-1, transmission clusters, molecular epidemiology, men who have sex with men, migrants**

OPEN ACCESS

***Edited by:***
*Xiaoxu Han,*
*The First Affiliated Hospital of China*
*Medical University, China*

***Reviewed by:***
*Frank Maldarelli,*
*National Cancer Institute (NIH),*
*United States*
*Carmen Maria Gonzalez Domenech,*
*University of Malaga, Spain*
*Vlad Novitsky,*
*HSPH, United States*
*Juan Angel Patino Galindo,*
*Columbia University, United States*

***Correspondence:**
*Michael M. Thomson*
*mthomson@isciii.es*
*Horacio Gil*
*hgil@isciii.es*

*[†] These authors have contributed equally to this work and share first authorship*

***Specialty section:***
*This article was submitted to*
*Virology,*
*a section of the journal*
*Frontiers in Microbiology*

***Received:** 24 September 2021*
***Accepted:** 14 March 2022*
***Published:** 31 March 2022*

***Citation:***
*Gil H, Delgado E, Benito S, Georgalis L, Montero V, Sánchez M, Cañada-García JE, García-Bodas E, Díaz A, Thomson MM and The Members of the Spanish Group for the Study of New HIV Diagnoses (2022) Transmission Clusters, Predominantly Associated With Men Who Have Sex With Men, Play a Main Role in the Propagation of HIV-1 in Northern Spain (2013–2018). Front. Microbiol. 13:782609. doi: 10.3389/fmicb.2022.782609*

## INTRODUCTION

The HIV-1 epidemic is still one of the major public health problems in Spain. Around 3,500–4,000 new diagnoses (NDs) of HIV-1 infection are reported every year, with an estimated incidence of NDs of 7.5 per 100,000 population in 2019 (HIV STI and Hepatitis Surveillance Unit, 2020). A decreasing trend in the incidence of NDs has been observed since 2010, although it is still higher than the average rate found in the EU/EEA (European Centre for Disease Prevention and Control and WHO Regional Office for Europe, 2020). In 2019, most (56%) of the NDs were diagnosed among men who had sex with men (MSM) and 36% of all reported NDs were in people born outside of Spain (HIV STI and Hepatitis Surveillance Unit, 2020).

Molecular epidemiology is an important tool for describing the HIV-1 epidemic (Brenner et al., 2013; Paraskevis et al., 2016; Wertheim et al., 2017; Board et al., 2020). Individuals whose transmission is related group phylogenetically in clades named transmission clusters (TCs). Due to the high genetic variability of HIV-1, phylogenetic analysis allows to reconstruct transmission events through the identified TCs and infer the history of the HIV-1 epidemic (Hué et al., 2004, 2005).

National databases of protease and reverse transcriptase (Pr-RT) sequences, primarily obtained for antiretroviral drug resistance testing, contain valuable data about HIV-1 expansion and have been used in molecular epidemiology studies (Ragonnet-Cronin et al., 2016a; Parczewski et al., 2017; Oster et al., 2018; Petersen et al., 2018; Verhofstede et al., 2018; Pineda-Peña et al., 2019; Fabeni et al., 2021). The phylogenetic analyses performed in these studies combined with clinical and epidemiological data of the patients provide relevant public health information for the implementation of control measures and for monitoring the HIV-1 epidemic (Brenner et al., 2013; Vasylyeva et al., 2016, 2020; Paraskevis et al., 2019; Board et al., 2020; Campbell et al., 2020).

Spanish clinical guidelines recommend to perform a genotypic drug resistance test before starting antiretroviral therapy (ART) in all HIV-1-diagnosed patients and in ART-failing cases [AIDS study group (GESIDA) of the Spanish Society of Infectious Diseases and Clinical Microbiology and the National AIDS Plan, 2020]. Pr-RT sequences obtained for these tests have been analyzed in different molecular epidemiology studies, describing the genetic features of the HIV-1 epidemic in different regions of Spain (Thomson et al., 2001; Holguín et al., 2007; Cuevas et al., 2009a,b; González-Alba et al., 2011; Yebra et al., 2013; Pérez-Parra et al., 2015, 2016; Patiño-Galindo et al., 2016, 2017b; González-Domenech et al., 2020). The phylogenetic studies have also allowed the identification of large TCs among MSM which are actively growing in Spain (Delgado et al., 2015, 2019; Patiño-Galindo et al., 2017a; González-Domenech et al., 2018), as well as transmitted drug resistance mutations which are spreading in TCs (Cuevas et al., 2009b; Vega et al., 2015; Viciana et al., 2016; González-Domenech et al., 2018, 2020), highlighting the value of molecular epidemiology as a tool for HIV surveillance.

The Basque Country and Galicia are two regions located in northern Spain, comparable in terms of population and HIV-1 diagnosis rates. During 2013–2018, the mean resident population and rates of new HIV diagnoses were estimated in 2,2 million and 6,9 per 100,000 inhabitants for the Basque Country and 2,7 million and 5,6 per 100,000 inhabitants for Galicia (HIV STI and Hepatitis Surveillance Unit, 2020; Instituto Nacional de Estadistica, 2021). In this study, we analyzed the frequency of grouping in TCs among patients diagnosed in 2013–2018 in these northern Spanish regions, identifying the different features associated with TCs in the HIV-1 epidemics in these regions.

## MATERIALS AND METHODS

### Patients

Samples from individuals newly diagnosed of HIV-1 infection in the northern Spanish regions of Galicia and Basque Country during 2013–2018, sent to the HIV Biology and Variability Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, were included in the study.

The representativeness of the NDs of our cohort was estimated comparing them to the NDs reported to the Spanish information system on new HIV diagnoses (SINIVIH by its Spanish acronym; HIV STI and Hepatitis Surveillance Unit, 2020), from the same period and studied regions. As patients in both databases cannot be linked, gender, age group, transmission route and region of birth frequencies were compared between both groups of individuals using the Chi-squared test.

The use of anonymized, de-identified clinical/demographic and sequence data was reviewed and approved under an exempt protocol by the Bioethics and Animal Well-being Committee of Instituto de Salud Carlos III, with report numbers CEI PI 38_2016-v3 (dated 20 June 2016) and CEI PI 31_2019-v5 (dated 6 November 2019). This study did not require written informed consent by study participants, except for those participants who required additional samples different from the ones obtained in the routine clinical practice.

### Nucleic Acid Extraction, Amplification, and Sequencing

Nucleic acid was extracted from plasma or whole blood samples. RNA was extracted from 1 ml plasma using NUCLISENS® easyMAG® (BioMérieux, Marcy l'Etoile, France) and DNA was extracted from 200 µl whole blood using QIAamp® DNA DSP blood mini kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. A Pr-RT fragment of *pol* (HXB2 positions 2253–3629) was amplified by RT-PCR followed by nested PCR from RNA or by nested PCR from DNA. Reagents, PCR thermal profiles, and primers are described in **Supplementary Tables 1, 2**.

Population sequencing was performed with ABI Prism BigDye Terminator Cycle Sequencing kit and ABI 3730 XL sequencer (Applied Biosystems, Foster City, CA, United States) in the Genomic Unit of Instituto de Salud Carlos III. Sequences were assembled with SeqMan Pro v.12.2.1 (DNA STAR Lasergene, Madison, WI, United States) and edited with BioEdit v.7.2.5 (Hall, 1999).

## Phylogenetic Analyses and Transmission Cluster Identification

Partial *pol* sequences from the patients included in this study and sequences obtained by us from another 10,687 HIV-1-infected individuals attended in clinical centers from Spain whose samples were collected in the period 1999–2019 were included in the analyses. A single sequence per patient was used in the phylogenetic analysis. In cases where several sequences were available for one patient, that closer to the date of HIV-1 diagnosis was selected.

Reference sequences from the different subtypes and CRFs retrieved from the Los Alamos HIV Sequence Database[1] were included for the phylogenetic and bootscanning analysis. In addition, we conducted BLAST (Altschul et al., 1990) searches of each of the sequences obtained by us against the Los Alamos database, including in the phylogenetic analysis up to 10 most similar sequences from each search.

Sequences were analyzed phylogenetically by an approximately maximum-likelihood method using FastTree2 (Price et al., 2010), running it in a local desktop computer. In these analyses, the general time reversible model of nucleotide substitution with CAT approximation to account for among-site heterogeneity in substitution rates was used, and the reliability of nodes was assessed with Shimodaira–Hasegawa (SH)-like local support values (Shimodaira and Hasegawa, 1999; Guindon et al., 2010). Classification of sequences in subtypes and recombinant forms was based on clustering with clade references in approximately maximum-likelihood trees. Sequences suspected of intersubtype recombination were subsequently analyzed by bootscanning with SimPlot v3.5 (Lole et al., 1999).

Transmission clusters were defined as those comprising viruses from four or more individuals, at least 50% of them Spanish, and whose sequences grouped in the phylogenetic tree with a SH-like node support value $\geq 0.95$. This limit in the percentage of migrants included in TCs was used to minimize the effect of viruses which are circulating in the countries of origin of the migrants, grouping phylogenetically in TCs in our sequence dataset due to the relatedness of the viruses of those countries, although they are not actually spreading in Spain.

## Statistical Analysis

To evaluate the association between variables, Chi-squared and Fisher's exact tests were used, with associations being considered statistically significant at a value of $p < 0.05$. A multivariable logistic regression model was performed to identify factors associated to TCs. The model was adjusted by gender, age group, transmission route, country of origin of the patient, Spanish region of sample collection and HIV-1 genetic form. Associations were measured using the odds ratio (OR) and its 95% CI. Data analyses were performed using the STATA statistical software package Version 16 (Stata Corporation, College Station, TX, United States).

## Sequence Accession Numbers

To avoid the identification of transmission networks and the potential breech of patient confidentiality, only 19% of the sequences have been deposited in GenBank, including at least two sequences from each TC comprising five or more patients with new HIV-1 diagnoses in the period of study. The sequences are under the following accession numbers: KT276255, KT276260, KT276263-KT276264, KT276266-KT276267, KT276270-KT276271, KU685562-KU685565, KU685567, KU685569-KU685575, KU685577, KU685581-KU685582, KU685586, KU685588-KU685590, KU685592, KX534325, KX534329, KX534331, KY465968, KY496624, KY514084, KY989950, MF999250-MF999256, MF999258-MF999261, MK177721-MK177733, MK177735-MK177752, MK177754-MK177757, MK177761-MK177772, MK177783-MK17785, MK177787-MK177790, MK177792, MK177795-MK177796, MK177798-MK177799, MK177801-MK177803, MK177805-MK177807, MK298150, MT436242, MT436244, MT436246-MT436247, MT436249-MT436250, MW344920-MW344921, MW584217-MW584224, OK011532, OK011542, OK011545-OK011546, OK011549, OK148912, OK148914, OK148917-OK148919, OK148921, OK148941-OK148942, OK148953-OK148957, OK148961, OK148965, OK148968-OK148972, OK148974, OL982314-OL982315, and OM914651-OM914711.

# RESULTS

## Study Population

A total of 1,158 HIV-1 NDs from Galicia and Basque Country diagnosed in 2013–2018 were included in the study (**Table 1**). The majority of the patients were male (82%), with MSM being the most common transmission route (46%). Heterosexuals represented 34% and 15% of the individuals were male whose transmission route was reported as sexual without specifying whether they were in the MSM or heterosexual category. Subtype B (63%) was the most frequent genetic form among the individuals included in the study, followed by unique recombinant forms (URFs; 8.2%), CRF02_AG (7.7%), and subtype F (7.6%; **Table 1**).

Seventy percent of the patients were native Spanish (**Table 1**), followed by Latin Americans (17%) and Sub-Saharan Africans (8.0%). There were important differences in gender and transmission route proportions between these two migrant populations, with female [52% (44/85) vs. 26% (37/180)] and heterosexual transmission [88% (65/74) vs. 33% (54/165)] being higher in Sub-Saharan Africans than in Latin Americans.

The distribution of patients by gender and transmission route was similar in both studied regions (**Table 1**), although the frequency of people who inject drugs (PWID) was slightly higher in Galicia (5.3 vs. 3.2%), but the difference was not statistically significant ($p = 0.13$). The percentage of patients of non-Spanish origin was 35% in the Basque Country (13% corresponding to Africans), a proportion which is statistically higher ($p < 0.001$) compared to the 17% found in Galicia (2.2% Africans). Statistical differences in the distribution of genetic forms were also observed between both regions ($p < 0.001$), with the frequencies of subtype F (13%) and CRF02_AG (9.9%) being higher in Galicia and the Basque Country, respectively (**Table 1**).

[1] https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html

**TABLE 1** | Characteristics of the HIV-1 newly diagnosed individuals included in the study.

| | | BC* | GA† | All | |
|---|---|---|---|---|---|
| | | *N* (%) | *N* (%) | *N* (%) | *p*-value |
| Gender | Male | 626 (82) | 306 (83) | 932 (82) | 0.71 |
| | Female | 133 (17) | 65 (17) | 198 (17) | |
| | Transexual | 5 (0.65) | 0 (0) | 5 (0.44) | |
| | Unknown | 12 | 11 | 23 | |
| Transmission route | MSM‡ | 317 (47) | 142 (44) | 459 (46) | 0.61 |
| | Heterosexual | 233 (34) | 112 (35) | 345 (34) | |
| | MNSST§ | 103 (15) | 48 (15) | 151 (15) | |
| | PWID¶ | 22 (3.2) | 17 (5.3) | 39 (3.9) | |
| | Other | 6 (0.88) | 3 (0.93) | 9 (0.90) | |
| | Unknown | 95 | 60 | 155 | |
| Region of origin | Spain | 477 (65) | 268 (83) | 745 (70) | <0.001 |
| | Latin America | 147 (20) | 37 (11) | 184 (17) | |
| | Sub-Saharan Africa | 79 (11) | 6 (1.9) | 85 (8.0) | |
| | North Africa | 14 (1.9) | 1 (0.3) | 15 (1.4) | |
| | Europe‖ | 16 (2.2) | 12 (3.7) | 28 (2.6) | |
| | Other | 5 (0.68) | 1 (0.31) | 6 (0.56) | |
| | Unknown | 38 | 57 | 95 | |
| Genetic form | B | 489 (63) | 235 (62) | 724 (63) | <0.001 |
| | F | 39 (5.0) | 49 (13) | 88 (7.6) | |
| | A | 21 (2.7) | 12 (3.1) | 33 (2.9) | |
| | C | 19 (2.5) | 20 (5.2) | 39 (3.4) | |
| | G | 12 (1.6) | 8 (2.1) | 20 (1.7) | |
| | CRF02_AG | 77 (9.9) | 12 (3.1) | 89 (7.7) | |
| | CRF_BF | 20 (2.6) | 8 (2.1) | 28 (2.4) | |
| | URF | 65 (8.4) | 30 (7.9) | 95 (8.2) | |
| | Other | 34 (4.4) | 8 (2.1) | 42 (3.6) | |
| Total | | 776 | 382 | 1158 | |

**BC, Basque Country.*

*†GA, Galicia.*

*‡MSM, men who have sex with men.*

*§MNSST, men who have non-specified sexual transmission.*

*¶PWID, person who injects drugs.*

*‖Other than Spain.*

## Representativeness of the Patients Included in the Study

Our cohort of newly diagnosed HIV-1-infected patients represents 64% of the notified cases in 2013–2018 (**Supplementary Table 3**), assuming that all NDs included in our study have been reported to the SINIVIH. This percentage was higher in the Basque Country, where it reached 86%, than in Galicia, where it was 42% of the notified cases. No statistical differences were found between both patient groups regarding the distribution of individuals by gender, transmission route, region of origin, or age group (**Supplementary Table 3**). A map with the studied regions, with the numbers of sequences included in our dataset and the representativeness in each region of the NDs analyzed in this study is shown in **Supplementary Figure 1**.

## TCs Identified in the Study

Fifty-one percent (594/1158) of the individuals belonged to 162 different TCs. Regarding regions, 45% (351/776) of the patients were included in 98 TCs in the Basque Country and 64% (243/382) belonged to 82 TCs in Galicia (**Table 2**). Eighteen (11%) TCs comprised patients from both regions. The sizes of the TCs ranged from 4 to 205 patients. MSM was the main transmission route associated with 59% of the TCs (**Table 3**). Transmission routes other than MSM were more frequently associated with TCs in Galicia (44%) than in the Basque County (35%; **Table 2**), although the difference was not statistically significant ($p = 0.207$).

Non-subtype B genetic forms were found in 33 (21%) of the 162 TCs identified in the study. Sixteen of them derived from genetic forms circulating in Latin America (BF1 recombinants, Brazilian subtype C and F1 variants, CRF19_cpx and CRF20_BG) and 10 from genetic forms circulating in Africa (CRF02_AG, CRF02/A3, CRF02/G, CRF06_cpx and subtype C strains of African origin). Twenty-nine percent (138/474) of the Spaniards belonging to TCs were members of non-subtype B TCs.

A total of 15 large TCs (here defined as those with ≥30 individuals) were identified in the study (**Table 3**), 10 of them of subtype B. MSM was the associated transmission route in 14 (93%) of them. Seventy-six percent (67/88) of subtype F infections belonged to the MSM-associated TCs F1_1 and F1_3. Most of these large TCs have spread mainly in a specific region. Thus, TCs B07, B08, B09, B10, B12, B70, and F1_3 were found mainly in the Basque Country, while TCs F1_1, B13, BG_2, and B05 were spread mainly in Galicia (**Table 3**). In addition, A1_1, CRF02_1, B50 and B31 have a global

**TABLE 2** | Characteristic of the identified transmission clusters (TCs) of the study.

| | | Basque Country | | Galicia | | Both regions | |
|---|---|---|---|---|---|---|---|
| | | *N* | % | *N* | % | *N* | % |
| Transmission route* | MSM† | 64 | 65 | 46 | 56 | 95 | 59 |
| | Heterosexual | 10 | 10 | 14 | 17 | 24 | 15 |
| | Sexual‡ | 16 | 16 | 10 | 12 | 23 | 14 |
| | PWID§ | 8 | 8.2 | 12 | 15 | 20 | 12 |
| Total | | 98 | 100 | 82 | 100 | 162 | 100 |

*Main mode of HIV transmission among individuals in the TC.

†MSM, men who have sex with men.

‡There is not predominance of MSM or heterosexual transmission.

§PWID, persons who inject drugs.

expansion in Spain, with a high number of patients outside of both studied regions. Four TCs had greater than 50% increase in NDs during the period 2013–2018: F1_3 (94%), A1_1 (76%), B70 (72%), and CRF02_1 (53%; **Table 3**).

A phylogenetic tree which includes all the TCs comprising five or more patients newly diagnosed in the studied period in Galicia and Basque Country is shown in **Supplementary Figure 2**.

## Factors Associated With TCs

The percentage of patients in TCs was higher in males (57%) than in females (22%), and in MSM (67%) than in heterosexuals (31%). This percentage was also higher in Spaniards (64%) than in Latin Americans (35%) or Sub-Saharan Africans (1.2%), and in individuals infected with viruses of subtypes F (76%) or B (58%) than in those infected with CRF02_AG viruses (17%). The different distribution of TCs according to gender, mode of HIV-1 transmission, country of origin of the patient, and HIV-1 genetic form were statistically significant (**Table 4**). The age group was also associated to TCs ($p=0.036$). The age range of 20–29 years showed an increased percentage of patients in TCs (59%), which was especially high in Galicia (81%, $p=0.003$) but was not statistical significant in the Basque Country, and no statistical differences were found in the distribution of main variables.

Along the studied period, there was no clear trend in the proportion of patients associated with TCs (**Supplementary Figure 3**). In Galicia, a decrease of notified HIV-1 NDs (**Supplementary Figure 3**, panel A) with a slight reduction of the number of TCs and an increase in the ratio of patients per TC (**Supplementary Figure 3**, panel D) was observed during the study period.

In the multivariate analysis, males (OR: 2.6; 95% CI: 1.5–4.7) and MSM (OR: 2.1; 95% CI: 1.4–3.2) had higher odds of belonging to a TC compared to females and heterosexuals, respectively (**Table 5**). Patients from Galicia (OR: 1.5; 95% CI: 1.1–2.2) and subtype F infections (OR: 2.5; 95% CI: 1.3–5.0) were also factors associated with TCs (**Table 5**).

Individuals from Latin America (OR: 0.3; 95% CI: 0.2–0.4), North Africa (OR: 0.4; 95% CI: 0.2–1.0) and especially from Sub-Saharan Africa (OR: 0.02; 95% CI: 0.003–0.2) were inversely associated to belonging to TCs compared to Spanish patients (**Table 5**). Similarly, patients over 29 years old (OR: 0.6; 95% CI: 0.4–0.9) and infections with a virus of a CRF_BF (OR: 0.1; 95% CI: 0.03–0.4) or of a genetic form classified in the "other" category (OR: 0.3; 95% CI: 0.2–0.8) had lower odds of belonging to TCs than patients in the age range of 20–29 years and with subtype B infections, respectively (**Table 5**). These associations, although with slightly lower OR, where also observed when the criterion on the country of origin of the patients was excluded from the TC definition (**Supplementary Table 4**).

# DISCUSSION

The phylogenetic analysis of sequences from almost 12,000 HIV-1-infected patients from our database has allowed us to investigate the transmission events occurring during the 2013–2018 period in two regions of northern Spain, Galicia and Basque Country, identifying TCs and their associated factors. Our ND cohort had a good representativeness, corresponding to 64% of the NDs reported to the SINIVIH from these regions during that period, especially in the Basque Country where it reached 86%.

Our analysis has assigned 51% of the NDs from Galicia and Basque Country to 162 different TCs, indicating that TCs are playing an important role in the spread of HIV-1 infections in both regions. There is no consensus on the criteria for defining a TC, as these should depend on the aim of the study (Hassan et al., 2017). We considered only TCs including at least 50% Spaniards and at least four documented infections, because they may be more epidemiologically relevant to describe HIV-1 strains spreading in an area. Smaller clusters comprising two ("pairs") or three ("triplets") individuals are much more numerous than larger ones (Hughes et al., 2009; Brown et al., 2011; Yebra et al., 2013) and in some studies are distinguished from larger clusters (Hoenigl et al., 2016; Fabeni et al., 2020). They could represent sexual couples with or without a second relationship by one of the members of the couple, and their epidemiologic relevance is uncertain. The inclusion of these requirements in the TC definition has allowed us to better identify the structure and factors associated to HIV-1 strains whose transmission is established in Spain.

Transmission clusters were found associated with male gender and MSM. In fact, almost all the largest TCs (≥30 patients) identified in our study were associated with MSM. HIV-1 molecular epidemiological studies in different countries, independently of the methodology used, have found a similar

**TABLE 3** | Characteristics of the transmission clusters comprising ≥30 patients in the Basque Country and Galicia (2013–2018).

| Cluster | Genetic form | No. patients | Regional distribution | | | | | | % Spanish | % Male | % Transmission route[‡] (MSM/HT/MNSST/PWID) | Patients (2013–18)[§] | | Period[¶] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | BC* | | GA[†] | | Others | | | | | | | |
| | | | *N* | % | *N* | % | *N* | % | | | | *N* | % | |
| F1_1 | F1 | 205 | 9 | 4.4 | 138 | 67 | 58 | 28 | 85 | 97 | 72/13/14/0 | 78 | 38 | 2000–2019 |
| CRF02_1 | CRF02_AG | 83 | 34 | 41 | 3 | 3.6 | 46 | 55 | 79 | 92 | 64/22/12/2 | 44 | 53 | 2008–2019 |
| B50 | B | 74 | 15 | 20 | 3 | 4.1 | 56 | 76 | 85 | 100 | 69/9/22/0 | 34 | 46 | 2007–2019 |
| A1_1 | A1 | 63 | 8 | 13 | 24 | 38 | 31 | 49 | 71 | 95 | 77/12/11/0 | 48 | 76 | 2006–2019 |
| B70 | B | 61 | 56 | 92 | | | 5 | 8.2 | 88 | 98 | 64/9/27/0 | 44 | 72 | 2008–2019 |
| B08 | B | 60 | 46 | 77 | 6 | 10 | 8 | 13 | 92 | 98 | 74/6/19/2 | 13 | 22 | 2006–2019 |
| B09 | B | 57 | 55 | 97 | | | 2 | 3.5 | 88 | 98 | 62/15/21/2 | 10 | 18 | 1991–2019 |
| B13 | B | 55 | | | 54 | 98 | 1 | 1.8 | 89 | 94 | 62/19/19/0 | 16 | 29 | 2004–2017 |
| B31 | B | 49 | 19 | 39 | 2 | 4.1 | 28 | 57 | 77 | 98 | 73/7/20/0 | 23 | 47 | 2001–2019 |
| B07 | B | 38 | 36 | 95 | | | 2 | 5.3 | 91 | 95 | 56/12/29/3 | 3 | 7.9 | 2000–2019 |
| B10 | B | 36 | 29 | 81 | | | 7 | 19 | 90 | 97 | 67/12/18/3 | 7 | 19 | 1991–2019 |
| F1_3 | F1 | 36 | 29 | 81 | | | 7 | 19 | 76 | 100 | 61/11/29/0 | 34 | 94 | 2014–2018 |
| B12 | B | 35 | 25 | 71 | 1 | 2.9 | 9 | 26 | 63 | 100 | 67/7/26/0 | 11 | 31 | 2009–2019 |
| BG_2 | URF_BG | 33 | | | 33 | 100 | | | 85 | 80 | 0/45/3/52 | 2 | 6.1 | 1994–2019 |
| B05 | B | 31 | | | 31 | 100 | | | 93 | 94 | 56/26/19/0 | 10 | 32 | 2003–2019 |

**BC, Basque Country.*

*[†]GA, Galicia.*

*[‡]Percentages were calculated from the total of patients with known transmission route. MSM, men who have sex with men; HT, heterosexual; MNSST, men who have non-specified sexual transmission; and PWID, people who inject drugs.*

*[§]Newly-diagnosed HIV-1 patients during 2013–2018.*

*[¶]Year of HIV-1 diagnosis of the patients in the TC.*

**TABLE 4 |** Characteristics of the new HIV diagnoses included in TCs (2013–2018).

| | | Basque Country | | | | | Galicia | | | | | Both regions | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Categories** | **No** | | **Yes** | | ***P*-value** | **No** | | **Yes** | | ***P*-value** | **No** | | **Yes** | | ***P*-value** |
| | | ***N***[*] | **%** | ***N*** | **%** | | ***N*** | **%** | ***N*** | **%** | | ***N*** | **%** | ***N*** | **%** | |
| Gender (*n* = 1135, no data = 23) | Female | 115 | 86 | 18 | 14 | <0.001 | 40 | 62 | 25 | 38 | <0.001 | 155 | 78 | 43 | 22 | <0.001 |
| | Male | 303 | 48 | 323 | 52 | | 95 | 31 | 211 | 69 | | 398 | 43 | 534 | 57 | |
| | Transexual | 3 | 60 | 2 | 40 | | 0 | 0 | 0 | 0 | | 3 | 60 | 2 | 40 | |
| Transmission route (*n* = 1003, no data = 155) | Heterosexual | 178 | 76 | 55 | 24 | <0.001 | 60 | 54 | 52 | 46 | <0.001 | 238 | 69 | 107 | 31 | <0.001 |
| | MSM[†] | 120 | 38 | 197 | 62 | | 33 | 23 | 109 | 77 | | 153 | 33 | 306 | 67 | |
| | MNSST[‡] | 48 | 47 | 55 | 53 | | 18 | 37 | 30 | 63 | | 66 | 43 | 85 | 56 | |
| | PWID[§] | 15 | 68 | 7 | 32 | | 8 | 47 | 9 | 53 | | 23 | 59 | 16 | 41 | |
| | Other | 4 | 67 | 2 | 33 | | 3 | 100 | 0 | 0 | | 7 | 78 | 2 | 22 | |
| Region of origin (*n* = 1063, no data = 95) | Spain | 189 | 40 | 288 | 60 | <0.001 | 82 | 31 | 186 | 69 | 0.001 | 271 | 37 | 474 | 64 | <0.001 |
| | Latin America | 102 | 69 | 45 | 31 | | 18 | 49 | 19 | 51 | | 120 | 65 | 64 | 35 | |
| | Sub-Saharan Africa | 79 | 100 | 0 | 0 | | 5 | 83 | 1 | 17 | | 84 | 99 | 1 | 1.2 | |
| | North Africa | 9 | 64 | 5 | 36 | | 1 | 100 | 0 | 0 | | 10 | 67 | 5 | 33 | |
| | Europe[¶] | 11 | 69 | 5 | 31 | | 7 | 58 | 5 | 42 | | 18 | 64 | 10 | 36 | |
| | Other | 4 | 80 | 1 | 20 | | 1 | 100 | 0 | 0 | | 5 | 83 | 1 | 17 | |
| Age group (*n* = 1133, no data = 25) | <20 | 10 | 53 | 9 | 47 | 0.44 | 3 | 50 | 3 | 50 | 0.003 | 13 | 52 | 12 | 48 | 0.036 |
| | 20–29 | 92 | 51 | 90 | 49 | | 14 | 19 | 60 | 81 | | 106 | 41 | 150 | 59 | |
| | 30–39 | 160 | 58 | 115 | 41 | | 49 | 41 | 72 | 59 | | 209 | 53 | 187 | 47 | |
| | ≥40 | 161 | 55 | 137 | 45 | | 68 | 43 | 90 | 57 | | 229 | 50 | 227 | 50 | |
| Genetic form (*n* = 1158, no data = 0) | A | 14 | 67 | 7 | 33 | <0.001 | 3 | 25 | 9 | 75 | 0.002 | 17 | 52 | 16 | 48 | <0.001 |
| | B | 224 | 46 | 265 | 54 | | 83 | 35 | 152 | 65 | | 307 | 42 | 417 | 58 | |
| | C | 11 | 58 | 8 | 42 | | 8 | 40 | 12 | 60 | | 19 | 49 | 20 | 51 | |
| | F | 10 | 26 | 29 | 74 | | 11 | 23 | 38 | 78 | | 21 | 24 | 67 | 76 | |
| | G | 10 | 83 | 2 | 17 | | 5 | 63 | 3 | 37 | | 15 | 75 | 5 | 25 | |
| | CRF02_AG | 63 | 82 | 14 | 18 | | 11 | 92 | 1 | 8.3 | | 74 | 83 | 15 | 17 | |
| | CRF_BF | 18 | 90 | 2 | 10 | | 4 | 50 | 4 | 50 | | 22 | 79 | 6 | 21 | |
| | URF | 47 | 72 | 18 | 28 | | 10 | 33 | 20 | 67 | | 57 | 60 | 38 | 40 | |
| | Other | 28 | 82 | 6 | 18 | | 4 | 50 | 4 | 50 | | 32 | 76 | 10 | 24 | |

[*] *N, number of patients.*

[†] *MSM, men who have sex with men.*

[‡] *MNSST, men who have non-specified sexual transmission.*

[§] *PWID, people who inject drugs.*

[¶] *Spain not included.*

**TABLE 5** | Factors associated with transmission clusters, univariate/multivariate analysis.

| | | **Univariate analysis** | | **Multivariate analysis** | |
|---|---|---|---|---|---|
| **Variables** | | **OR[*]** | **95% CI[†]** | **Adjusted OR** | **95% CI** |
| Gender | Female | Reference | | Reference | |
| | Male | 4.8 | 3.3–7.1 | 2.6 | 1.5–4.7 |
| | Transexual | 2.4 | 0.4–15 | NA[‡] | NA |
| Transmission route | Heterosexual | Reference | | Reference | |
| | MSM[§] | 4.5 | 3.2-6.1 | 2.1 | 1.4–3.2 |
| | MNNSST[¶] | 2.9 | 1.9-4.3 | 1.4 | 0.8–2.3 |
| | PWID[‖] | 1.6 | 0.8-3.1 | NA | NA |
| | Other | 0.6 | 0.1–3.1 | NA | NA |
| Region of origin | Spain | Reference | | Reference | |
| | Latin America | 0.3 | 0.2–0.4 | 0.3 | 0.2–0.4 |
| | Sub-Saharan Africa | <0.01 | <0.001–0.06 | 0.02 | 0.003–0.2 |
| | North Africa | 0.3 | 0.1–0.9 | 0.4 | 0.2–1.0 |
| | Europe[ℓ] | 0.3 | 0.1-0.7 | 0.3 | 0.03–3.5 |
| | Other | 0.1 | 0.1–1.0 | 0.4 | 0.1–1.6 |
| Age group | <20 | 0.7 | 0.3–1.5 | NA | NA |
| | 20–29 | Reference | | Reference | |
| | 30–39 | 0.6 | 0.5–0.9 | 0.6 | 0.4–0.9 |
| | ≥40 | 0.7 | 0.5–1.0 | 0.6 | 0.4–0.9 |
| Spanish region | Basque Country | Reference | | Reference | |
| | Galicia | 2.1 | 1.6–2.7 | 1.5 | 1.1–2.2 |
| Genetic form | A | 0.7 | 0.3–1.4 | NA | NA |
| | B | Reference | | Reference | |
| | C | 0.8 | 0.4–1.5 | NA | NA |
| | F | 2.4 | 1.4–3.9 | 2.5 | 1.3–5.0 |
| | G | 0.3 | 0.1–0.7 | 0.4 | 0.1–1.5 |
| | CRF02_AG | 0.2 | 0.1–0.3 | 0.6 | 0.3–1.2 |
| | CRF_BF | 0.2 | 0.1–0.5 | 0.1 | 0.03–0.4 |
| | URF | 0.5 | 0.3–0.8 | 0.8 | 0.4–1.4 |
| | Other | 0.4 | 0.3–0.7 | 0.3 | 0.2–0.8 |

[*] *OR, odds ratio.*

[†] *95% CI, 95% confidence interval.*

[‡] *NA, not applicable. Adjusted OR were not calculated for categories with p<0.1 in the univariate analysis.*

[§] *MSM, men having sex with men.*

[¶] *MNSST, men who have non-specified sexual transmission.*

[‖] *PWID, people who inject drugs.*

[ℓ] *Spain not included.*

association (Parczewski et al., 2017; Wertheim et al., 2017; Ramirez et al., 2021). Also, studies in other Spanish regions have found this association of TCs with MSM (Pérez-Parra et al., 2015; González-Domenech et al., 2020). These findings indicate that MSM TCs are playing a main role in spreading HIV-1 infections and are consistent with the fact that MSM is the main frequent mode of HIV-1 transmission in Spain since the early 2000s (Nuñez et al., 2018).

In Spain, 36% of HIV-1 infections diagnosed in 2019 were in migrants (HIV STI and Hepatitis Surveillance Unit, 2020). Migrants play an important role in HIV-1 epidemic spread, contributing to bridging between networks from different geographic areas (Grossman et al., 2015; Su et al., 2018; Grabowski et al., 2020). In this study, we have examined the frequency of membership in Spanish TCs of migrants from different geographic areas. We have found that Latin Americans and Africans showed lower odds of belonging to TCs than Spaniards. However, there were important differences between these migrant populations, as 35% of Latin Americans were included in TCs, compared to only 1.2% of Sub-Saharan Africans, suggesting that HIV-1 transmission between Latin Americans and Spaniards is relatively common, presumably derived from cultural and linguistic affinities (Osorno-González de León et al., 2021). Similarly to our results, a low percentage of Sub-Saharan Africans in TCs have been also reported in Italy (Fabeni et al., 2020), Belgium (Verhofstede et al., 2018), and Madrid, Spain (Yebra et al., 2013), while in Israel, cross-ethnic HIV-1 spread from Israeli-born individuals to Ethiopian-born immigrants was rare (Grossman et al., 2015). In our study, TCs are mainly MSM-driven. In contrast, we have found high frequencies of females (52%) and heterosexuals (88%) in Sub-Saharan Africans, which, in addition to the likely HIV-1 acquisition in their countries of origin, can explain the low proportion of individuals in this migrant population belonging to TCs. Indeed, we have observed that Sub-Saharan Africans are infected frequently with genetic forms circulating in Africa.

Determining the origin of HIV-1 infections in the migrant population would allow proper allocation of resources for preventive measures (Fakoya et al., 2015; Álvarez del Arco et al., 2017). Previous studies based on models using clinical and behavioral data estimated high levels of post-migration HIV acquisition in Europe (Desgrées-du-Loû et al., 2015;

Álvarez del Arco et al., 2017), up to 71% among migrants in Spain (Álvarez del Arco et al., 2017). This figure could be an overestimate due to biases in data collection and failure to consider infections acquired by migrants when visiting their countries, as migrants' mobility is associated with increased risk for HIV acquisition (Fenton et al., 2001; Dias et al., 2020). Phylogenetic analyses can provide additional information on the geographical area of HIV-1 acquisition. In our study, the lower percentage of foreigners found in TCs could be due to infections with virus strains which are not circulating in Spain, suggestive of imported infections, which would be especially frequent in Sub-Saharan Africans. Further studies using a combined phylogenetic and epidemiological approach can address the accuracy of the origin of the HIV-1 infections in the migrant populations in Spain.

With regard to the role of immigration on the origin of HIV-1 TCs in Spain, it is also interesting to note that 21% TCs identified in this study are of non-subtype B genetic forms, with their ancestry showing some correlation with the predominant geographic origins of the immigrant population in Spain. Thus, the fact that 16 TCs derive from genetic forms circulating in Latin America and Africa may correlate to the high number of immigrants from these geographic areas living in Spain, representing around 36 and 18%, respectively, of the approximately 7.3 million immigrants (Instituto Nacional de Estadistica, 2021). Similarly, the expansion of a large F1 subsubtype cluster of Romanian origin (F1_3) in Spain (Delgado et al., 2019) may relate to the fact that among European immigrants in Spain, Romanians are the most numerous, representing around 8% of all immigrants (Instituto Nacional de Estadistica, 2021).

Subtype B was the most represented genetic form associated to TCs. Only subtype F infections (76% of NDs in TCs) showed a higher frequency than subtype B infections (56% of NDs in TCs) of belonging to TCs. This was mainly due to the presence of two large TCs of F1 subsubtype, designated F1_1 (205 individuals) and F1_3 (36 individuals), which have spread among MSM mainly in Galicia and Basque Country, respectively (Delgado et al., 2015, 2019). Interestingly, two other large non-subtype B TCs associated with MSM, of CRF02_AG and A1 subsubtype, are also present in these regions (Delgado et al., 2019). In Western Europe, the HIV-1 epidemic among MSM is dominated by subtype B. However, the epidemic among MSM in Spain is becoming increasingly diverse through the expansion of multiple non-subtype B TCs, comprising or related to viruses circulating in other countries (González-Domenech et al., 2018; Delgado et al., 2019). This phenomenon has been also documented in other European countries (Dauwe et al., 2015; Ragonnet-Cronin et al., 2016b; Verhofstede et al., 2018; Fabeni et al., 2019; Ramirez et al., 2021), where viruses introduced from abroad have expanded successfully among the MSM population.

In general, there was not a clear temporal trend in the proportion of patients belonging to TC, although in Galicia our data show an increase in the proportion of patients in TCs in the last years. However, trends are difficult to evaluate due to the presence of fast-growing TCs in some periods.

A different distribution of TCs is responsible for the spread of HIV-1 in Galicia and the Basque Country. Large MSM TCs have spread locally in both regions and the majority of the identified TCs were associated with this population group. However, we found a higher frequency of TCs associated with PWID or heterosexual transmission in Galicia. Galician patients have 1.5 higher odds of belonging to TCs than individuals from the Basque Country. This can be related to the low percentage of migrants in Galicia, who have a reduced association to TCs, and the high number of patients belonging to the large F1_1 TC, currently comprising 205 individuals, which has successfully spread in Galicia and other Spanish regions (Delgado et al., 2015), even in other European countries (Delgado et al., 2015; Vinken et al., 2019). We have found an association of young patients (age range 20–29 years) to TCs in Galicia, similarly to the findings in other studies (Brenner et al., 2017; Dennis et al., 2018; Verhofstede et al., 2018; Fabeni et al., 2019; Paraskevis et al., 2019; González-Domenech et al., 2020), but not in the Basque Country. Public health interventions for reducing high risk behavior and control measures in this age group in Galicia should be implemented to decrease HIV-1 transmissions in the region. Moreover, our results highlight that even regions with comparable demographic features can show a different HIV-1 epidemic pattern, likely due to specific social and cultural differences at the regional level.

In this study, we have shown that MSM TCs are a keystone of the HIV-1 epidemic at regional and probably also at the country level, according to the propagation of the large MSM TCs identified in this study in different regions of Spain. Consequently, reinforcement of public health measures for preventing HIV-1 infections in MSM, such as behavioral interventions to reduce risky practices, pre-exposure prophylaxis (Grant et al., 2010; McCormack et al., 2016; Riddell et al., 2018), and early diagnosis and treatment of HIV-1 infections (European Centre for Disease Prevention and Control, 2015; United Nations Population Fund Global Forum on MSM & HIV United Nations Development Programme World Health Organization United States Agency for International Development World Bank, 2015) are recommended.

Implementation of preventive public health measures targeting MSM could reduce indirectly the spread of HIV-1 in Spain, which is mainly MSM-driven, in other population groups. National surveillance data in the United States suggest that infections among heterosexual women predominantly originate from MSM (Oster et al., 2015). Moreover, the presence of female and self-declared heterosexual male individuals in MSM-associated TCs is frequent (Hué et al., 2014; Esbjörnsson et al., 2016; Ragonnet-Cronin et al., 2018; Verhofstede et al., 2018), as we have also observed in the large TCs in our study (**Table 3**). Mixed demographics in TCs allows the spread of HIV-1 infections between different population groups, as we have seen in the expansion of MSM-associated TCs to heterosexuals in Spain (Delgado et al., 2019), indicating the importance of monitoring the evolution of TCs for implementing proper control measures in each population group. Studies to assess the contribution of MSMs to HIV-1 transmission to other population groups are needed in Spain.

The high representativeness of our cohort, especially in the Basque Country, suggests that the largest and most spread

TCs in these regions have been identified, which supports the conclusions of the study. However, our data may have some limitations as some of the NDs from Basque Country and Galicia during the studied period were not analyzed and not all the TCs propagating in Spain were detected in our phylogenetic analysis.

Our study has provided valuable data on the HIV-1 epidemic in Spain, identifying specific features at geographical and population levels for tailoring more specific public health interventions. The implementation of a molecular HIV-1 surveillance system to monitor the evolution of the epidemic in Spain could allow to promptly detect rapidly expanding TCs needing urgent investigation and the implementation of additional public health measures to prevent their spread (Centers for Disease Control and Prevention, 2018), as well as to readjust the control measures in place according to the evolution of the TCs.

## MEMBERS OF THE SPANISH GROUP FOR THE STUDY OF NEW HIV DIAGNOSES

Basque Country: Hospital Universitario de Araba, Vitoria: Andrés Canut-Blasco, José Joaquín Portu, Carmen Gómez-González; Hospital Universitario de Basurto, Bilbao: Josefa Muñoz, Mª Carmen Nieto, María Zuriñe Zubero, Silvia Hernáez-Crespo, Estibaliz Ugalde; Hospital Universitario de Cruces, Barakaldo: Luis Elorduy, Elena Bereciartua, Leyre López Soria; Hospital de Galdakao: Mª José López de Goicoechea, José Mayo; Hospital Universitario Donostia: Gustavo Cilla, Julio Arrizabalaga, José Antonio Iribarren, Mª Jesús Bustinduy, Mª Julia Echevarría. Mª Yolanda Salicio, David Grandioso. Galicia: Área sanitaria de Ferrol: Ana Mariño, Patricia Ordóñez, Hortensia Álvarez, Nieves Valcarce; Complejo Hospitalario Universitario de A Coruña: Ángeles Cañizares, Mª Ángeles Castro. Hospital Universitario Lucus Augusti, Lugo: Ramón Rabuñal-Rey, María José García-País, Mª José Gude-González, Pilar Alonso-García, Antonio Moreno-Flores; Complejo Hospitalario Universitario de Ourense: Juan García Costa, Ricardo Fernández-Rodríguez, Raúl Rodríguez-Pérez, Jorge Guitián, María Dolores Díaz-López, María Genoveva Naval-Calviño; Complejo Hospitalario Universitario de Vigo: Celia Miralles, Antonio Ocampo, Sonia Pérez-Castro, Jorge Julio Cabrera; Complejo Hospitalario Universitario de Pontevedra: Julio Díz-Arén, Matilde Trigo, Mª Ángeles Pallarés. Navarra: Complejo Hospitalario de Navarra, Pamplona: Carmen Ezpeleta Baquedano, Carmen Martín Salas, Irati Arregui García, María Gracia Ruiz de Alda. Madrid: Centro Sanitario Sandoval, Madrid: Jorge del Romero, Carmen Rodríguez, Mar Vera, Óskar Ayerdi, Eva Orviz; Hospital Universitario de Fuenlabrada: María Isabel García-Arata, Santiago Prieto-Menchero; Hospital Clínico Universitario San Carlos, Madrid: Esther Culebras, Icíar Rodríguez-Avial; Hospital Universitario Fundación Jiménez Díaz, Madrid: Raquel Téllez-Pérez, Olalla Calabia-González, Alfonso Cabello-Úbeda, Miguel Górgolas Hernández-Mora; Hospital Universitario Severo Ochoa, Leganés: Sara María Quevedo, Lucía Puente, Manuel del Álamo; Hospital Fundación Alcorcón, Madrid: Carolina Campelo Gutiérrez, María José Goyanes Galán; Castilla y León: Hospital Clínico Universitario de Valladolid: Carmen Hinojosa, Carlos Dueñas, Begoña Monteagudo, Edita Sánchez; Hospital Río Hortega, Valladolid: Jessica Abadía, Belén Lorenzo Vidal; Hospital Virgen de la Concha, Zamora: Teresa Martín-Domínguez, Rosa Martínez-González. La Rioja: Hospital San Pedro, Logroño: José Ramón Blanco, Miriam Blasco. Aragón: Hospital Universitario Miguel Servet, Zaragoza: Ana María Martínez-Sapiña, Piedad Arazo. Castilla-La Mancha: Hospital Universitario de Guadalajara: Alejandro González Praetorius, Complejo Hospitalario de Toledo: César Gómez-Hernando, José Largo-Pau; Comunitat Valenciana: Hospital Universitari Sant Joan d'Alacant: Fernando Buñuel, Ana Infante.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/genbank/, KT276255, KT276260, KT276263-KT276264, KT276266-KT276267, KT276270-KT276271, KU685562-KU685565, KU685567, KU685569-KU685575, KU685577, KU685581-KU685582, KU685586, KU685588-KU685590, KU685592, KX534325, KX534329, KX534331, KY465968, KY496624, KY514084, KY989950, MF999250-MF999256, MF999258-MF999261, MK177721-MK177733, MK177735-MK177752, MK177754-MK177757, MK177761-MK177772, MK177783-MK17785, MK177787-MK177790, MK177792, MK177795-MK177796, MK177798-MK177799, MK177801-MK177803, MK177805-MK177807, MK298150, MT436242, MT436244, MT436246-MT436247, MT436249-MT436250, MW344920-MW344921, MW584217-MW584224, OK011532, OK011542, OK011545-OK011546, OK011549, OK148912, OK148914, OK148917-OK148919, OK148921, OK148941-OK148942, OK148953-OK148957, OK148961, OK148965, OK148968-OK148972, OK148974, OL982314-OL982315, and OM914651-OM914711.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Bioethics and Animal Well-being Committee of Instituto de Salud Carlos III. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

HG, ED, and MT conceived the study and supervised the experimental work. ED, HG, MT, and JC-G processed sequences and performed phylogenetic analyses. HG and ED performed data curation. HG, AD, and LG performed statistical analyses. SB, MS, VM, and EG-B performed experimental work. The members of the Spanish Group for the Study of New HIV Diagnoses recruited patients and obtained epidemiological and clinical data. HG wrote the manuscript draft, with contributions to the text by AD, MT, and ED. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was funded through Acción Estratégica en Salud Intramural (AESI), Instituto de Salud Carlos III, Project "Estudios sobre vigilancia epidemiológica molecular del VIH-1 en España," PI16CIII/00033 and Project "Epidemiología molecular del VIH-1 en España y su utilidad para investigaciones biológicas y en vacunas"PI19CIII/0042; Red de Investigación en SIDA (RIS), Instituto de Salud Carlos III, Subdirección General de Evaluación y Fondo Europeo de Desarrollo Regional (FEDER), Plan Nacional ICDCI, project RD16ISCIII/0002/0004; and scientific agreements with Consellería de Sanidade, Government of Galicia (MVI 1004/16), and Osakidetza-Servicio Vasco de Salud, Government of Basque Country (MVI 1001/16).

## ACKNOWLEDGMENTS

The authors would like to thank José Antonio Taboada, from Consellería de Sanidade, Xunta de Galicia, and Daniel Zulaika, from Osakidetza-Servicio Vasco de Salud, for their support of this study, and the personnel at the Genomic Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, for technical assistance in sequencing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.782609/full#supplementary-material

## REFERENCES

AIDS Study Group (GESIDA) of the Spanish Society of Infectious Diseases and Clinical Microbiology and the National AIDS Plan. (2020). Consensus document on antirretroviral therapy in adults infected by the human inmunodeficiency virus (Updated 2020). Available at: https://gesida-seimc.org/wp-content/uploads/2020/07/TAR_GUIA_GESIDA_2020_COMPLETA_Julio.pdf (Accessed July 16, 2020).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Álvarez del Arco, D., Fakoya, I., Thomadakis, C., Pantazis, N., Touloumi, G., Gennotte, A. F., et al. (2017). High levels of postmigration HIV acquisition within nine European countries. *AIDS* 31, 1979–1988. doi: 10.1097/QAD.0000000000001571

Board, A. R., Oster, A. M., Song, R., Gant, Z., Linley, L., Watson, M., et al. (2020). Geographic distribution of HIV transmission networks in the United States. *J. Acquir. Immune Defic. Syndr.* 85, e32–e40. doi: 10.1097/QAI.0000000000002448

Brenner, B. G., Ibanescu, R. I., Hardy, I., Stephens, D., Otis, J., Moodie, E., et al. (2017). Large cluster outbreaks sustain the HIV epidemic among MSM in Quebec. *AIDS* 31, 707–717. doi: 10.1097/QAD.0000000000001383

Brenner, B., Wainberg, M. A., and Roger, M. (2013). Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. *AIDS* 27, 1045–1057. doi: 10.1097/QAD.0b013e32835cffd9

Brown, A. J. L., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., and Dunn, D. T. (2011). Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* 204, 1463–1469. doi: 10.1093/infdis/jir550

Campbell, E. M., Patala, A., Shankar, A., Li, J. F., Johnson, J. A., Westheimer, E., et al. (2020). Phylodynamic analysis complements partner services by identifying acute and unreported HIV transmission. *Viruses* 12:145. doi: 10.3390/v12020145

Centers for Disease Control and Prevention (2018). Detecting, investigating, and responding to HIV transmission clusters. Available at: https://www.cdc.gov/hiv/pdf/funding/announcements/ps18-1802/CDC-HIV-PS18-1802-AttachmentE-Detecting-Investigating-and-Responding-to-HIV-Transmission-Clusters.pdf (Accessed September 13, 2018).

Cuevas, M. T., Fernández-García, A., Sánchez-García, A., González-Galeano, M., Pinilla, M., Sánchez-Martínez, M., et al. (2009a). Incidence of non-B subtypes of HIV-1 in Galicia, Spain: high frequency and diversity of HIV-1 among men who have sex with men. *Euro Surveill.* 14:19413. doi: 10.2807/ese.14.47.19413-en

Cuevas, M. T., Muñoz-Nieto, M., Thomson, M. M., Delgado, E., Iribarren, J. A., Cilla, G., et al. (2009b). HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. *J. Acquir. Immune Defic. Syndr.* 51, 99–103. doi: 10.1097/QAI.0b013e318199063e

Dauwe, K., Mortier, V., Schauvliege, M., Van Den Heuvel, A., Fransen, K., Servais, J. Y., et al. (2015). Characteristics and spread to the native population of HIV-1 non-B subtypes in two European countries with high migration rate. *BMC Infect. Dis.* 15:524. doi: 10.1186/s12879-015-1217-0

Delgado, E., Benito, S., Montero, V., Cuevas, M. T., Fernández-García, A., Sánchez-Martínez, M., et al. (2019). Diverse large HIV-1 non-subtype B clusters are spreading among men who have sex with men in Spain. *Front. Microbiol.* 10:655. doi: 10.3389/fmicb.2019.00655

Delgado, E., Cuevas, M. T., Domínguez, F., Vega, Y., Cabello, M., Fernández-García, A., et al. (2015). Phylogeny and phylogeography of a recent HIV-1 subtype F outbreak among men who have sex with men in Spain deriving from a cluster with a wide geographic circulation in Western Europe. *PLoS One* 10:e0143325. doi: 10.1371/journal.pone.0143325

Dennis, A. M., Volz, E., Frost, A. S. M. S., Hossain, M., Poon, A. F. Y., Rebeiro, P. F., et al. (2018). HIV-1 transmission clustering and phylodynamics highlight the important role of young men who have sex with men. *AIDS Res. Hum. Retrovir.* 34, 879–888. doi: 10.1089/AID.2018.0039

Desgrées-du-Loû, A., Pannetier, J., Ravalihasy, A., Gosselin, A., Supervie, V., Panjo, H., et al. (2015). Sub-Saharan African migrants living with HIV acquired after migration, France, ANRS PARCOURS study, 2012 to 2013. *Euro Surveill.* 20:30065. doi: 10.2807/1560-7917.ES.2015.20.46.30065

Dias, S., Gama, A., Loos, J., Roxo, L., Simões, D., and Nöstlinger, C. (2020). The role of mobility in sexual risk behaviour and HIV acquisition among sub-Saharan African migrants residing in two European cities. *PLoS One* 15:e0228584. doi: 10.1371/journal.pone.0228584

Esbjörnsson, J., Mild, M., Audelin, A., Fonager, J., Skar, H., Bruun-Jørgensen, L., et al. (2016). HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic countries. *Virus Evol.* 2:vew010. doi: 10.1093/ve/vew010

European Centre for Disease Prevention and Control (2015). ECDC Guidance: HIV and STI prevention among men who have sex with men. Available at: https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/hiv-sti-prevention-among-men-who-have-sex-with-men-guidance.pdf (Accessed June 17, 2015).

European Centre for Disease Prevention and Control and WHO Regional Office for Europe (2020). HIV/AIDS surveillance in Europe 2020–2019 data. Available at: https://www.ecdc.europa.eu/sites/default/files/documents/hiv-surveillance-report-2020.pdf (Accessed November 26, 2020).

Fabeni, L., Alteri, C., Berno, G., Scutari, R., Orchi, N., De, C. G., et al. (2019). Characterisation of HIV-1 molecular transmission clusters among newly diagnosed individuals infected with non-B subtypes in Italy. *Sex. Transm. Infect.* 95, 619–625. doi: 10.1136/sextrans-2019-054017

Fabeni, L., Rozera, G., Berno, G., Giombini, E., Gori, C., Orchi, N., et al. (2021). Molecular transmission dynamics of primary HIV infections in Lazio region, years 2013-2020. *Viruses* 13:176. doi: 10.3390/v13020176

Fabeni, L., Santoro, M. M., Lorenzini, P., Rusconi, S., Gianotti, N., Costantini, A., et al. (2020). Evaluation of HIV transmission clusters among natives and foreigners living in Italy. *Viruses* 12:791. doi: 10.3390/v12080791

Fakoya, I., Ávarez del Arco, D., Woode-Owusu, M., Monge, S., Rivero-Montesdeoca, Y., Delpech, V., et al. (2015). A systematic review of post-migration acquisition of HIV among migrants from countries with generalised HIV epidemics living in Europe: implications for effectively managing HIV prevention programmes and policy. *BMC Public Health* 15:561. doi: 10.1186/s12889-015-1852-9

Fenton, K. A., Chinouya, M., Davidson, O., and Copas, A. (2001). HIV transmission risk among sub-Saharan Africans in London travelling to their countries of origin. *AIDS* 15, 1442–1445. doi: 10.1097/00002030-200107270-00017

González-Alba, J. M., Holguín, A., García, R., García-Bujalance, S., Alonso, R., Suáez, A., et al. (2011). Molecular surveillance of HIV-1 in Madrid, Spain: a phylogeographic analysis. *J. Virol.* 85, 10755–10763. doi: 10.1128/JVI.00454-11

González-Domenech, C. M., Sena-Corrales, G., Viciana-Ramos, I., Palacios-Muñoz, R., Mora-Navas, L., Clavijo-Frutos, E., et al. (2020). High prevalence of sequences included in transmission clusters within newly diagnosed HIV-1 patients in southern Spain (2004-2015). *Microb. Drug Resist.* 26, 1090–1097. doi: 10.1089/mdr.2019.0344

González-Domenech, C. M., Viciana, I., Delaye, L., Mayorga, M. L., Palacios, R., de la Torre, J., et al. (2018). Emergence as an outbreak of the HIV-1 CRF19_cpx variant in treatment-naïve patients in southern Spain. *PLoS One* 13:e0190544. doi: 10.1371/journal.pone.0190544

Grabowski, M. K., Lessler, J., Bazaale, J., Nabukalu, D., Nankinga, J., Nantume, B., et al. (2020). Migration, hotspots, and dispersal of HIV infection in Rakai, Ugand. *Nat. Commun.* 11:976. doi: 10.1038/s41467-020-14636-y

Grant, R. M., Lama, J. R., Anderson, P. L., McMahan, V., Liu, A. Y., Vargas, L., et al. (2010). Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *N. Engl. J. Med.* 363, 2587–2599. doi: 10.1056/NEJMoa1011205

Grossman, Z., Avidor, B., Mor, Z., Chowers, M., Levy, I., Shahar, E., et al. (2015). A population-structured HIV epidemic in Israel: roles of risk and ethnicity. *PLoS One* 10:e0135061. doi: 10.1371/journal.pone.0135061

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.

Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J., and Esbjornsson, J. (2017). Defining HIV-1 transmission clusters based on sequence data. *AIDS* 31, 1211–1222. doi: 10.1097/QAD.0000000000001470

HIV STI and Hepatitis Surveillance Unit (2020). Vigilancia del VIH y sida en España: Sistema de información sobre nuevos diagnósticos de VIH y registro nacional de casos de sida. Plan Nacional sobre el Sida. Available at: https://www.mscbs.gob.es/ca//ciudadanos/enfLesiones/enfTransmisibles/sida/vigilancia/Informe_VIH_SIDA_20201130.pdf (Accessed November 27, 2020).

Hoenigl, M., Chaillon, A., Kessler, H. H., Haas, B., Stelzl, E., Weninger, K., et al. (2016). Characterization of HIV transmission in south-East Austria. *PLoS One* 11:e0151478. doi: 10.1371/journal.pone.0151478

Holguín, A., Pena, M. J., Troncoso, F., and Soriano, V. (2007). Introduction of non-B subtypes among Spaniards newly diagnosed with HIV type 1 in the Canary Islands. *AIDS Res. Hum. Retrovir.* 23, 498–502. doi: 10.1089/aid.2006.0191

Hué, S., Brown, A. E., Ragonnet-Cronin, M., Lycett, S. J., Dunn, D. T., Fearnhill, E., et al. (2014). Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS* 28, 1967–1975. doi: 10.1097/QAD.0000000000000383

Hué, S., Clewley, J. P., Cane, P. A., and Pillay, D. (2004). HIV-1 *pol* gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18, 719–728. doi: 10.1097/00002030-200403260-00002

Hué, S., Pillay, D., Clewley, J. P., and Pybus, O. G. (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4425–4429. doi: 10.1073/pnas.0407534102

Hughes, G. J., Fearnhill, E., Dunn, D., Lycett, S. J., Rambaut, A., and Leigh Brown, A. J. (2009). Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog.* 5:e1000590. doi: 10.1371/journal.ppat.1000590

Instituto Nacional de Estadistica (2021). Población residente. Available at: https://www.ine.es/jaxiT3/Datos.htm?t=9681 (Accessed September 8, 2021).

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/JVI.73.1.152-160.1999

McCormack, S., Dunn, D. T., Desai, M., Dolling, D. I., Gafos, M., Gilson, R., et al. (2016). Pre-exposure prophylaxis to prevent the acquisition of HIV-1 infection (PROUD): effectiveness results from the pilot phase of a pragmatic open-label randomised trial. *Lancet* 387, 53–60. doi: 10.1016/S0140-6736(15)00056-2

Nuñez, O., Hernando, V., and Díaz, A. (2018). Estimating the number of people living with HIV and the undiagnosed fraction in Spain in 2013. *AIDS* 32, 2573–2581. doi: 10.1097/QAD.0000000000001989

Osorno-González de León, M. F., Delgado, E., Benito, S., Sánchez, M., Gil, H., Montero, V., et al. (2021). 11th Conference on HIV Science, abstract PEC252. Available at: https://ias2021.org/wp-content/uploads/2021/07/IAS2021_Abstracts_web.pdf (Accessed July 18–21, 2021).

Oster, A. M., France, A. M., Panneer, N., Bañez-Ocfemia, M. C., Campbell, E., Dasgupta, S., et al. (2018). Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. *J. Acquir. Immune Defic. Syndr.* 79, 543–550. doi: 10.1097/QAI.0000000000001856

Oster, A. M., Wertheim, J. O., Hernandez, A. L., Ocfemia, M. C., Saduvala, N., and Hall, H. I. (2015). Using molecular HIV surveillance data to understand transmission between subpopulations in the United States. *J. Acquir. Immune Defic. Syndr.* 70, 444–451. doi: 10.1097/QAI.0000000000000809

Paraskevis, D., Beloukas, A., Stasinos, K., Pantazis, N., de Mendoza, C., Bannert, N., et al. (2019). HIV-1 molecular transmission clusters in nine European countries and Canada: association with demographic and clinical factors. *BMC Med.* 17:4. doi: 10.1186/s12916-018-1241-1

Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I., and Hatzakis, A. (2016). The application of HIV molecular epidemiology to public health. *Infect. Genet. Evol.* 46, 159–168. doi: 10.1016/j.meegid.2016.06.021

Parczewski, M., Leszczyszyn-Pynka, M., Witak-Jedra, M., Szetela, B., Gasiorowski, J., Knysz, B., et al. (2017). Expanding HIV-1 subtype B transmission networks among men who have sex with men in Poland. *PLoS One* 12:e0172473. doi: 10.1371/journal.pone.0172473

Patiño-Galindo, J. A., Thomson, M. M., Pérez-Álvarez, L., Delgado, E., Cuevas, M. T., Fernández-García, A., et al. (2016). Transmission dynamics of HIV-1 subtype B in the Basque Country, Spain. *Infect. Genet. Evol.* 40, 91–97. doi: 10.1016/j.meegid.2016.02.028

Patiño-Galindo, J. A., Torres-Puente, M., Bracho, M. A., Alastrué, I., Juan, A., Navarro, D., et al. (2017a). Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain. *PLoS One* 12:e0171062. doi: 10.1371/journal.pone.0171062

Patiño-Galindo, J. A., Torres-Puente, M., Bracho, M. A., Alastrué, I., Juan, A., Navarro, D., et al. (2017b). The molecular epidemiology of HIV-1 in the Comunidad Valenciana (Spain): analysis of transmission clusters. *Sci. Rep.* 7:11584. doi: 10.1038/s41598-017-10286-1

Pérez-Parra, S., Chueca, N., Ávarez, M., Pasquau, J., Omar, M., Collado, A., et al. (2016). Phylodynamic and phylogeographic profiles of subtype B HIV-1 epidemics in South Spain. *PLoS One* 11:e0168099. doi: 10.1371/journal.pone.0168099

Pérez-Parra, S., Chueca-Porcuna, N., Ávarez-Estevez, M., Pasquau, J., Omar, M., Collado, A., et al. (2015). Los estudios de resistencia a antirretrovirales como herramienta para el análsis de los clusters de transmisión del virus de la inmunodefiencia humana. *Enferm. Infecc. Microbiol. Clin.* 33, 603–608. doi: 10.1016/j.eimc.2014.11.016

Petersen, A., Cowan, S. A., Nielsen, J., Fischer, T. K., and Fonager, J. (2018). Characterisation of HIV-1 transmission clusters and drug-resistant mutations in Denmark, 2004 to 2016. *Euro Surveill.* 23:1700633. doi: 10.2807/1560-7917.ES.2018.23.44.1700633

Pineda-Peña, A. C., Pingarilho, M., Li, G., Vrancken, B., Libin, P., Gomes, P., et al. (2019). Drivers of HIV-1 transmission: the Portuguese case. *PLoS One* 14:e0218226. doi: 10.1371/journal.pone.0218226

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Ragonnet-Cronin, M., Hué, S., Hodcroft, E. B., Tostevin, A., Dunn, D., Fawcett, T., et al. (2018). Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *Lancet HIV* 5, e309–e316. doi: 10.1016/S2352-3018(18)30062-6

Ragonnet-Cronin, M., Lycett, S. J., Hodcroft, E. B., Hué, S., Fearnhill, E., Brown, A. E., et al. (2016a). Transmission of non-B HIV subtypes in the United Kingdom is increasingly driven by large non-heterosexual transmission clusters. *J. Infect. Dis.* 213, 1410–1418. doi: 10.1093/infdis/jiv758

Ragonnet-Cronin, M., Shilaih, M., Günthard, H. F., Hodcroft, E. B., Böni, J., FearnHill, E., et al. (2016b). A direct comparision of two densenly sampled HIV epidemics: the UK and Switzerland. *Sci. Rep.* 6:32251. doi: 10.1038/SREP32251

Ramirez, J. J. D., Ballouz, T., Nguyen, H., Kusejko, K., Chaudron, S. E., Huber, M., et al. (2021). Increasing frequency and transmission of HIV-1 non-B subtypes among men who have sex with men in the Swiss HIV cohort study. *J. Infect. Dis.* 225, 306–316. doi: 10.1093/infdis/jiab360

Riddell, J., Amico, K. R., and Mayer, K. H. (2018). HIV preexposure prophylaxis: a review. *JAMA* 319, 1261–1268. doi: 10.1001/jama.2018.1917

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201

Su, L., Liang, S., Hou, X., Zhong, P., Wei, D., Fu, Y., et al. (2018). Impact of worker emigration on HIV epidemics in labour export areas: a molecular epidemiology investigation in Guangyuan, China. *Sci. Rep.* 8:16046. doi: 10.1038/s41598-018-33996-6

Thomson, M. M., Delgado, E., Manjón, N., Ocampo, A., Villahermosa, M. L., Mariño, A., et al. (2001). HIV-1 genetic diversity in Galicia Spain: BG intersubtype recombinant viruses circulating among injecting drug users. *AIDS* 15, 509–516. doi: 10.1097/00002030-200103090-00010

United Nations Population Fund Global Forum on MSM & HIV United Nations Development Programme World Health Organization United States Agency for International Development World Bank (2015). Implementing comprehensive HIV and STI programmes with men who have sex with men: practical guidance for collaborative interventions. Available at: https://www.unfpa.org/es/node/13155 (Accessed September 8, 2015).

Vasylyeva, T. I., Friedman, S. R., Paraskevis, D., and Magiorkinis, G. (2016). Integrating molecular epidemiology and social network analysis to study infectious diseases: towards a socio-molecular era for public health. *Infect. Genet. Evol.* 46, 248–255. doi: 10.1016/j.meegid.2016.05.042

Vasylyeva, T. I., Zarebski, A., Smyrnov, P., Williams, L. D., Korobchuk, A., Liulchuk, M., et al. (2020). Phylodynamics helps to evaluate the impact of an HIV prevention intervention. *Viruses* 12:469. doi: 10.3390/v12040469

Vega, Y., Delgado, E., Fernández-García, A., Cuevas, M. T., Thomson, M. M., Montero, V., et al. (2015). Epidemiological surveillance of HIV-1 transmitted drug resistance in Spain in 2004-2012: relevance of transmission clusters in the propagation of resistance mutations. *PLoS One* 10:e0125699. doi: 10.1371/journal.pone.0125699

Verhofstede, C., Dauwe, K., Fransen, K., Van, L. K., Van den Wijngaert, S., Ruelle, J., et al. (2018). Phylogenetic analysis of the Belgian HIV-1 epidemic reveals that local transmission is almost exclusively driven by men having sex with men despite presence of large African migrant communities. *Infect. Genet. Evol.* 61, 36–44. doi: 10.1016/j.meegid.2018.03.002

Viciana, I., González-Domenech, C. M., Palacios, R., Delgado, M., Del, A. A., Tellez, F., et al. (2016). Clinical, virological and phylogenetic characterization of a multiresistant HIV-1 strain outbreak in naive patients in southern Spain. *J. Antimicrob. Chemother.* 71, 357–361. doi: 10.1093/jac/dkv332

Vinken, L., Fransen, K., Cuypers, L., Alexiev, I., Balotta, C., Debaisieux, L., et al. (2019). Earlier initiation of antiretroviral treatment coincides with an initial control of the HIV-1 sub-subtype F1 outbreak among men-having-sex-with-men in Flanders, Belgium. *Front. Microbiol.* 10:613. doi: 10.3389/fmicb.2019.00613

Wertheim, J. O., Kosakovsky Pond, S. L., Forgione, L. A., Mehta, S. R., Murrell, B., Shah, S., et al. (2017). Social and genetic networks of HIV-1 transmission in New York City. *PLoS Pathog.* 13:e1006000. doi: 10.1371/journal.ppat.1006000

Yebra, G., Holguín, A., Pillay, D., and Hué, S. (2013). Phylogenetic and demographic characterization of HIV-1 transmission in Madrid, Spain. *Infect. Genet. Evol.* 14, 232–239. doi: 10.1016/j.meegid.2012.12.006

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 12 April 2022
doi: 10.3389/fmicb.2022.854974

# HIV Capsid Protein Genetic Diversity Across HIV-1 Variants and Impact on New Capsid-Inhibitor Lenacapavir

*Paloma Troyano-Hernáez, Roberto Reinosa and África Holguín**

*HIV-1 Molecular Epidemiology Laboratory, Department of Microbiology, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP), Red en Investigación Translacional en Infecciones Pediátricas (RITIP), Madrid, Spain*

The HIV p24 capsid protein has an essential, structural, and functional role in the viral replication cycle, being an interesting target for vaccine design, diagnostic tests, and new antiretroviral drugs (ARVs). The HIV-1 variability poses a challenge for the accuracy and efficiency of diagnostic and treatment tools. This study analyzes p24 diversity among HIV-1 variants and within its secondary structure in HIV-1 M, O, P, and N groups. All available HIV-1 p24 nucleotide sequences were downloaded from the Los Alamos HIV Sequence Database, selecting 23,671 sequences belonging to groups O, N, P, and M (9 subtypes, 7 sub-sub types, and 109 circulating recombinant forms or CRFs). Using a bioinformatics tool developed in our laboratory (EpiMolBio program), we analyzed the amino acid conservation compared to the HXB2 subtype B reference sequence and the V-markers, or amino acid changes that were specific for each variant with at least 10 available sequences. We inferred the p24 consensus sequence for HIV-1 and for each group to analyze the overall conservation in p24 main structural regions, reporting the percentage of substitutions per variant affecting the capsid assembly and molecule-binding, including those associated with resistance to the new capsid-inhibitor lenacapavir, and the key residues involved in lenacapavir-p24 interaction, according to the bibliography. Although the overall structure of p24 was highly conserved, the conservation in the secondary structure varied between HIV-1 variants and the type of secondary structure. All HIV-1 variants presented >80% amino acid conservation vs. HXB2 reference sequence, except for group M sub-subtype F1 (69.27%). Mutants affecting the capsid assembly or lenacapavir capsid-binding were found in <1% of the p24 consensus sequence. Our study reports the HIV-1 variants carrying 14 unique single V-markers in 9/38 group M variants and the level of p24 conservation in each secondary structure region among the 4 HIV-1 groups and group M variants, revealing no natural resistance to lenacapavir in any HIV-1 variant. We present a thorough analysis of p24 variability among all HIV-1 variants circulating to date. Since p24 genetic variability can impact the viral replication cycle and the efficacy of new p24-based diagnostic, therapeutic, and vaccine strategies, conservation studies must consider all HIV-1 variants circulating worldwide.

**Keywords: HIV-1, p24, variants, conservation, lenacapavir, capsid**

***Edited by:***
*Michael M. Thomson,*
*Instituto de Salud Carlos III (ISCIII),*
*Spain*

***Reviewed by:***
*Matteo Negroni,*
*UPR 9002, Institut de Biologie*
*Moléculaire et Cellulaire (CNRS),*
*France*
*Masahiro Yamashita,*
*Columbia University Irving Medical*
*Center, United States*

****Correspondence:***
*África Holguín*
*africa.holguin@salud.madrid.org*

***Specialty section:***
*This article was submitted to*
*Virology,*
*a section of the journal*
*Frontiers in Microbiology*

***Received:*** *14 January 2022*
***Accepted:*** *09 March 2022*
***Published:*** *12 April 2022*

***Citation:***
*Troyano-Hernáez P, Reinosa R and Holguín Á (2022) HIV Capsid Protein Genetic Diversity Across HIV-1 Variants and Impact on New Capsid-Inhibitor Lenacapavir. Front. Microbiol. 13:854974. doi: 10.3389/fmicb.2022.854974*

## INTRODUCTION

The HIV-1 capsid houses the replicative enzymes and viral genomic RNA, protecting them from antiviral factors and cellular sensors of innate immunity (Le Sage et al., 2014), allowing their traffic from entry to near integration sites before fully uncoating (Burdick et al., 2020). Although previous models pointed to an early cytoplasmatic capsid disassembly (Miller et al., 1997; Mamede et al., 2017), recent studies have supported the possibility of a complete or almost intact capsid entering the nucleus (Burdick et al., 2020; Dharan et al., 2020; Li et al., 2021; Zila et al., 2021a). Increasing evidence suggests that the capsid participates in the translocation of viral genomic material into the host nucleus for integration through partial uncoating that allows higher plasticity of the capsid, through structural rearrangements of the nuclear pore, or by both mechanisms (Dharan et al., 2020; Toccafondi et al., 2021; Zhuang and Torbett, 2021; Zila et al., 2021b).

In addition to this structural function, the capsid is involved in different steps of the viral infectious cycle, such as HIV reverse transcription (RT), cytoplasmic trafficking through microtubules, decapsidation and nuclear import of the viral pre-integration complex, integration, and assembly (Sundquist and Kräusslich, 2012; Campbell and Hope, 2015; Novikova et al., 2019; Engelman, 2021; Zila et al., 2021b). Recent *in vitro* studies suggest that the capsid plays a key role in HIV RT contributing directly to its efficiency, maintaining sufficient concentrations of RT, and other core components needed for the long process of RT, excluding cytoplasmatic molecules that may degrade the viral nucleic acids, and interacting with certain molecules, such as IP6 (inositol-hexakisphosphate), which enhances RT *in vitro* by stabilizing the viral capsid (Mallery et al., 2018; Christensen et al., 2020; Aiken and Rousso, 2021). RT requires intact rings or capsid lattice sections with certain stability and geometry to produce double-stranded DNA genomes that are extruded from the ruptured capsid walls (Christensen et al., 2020) while promoting HIV-1 uncoating (Aiken and Rousso, 2021). Moreover, the capsid interacts with multiple host factors that can either promote or prevent virus infection (Stremlau et al., 2006; Fricke et al., 2014; Yamashita and Engelman, 2017; Novikova et al., 2019; Temple et al., 2020; Toccafondi et al., 2021; Wilbourne and Zhang, 2021).

The capsid monomers (p24) are encoded by the *gag* gene segment located between nucleotides 1186 and 1879 of the HIV-1 subtype B HXB2 isolate. The *gag* gene encodes different proteins involved in viral structure and trafficking, assembly, pol protein control, interaction with cellular proteins, and viral egress. Thus, *gag* determines the structure and enzymatic functions in HIV. The main Gag structural proteins are p17 or Matrix, p24 or Capsid (CA) (referred to as p26 in HIV-2), and p7 or Nucleocapsid. These proteins, p6 Gag, and two spacers (p1 and p2) are synthesized from a series of protease-mediated proteolytic reactions at specific cleavage sites (CS) located on the Gag precursor (Pr55$^{Gag}$) and GagPol (Pr160$^{GagPol}$) polyproteins. An additional ribosomal reading frame during translation of the Gag precursor generates the latter polyprotein (Jacks et al., 1988). Previous reports described different CS conservation across HIV-1 variants (Torrecilla et al., 2014).

The mature capsid structure has the shape of a fullerene cone that consists of approximately 1500 p24 monomers assembled in a hexamer lattice with 12 pentameric variations (Zhao et al., 2013). Each p24 comprises 231 amino acids (aa), has an N-terminal domain (NTD) of 145 aa with a β-hairpin and 7 α-helices (H), a C-terminal domain (CTD) of 85 aa with 4 α-helices, and an 11-residue unstructured region (Zhao et al., 2013). The domains are linked by a flexible linker or interdomain linker region (IDR) (aa 146–150) (Jiang et al., 2011). The NTD is responsible for intra-hexamer contacts, and the CTD forms binding dimers to adjacent hexamers (Rihn et al., 2013). In the center of each hexamer is a pore surrounded by six positively charged arginine residues, and the pore is covered by the β-hairpin that can change conformation to open or close it (Jacques et al., 2016). An IP6 molecule binds to the center of the pore, stabilizing the hexamer (Dick et al., 2018; Mallery et al., 2018; Márquez et al., 2018; Renner et al., 2021). The CA also has a major homology region or MHR (aa 153–172) in the CTD with 20 highly conserved aa (Mammano et al., 1994) and a loop (aa 85–93) in the NTD that binds to cyclophilin A or Cyp A (Gamble et al., 1996).

Previous studies of p24 with mutants generated by targeted mutagenesis have revealed that the capsid is extremely intolerant to non-synonymous substitutions, producing defective or reduced infectivity viruses (Rihn et al., 2013; Perilla and Gronenborn, 2016). This high conservation and the fact that it is the most abundant viral protein make of p24, a very interesting region for the design of serological and molecular diagnostic tests for HIV early detection (Gray et al., 2018; Parekh et al., 2019), is the target of new biosensors and nanotechnologies under development (Kosaka et al., 2017; Zhou and Rossi, 2017; Bala et al., 2018; Farzin et al., 2020; Rodríguez-Galet et al., 2022). In fact, p24 is recognized as an alternative early virological biomarker of infection (Gray et al., 2018). The molecular or serological detection of HIV-1 p24 in diagnostic tests relies on primers, labels, or antibodies binding to conserved areas of the protein across HIV variants, ideally detecting all circulating viral variants (Bbosa et al., 2019). However, naturally occurring aa variations and single aa changes in conserved epitopes can lead to the failure of p24 detection and false-negative results (Beelaert and Fransen, 2010; Ly et al., 2012; Vetter et al., 2015; Qiu et al., 2017). Thus, the evaluation of p24 natural variability, as well as of the performance of each different diagnostic assay across HIV-1 variants, is necessary to identify variants not correctly detected due to viral genetic variability (Alvarez et al., 2015, 2017; Kravitz Del Solar et al., 2018; Stone et al., 2018).

Since the capsid's stability and integrity are critical to the normal viral replication cycle and infectivity, the HIV capsid is an exciting target for the design of new ARVs (Rihn et al., 2013; Chen, 2016; Jacques et al., 2016; Novikova et al., 2019; McFadden et al., 2021; Saito and Yamashita, 2021). Furthermore, p24 can induce cellular immune responses [Los Alamos National Laboratory (LANL), 2021a] and has been included in some vaccine strategies (Larijani et al., 2018, 2021a,b). Among the numerous molecules that targeted the capsid, lenacapavir seems the most promising to date (Yant et al., 2019;

Dvory-Sobol et al., 2022; Margot et al., 2022). Capsid-targeting molecules bind to different sites in p24, aiming to alter the capsid stability and morphology, interfering in the assembly or disassembly processes, or competing with host factors resulting in the suppression of viral infectivity (Li et al., 2013; McFadden et al., 2021).

The HIV is one of the most genetically diverse pathogens due to its high mutation and recombination rates, large population size, and rapid replication rate (Hemelaar, 2012; Hemelaar et al., 2019). The HIV-1 is responsible for most HIV infections worldwide and has been divided into four groups, according to genetic homology: M (major or main), N (non-M, non-O) (Simon et al., 1998), O (outlier) (De Leys et al., 1990), and P (Plantier et al., 2009). Group M is the main HIV group related to the present HIV global pandemic (Hemelaar et al., 2019). This group has been subdivided into 10 subtypes (A-D, F-H, J-L) and 8 sub-sub types (A1, A2, A3, A4, A5, A6, F1, F2) (Robertson et al., 2000; Salminen et al., 2000; Yamaguchi et al., 2020), at least 118 circulating recombinant forms (CRFs) (Los Alamos National Laboratory, 2021b), and countless unique recombinant forms (URFs). The CRF is inter-subtype recombinant viruses detected in three or more not epidemiologically linked individuals (HIV sequence database, 2017), and URF is complex inter-subtype recombinant genomes found only in one HIV infected subject. The emergence of new HIV variants (mainly CRF and URF) and the spread of HIV-1 non-B subtypes and recombinants in this pandemic pose a challenge for the accuracy and efficiency of diagnostic and treatment tools.

Since sequence variation in p24 can influence the phenotypic properties of HIV-1 and its interactions with host factors and ARVs, a deep knowledge of p24 variability across circulating variants and the identification of highly conserved HIV epitopes could help for a more rational design of p24-based diagnostic tests, ARV, and vaccines, which, ideally, should be directed to highly conserved p24 sites across all HIV-1 variants (groups, subtypes, sub-sub types, and recombinants) (Ferrari et al., 2000; Hatziioannou et al., 2004; Thenin-Houssier and Valente, 2016; Saito and Yamashita, 2021). This descriptive study on p24 diversity across secondary structure elements in a large p24 sequence set from different HIV-1 variants reports differences and variant-specific markers in p24 across HIV-1 groups (M, O, N, and P) and group M variants. We also report the proportion of substitutions per variant in p24 residues affecting the capsid assembly (Lingappa et al., 2014), HIV uncoating (Hulme et al., 2015), infectivity (Yamashita et al., 2007; Koh et al., 2013), molecule binding (Qi et al., 2008; Ylinen et al., 2009; McFadden et al., 2021; Saito and Yamashita, 2021), and the new CA-inhibitor lenacapavir key capsid-interaction sites and described resistance mutations (Bester et al., 2020; Link et al., 2020; Sun et al., 2021), according to the bibliography.

## MATERIALS AND METHODS

In May 2021, we downloaded all available p24 nucleotide sequences from the Los Alamos National Laboratory (2021c), selecting one sequence per patient. The sequences were organized according to their HIV-1 variant in groups, subtypes, sub-subtypes, and CRF. The URF sequences were not included in this study. They were aligned, edited, and translated into aa with the MEGA v6.0 program (Molecular Evolutionary Genetics Analysis[1]) (Tamura et al., 2013) using HIV-1 reference sequence HXB2 (NCBI accession number: K03455.1).

Sequence analysis was performed with an in-house bioinformatics tool (EpiMolBio), previously designed and used in our laboratory for HIV genetic variability analysis and recently updated for severe acute respiratory syndrome-Corona virus-2 (SARS-CoV-2) sequences study (Burgos et al., 2019; Troyano-Hernáez et al., 2020, 2021a,b). This tool is programmed in JAVA OpenJDK version 11.0.9.1 using IDE NetBeans version 12.2. Among other functions, this tool calculates the conservation of a sequence set compared to a reference sequence, as well as the percentage of aa changes for each position within the studied protein. Furthermore, it can infer a consensus from a group of sequences or previously calculated consensus considering the total number of sequences and the frequency of any specific aa residue per position, avoiding the overestimation of polymorphisms present in variants with a small number of available sequences.

We inferred the p24 aa consensus sequence for HIV-1, each HIV-1 group (M, N, O, and P), and each HIV-1 variant (subtype, sub-subtype, and CRF) using all downloaded sequences. Group M consensus was inferred from the consensus of group M subtypes, sub-subtypes, and CRF. The HIV-1 consensus was inferred considering the 4 groups (M, N, O, and P) consensus. The inferred p24 consensus sequences for HIV-1 and each HIV-1 group were used to analyze the mean percentage or level of conserved aa in each p24 residue and each region of its secondary structure. Since the MHR is highly conserved among retroviruses (Gamble et al., 1997), we inferred the HIV-2 p26 consensus sequences by using 182 p26 LANL sequences to analyze this region.

Two analyses were performed only in those variants with at least 10 available p24 sequences to avoid biases due to a low number of sequences. In the first place, we studied the p24 average aa conservation compared to the HXB2 HIV-1 reference sequence in these variants. Secondly, we identified the presence of single V-markers, defined as the natural aa changes specific for each variant and present in >75% of the sequence set for a given position.

We calculated the Wu–Kabat protein variability coefficient (WK) for group M using all available p24 sequences belonging to this group and analyzed the results in the context of the proteins' domains. The WK coefficient allows studying the susceptibility of an aa position to evolutionary replacements (Kabat et al., 1977). It is calculated using the following formula: Variability = $N \times k/n$, where $N$ is the number of sequences in the alignment, $k$ is the number of different amino acids at a given position, and $n$ is the absolute frequency of the most common amino acid at that position. Therefore, a WK

[1] https://www.megasoftware.net/

**TABLE 1 |** HIV-1 p24 Los-Alamos National Library (LANL) sequences analyzed in the present study.

| HIV-1 variants | | | N° sequences | HIV-1 variants | | | N° sequences |
|---|---|---|---|---|---|---|---|
| Non-M groups | | N | 11 | Group M | CRF | 47_BF | 5 |
| | | O | 104 | | | 48_01B | 6 |
| | | P | 4 | | | 49_cpx | 9 |
| Group M | Subtypes | A | 61 | | | 50_A1D | 5 |
| | | A1 | 1408 | | | 51_01B | 7 |
| | | A2 | 71 | | | 52_01B | 3 |
| | | A3 | 17 | | | 53_01B | 4 |
| | | A4 | 3 | | | 54_01B | 3 |
| | | A6 | 191 | | | 55_01B | 22 |
| | | B | 9908 | | | 56_cpx | 4 |
| | | C | 4823 | | | 57_BC | 13 |
| | | D | 597 | | | 58_01B | 6 |
| | | F | 30 | | | 59_01B | 9 |
| | | F1 | 234 | | | 60_BC | 8 |
| | | F2 | 31 | | | 61_BC | 4 |
| | | G | 176 | | | 62_BC | 3 |
| | | H | 15 | | | 63_02A | 13 |
| | | J | 7 | | | 64_BC | 9 |
| | | K | 4 | | | 65_cpx | 17 |
| | CRF | 01_AE | 3782 | | | 66_BF1 | 3 |
| | | 02_AG | 539 | | | 67_01B | 4 |
| | | 03_AB | 7 | | | 68_01B | 3 |
| | | 04_cpx | 12 | | | 69_01B | 7 |
| | | 05_DF | 7 | | | 70_BF1 | 5 |
| | | 06_cpx | 42 | | | 71_BF1 | 16 |
| | | 07_BC | 621 | | | 72_BF1 | 6 |
| | | 08_BC | 296 | | | 73_BG | 2 |
| | | 09_cpx | 11 | | | 74_01B | 6 |
| | | 10_CD | 3 | | | 75_BF | 3 |
| | | 11_cpx | 29 | | | 76_01B | 2 |
| | | 12_BF | 25 | | | 77_cpx | 4 |
| | | 13_cpx | 12 | | | 78_cpx | 3 |
| | | 14_BG | 15 | | | 79_0107 | 4 |
| | | 15_01B | 8 | | | 80_0107 | 3 |
| | | 16_A2D | 4 | | | 81_cpx | 2 |
| | | 17_BF | 7 | | | 82_cpx | 6 |
| | | 18_cpx | 8 | | | 83_cpx | 11 |
| | | 19_cpx | 5 | | | 84_A1D | 3 |
| | | 20_BG | 4 | | | 85_BC | 12 |
| | | 21_A2D | 3 | | | 86_BC | 3 |
| | | 22_01A1 | 22 | | | 87_cpx | 4 |
| | | 23_BG | 2 | | | 88_BC | 3 |
| | | 24_BG | 11 | | | 89_BF | 3 |
| | | 25_cpx | 5 | | | 90_BF1 | 11 |
| | | 26_A5U | 4 | | | 92_C2U | 5 |
| | | 27_cpx | 5 | | | 93_cpx | 3 |
| | | 28_BF | 5 | | | 95_02B | 5 |
| | | 29_BF | 8 | | | 96_cpx | 4 |
| | | 31_BC | 3 | | | 98_06B | 2 |
| | | 32_06A6 | 5 | | | 99_BF | 3 |
| | | 33_01B | 18 | | | 100_01C | 3 |
| | | 34_01B | 3 | | | 101_01B | 3 |

*(Continued)*

**TABLE 1** | (Continued)

| HIV-1 variants | N° sequences | HIV-1 variants | N° sequences |
|---|---|---|---|
| 35_AD | 23 | 102_0107 | 2 |
| 36_cpx | 4 | 103_01B | 4 |
| 37_cpx | 5 | 104_0107 | 3 |
| 38_BF | 5 | 105_0108 | 5 |
| 39_BF | 3 | 106_cpx | 6 |
| 40_BF | 4 | 107_01B | 6 |
| 41_CD | 3 | 108_BC | 5 |
| 42_BF | 17 | 109_0107 | 2 |
| 43_02G | 7 | 112_0107 | 3 |
| 44_BF | 3 | 113_01B | 4 |
| 45_cpx | 10 | 114_01B | 3 |
| 46_BF | 8 | 115_01C | 3 |

*In red, variants with <10 sequences. N°, number; CRFs, circulating recombinant forms (Los Alamos National Laboratory, 2021c).*

of 1 indicates the same aa was found for that position in all the sequence sets, whereas a WK of >1 indicates relative variability of the respective site, with greater diversity as the WK value increases.

Finally, we identified the substitutions affecting the capsid assembly (Lingappa et al., 2014), HIV uncoating (Hulme et al., 2015), infectivity (Yamashita et al., 2007; Koh et al., 2013), molecule binding (Qi et al., 2008; Ylinen et al., 2009; McFadden et al., 2021; Saito and Yamashita, 2021), and lenacapavir key capsid-interaction sites and described resistance-associated mutations (Bester et al., 2020; Link et al., 2020; Sun et al., 2021) according to the bibliography.

# RESULTS

## p24 Sequences Analyzed and Inferred Consensus Sequences

After discarding incomplete sequences, 23,671 HIV-1 p24 LANL sequences were included in this study (**Table 1**). Among the HIV-1 p24 sequences, 119 belonged to non-M groups (N, O, and P) and 23,552 were ascribed to group M, including 9 subtypes, 7 sub-sub types, and 109 CRF. The p24 sequences from 9 HIV-1 variants were not available in LANL: CRF30_0206, CRF91_cpx, CRF94_cpx, CRF97_01B, CRF110_BC, CRF111_01C, CRF116_0108, CRF117_0107, and CRF118_BC. We inferred the consensus for each group M variant using all the available p24 LANL sequences for each variant (**Table 1**). Group M consensus was inferred using the 125 variants consensus. Consensus sequences for groups N, O, and P were generated with the 11, 104, and 4 available p24 LANL sequences for these groups. The HIV-1 p24 consensus sequence was generated using the four HIV-1 groups (M, N, O, and P) consensus. The consensus sequences for group M variants, HIV-1 groups, and HIV-1 inferred with EpiMolBio were aligned (**Supplementary Table 1**), indicating the most prevalent aa and its percentage of conservation for each p24 residue with a color code.

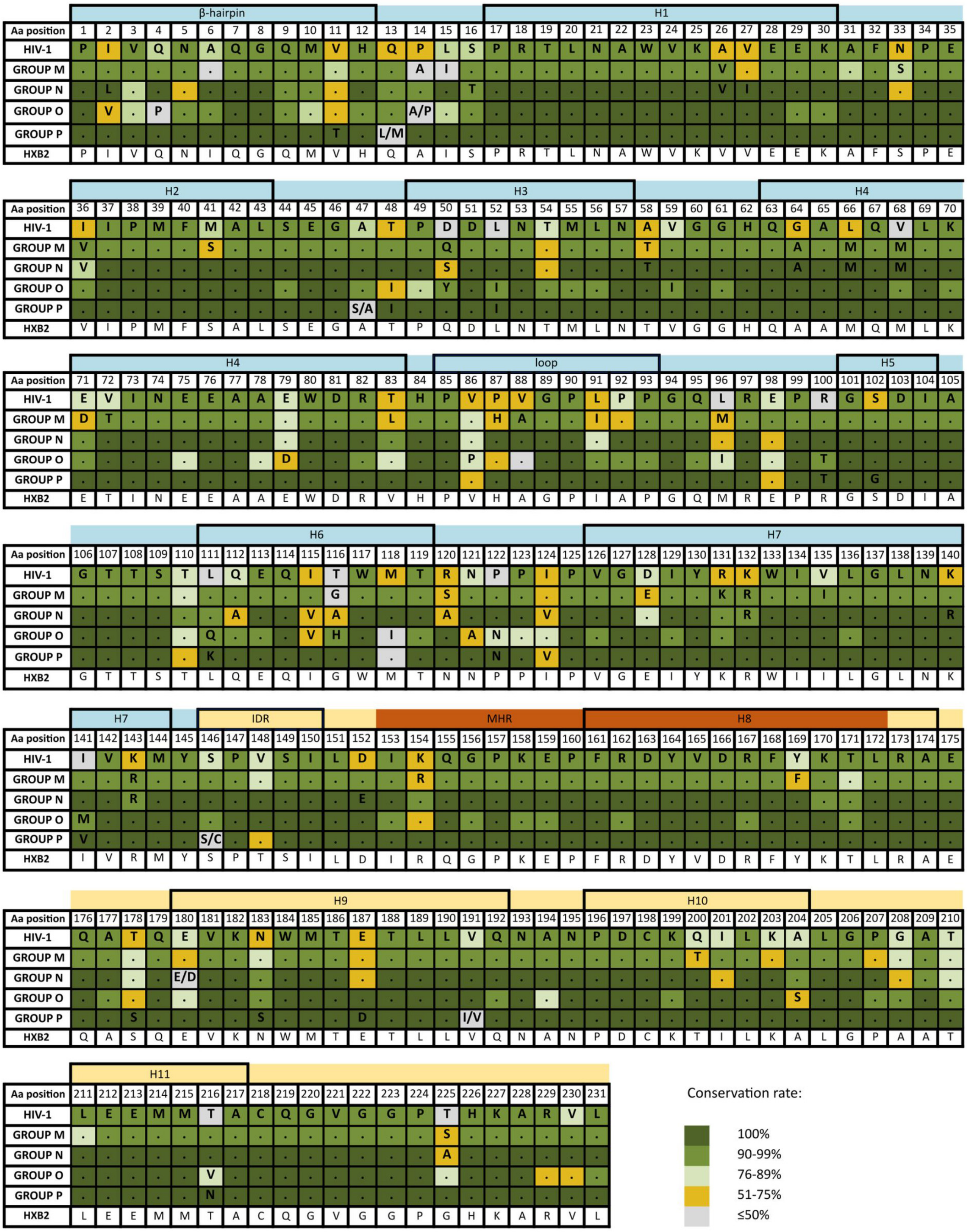
## p24 aa Conservation in Its Secondary Structure and Main Regions

The inferred p24 consensus sequences for HIV-1 and each HIV-1 group were used to analyze the aa variability of each residue in the p24 secondary structure domains, as illustrated in **Figure 1**. The HXB2 reference sequence was included for further guidance. The mean p24 aa conservation for the HIV-1 consensus was 89%. For groups M, N, O, and P, it was 94, 96, 95, and 98%, respectively. The mean aa conservation percentage of p24 compared to HXB2 reference sequence across HIV-1 groups was 93.6% (group M), 80.2% (group P), 79.1% (group O), and 85.4% (group N), respectively.

The aa conservation in the secondary p24 structure varied between HIV-1 variants, amino acid residues, and type of secondary structure. None of the structural regions presented 100% conservation across variants. In HIV-1 consensus, the most conserved structure was helix 8 (98%), followed by the MHR (97%). Both regions overlap at the beginning of p24 CTD (**Figure 2**). The least conserved region was helix 6 (76%), followed by the Cyp A-binding loop (83%). As for the HIV-1 groups (**Tables 2**, **3**), group P showed the lowest conservation vs. other groups in the IDR (83.3 vs. 94–100%), while group O showed the lowest conservation of the β-hairpin (84.4 vs. 90–100%). Both groups M and O presented lower conservation of the Cyp A-binding loop than groups N and P (86.3 and 86.4 vs. 95.8 and 97.2%). Overall, the Cyp A-binding loop and the β-hairpin were less conserved regions (**Figure 1**).

The Cyp A-binding loop presented 5 sites sharing the same aa residue in all the consensus: proline in sites 85, 90, 92, and 93, and glycine in site 89 (**Figure 3A**). The most variable residues were sites 86–88 and 91, showing polymorphisms in group O (V86P) and group M (P87H, V88A, and L91I) compared to HIV-1 consensus alignment. The most variable Cyp A-binding loop sites in group M consensus sequence were 87, 91, and 92 (61.5–69% conservation), with site 91 being the most polymorphic across the 125 group M variants (**Supplementary Table 1**).

The MHR (sites 153–172) was also highly conserved in the four HIV-1 groups (95–100% conservation) (**Figure 3B** and **Table 2**), with 18/20 of its sites showing >90% conservation in



| | β-hairpin | | | | | | | | | | | | | | | | H1 | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
| **HIV-1** | P | I | V | Q | N | A | Q | G | Q | M | V | H | Q | P | L | S | P | R | T | L | N | A | W | V | K | A | V | E | E | K | A | F | N | P | E |
| **GROUP M** | . | . | . | . | . | . | . | . | . | . | . | . | . | A | I | . | . | . | . | . | . | . | . | . | . | V | . | . | . | . | . | . | S | . | . |
| **GROUP N** | . | L | . | . | . | . | . | . | . | . | . | . | . | . | . | T | . | . | . | . | . | . | . | . | . | V | I | . | . | . | . | . | . | . | . |
| **GROUP O** | . | V | . | P | . | . | . | . | . | . | . | . | . | A/P | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP P** | . | . | . | . | . | . | . | . | . | . | T | . | L/M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **HXB2** | P | I | V | Q | N | I | Q | G | Q | M | V | H | Q | A | I | S | P | R | T | L | N | A | W | V | K | V | V | E | E | K | A | F | S | P | E |

| | H2 | | | | | | | | | | | | | H3 | | | | | | | | | | | | | | H4 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 |
| **HIV-1** | I | I | P | M | F | M | A | L | S | E | G | A | T | P | D | D | L | N | T | M | L | N | A | V | G | G | H | Q | G | A | L | Q | V | L | K |
| **GROUP M** | V | . | . | . | . | S | . | . | . | . | . | . | . | . | Q | . | . | . | . | . | . | . | T | . | . | . | . | . | A | . | M | . | M | . | . |
| **GROUP N** | V | . | . | . | . | . | . | . | . | . | . | . | . | . | S | . | . | . | . | . | . | . | T | . | . | . | . | . | A | . | M | . | M | . | . |
| **GROUP O** | . | . | . | . | . | . | . | . | . | . | . | . | I | . | Y | . | I | . | . | . | . | . | . | I | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP P** | . | . | . | . | . | . | . | . | . | . | . | S/A | I | . | . | . | I | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **HXB2** | V | I | P | M | F | S | A | L | S | E | G | A | T | P | Q | D | L | N | T | M | L | N | T | V | G | G | H | Q | A | A | M | Q | M | L | K |

| | H4 | | | | | | | | | | | | | | loop | | | | | | | | | | | | | | | | H5 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 |
| **HIV-1** | E | V | I | N | E | E | A | A | E | W | D | R | T | H | P | V | P | V | G | P | L | P | P | G | Q | L | R | E | P | R | G | S | D | I | A |
| **GROUP M** | D | T | . | . | . | . | . | . | . | . | . | . | L | . | . | . | H | A | . | . | I | . | . | . | . | M | . | . | . | . | . | . | . | . | . |
| **GROUP N** | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP O** | . | . | . | . | . | . | . | . | D | . | . | . | . | . | . | P | . | . | . | . | . | . | . | . | . | I | . | . | . | T | . | . | . | . | . |
| **GROUP P** | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | T | . | G | . | . | . |
| **HXB2** | E | T | I | N | E | E | A | A | E | W | D | R | V | H | P | V | H | A | G | P | I | A | P | G | Q | M | R | E | P | R | G | S | D | I | A |

| | | | | | | H6 | | | | | | | | | | | | | | | H7 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 106 | 107 | 108 | 109 | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 |
| **HIV-1** | G | T | T | S | T | L | Q | E | Q | I | T | W | M | T | R | N | P | P | I | P | V | G | D | I | Y | R | K | W | I | V | L | G | L | N | K |
| **GROUP M** | . | . | . | . | . | . | . | . | . | . | G | . | . | . | S | . | . | . | . | . | . | . | E | . | . | K | R | . | . | I | . | . | . | . | . |
| **GROUP N** | . | . | . | . | . | . | A | . | . | V | A | . | . | . | A | . | . | . | V | . | . | . | . | . | . | . | R | . | . | . | . | . | . | . | R |
| **GROUP O** | . | . | . | . | . | Q | . | . | . | V | H | . | I | . | . | A | N | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP P** | . | . | . | . | . | K | . | . | . | . | . | . | . | . | . | . | N | . | V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **HXB2** | G | T | T | S | T | L | Q | E | Q | I | G | W | M | T | N | N | P | P | I | P | V | G | E | I | Y | K | R | W | I | I | L | G | L | N | K |

| | H7 | | | | | IDR | | | | | | | MHR | | | | | | | | H8 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| **HIV-1** | I | V | K | M | Y | S | P | V | S | I | L | D | I | K | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | Y | K | T | L | R | A | E |
| **GROUP M** | . | . | R | . | . | . | . | . | . | . | . | . | . | R | . | . | . | . | . | . | . | . | . | . | . | . | . | . | F | . | . | . | . | . | . |
| **GROUP N** | . | . | R | . | . | . | . | . | . | . | . | E | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP O** | M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP P** | V | . | . | . | . | S/C | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **HXB2** | I | V | R | M | Y | S | P | T | S | I | L | D | I | R | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | Y | K | T | L | R | A | E |

| | | | | | H9 | | | | | | | | | | | | | | | | H10 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 | 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 | 208 | 209 | 210 |
| **HIV-1** | Q | A | T | Q | E | V | K | N | W | M | T | E | T | L | L | V | Q | N | A | N | P | D | C | K | Q | I | L | K | A | L | G | P | G | A | T |
| **GROUP M** | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | T | . | . | . | . | . | . | . | . | . | . |
| **GROUP N** | . | . | . | . | E/D | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP O** | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | S | . | . | . | . | . | . |
| **GROUP P** | . | . | S | . | . | . | . | S | . | . | . | D | . | . | . | I/V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **HXB2** | Q | A | S | Q | E | V | K | N | W | M | T | E | T | L | L | V | Q | N | A | N | P | D | C | K | T | I | L | K | A | L | G | P | A | A | T |

| | H11 | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aa position** | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 | 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 |
| **HIV-1** | L | E | E | M | M | T | A | C | Q | G | V | G | G | P | T | H | K | A | R | V | L |
| **GROUP M** | . | . | . | . | . | . | . | . | . | . | . | . | . | . | S | . | . | . | . | . | . |
| **GROUP N** | . | . | . | . | . | . | . | . | . | . | . | . | . | . | A | . | . | . | . | . | . |
| **GROUP O** | . | . | . | . | . | V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **GROUP P** | . | . | . | . | . | N | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| **HXB2** | L | E | E | M | M | T | A | C | Q | G | V | G | G | P | G | H | K | A | R | V | L |

Conservation rate:

| | |
|---|---|
| ■ | 100% |
| ■ | 90-99% |
| ■ | 76-89% |
| ■ | 51-75% |
| ■ | ≤50% |

**FIGURE 1 |** Amino acid conservation percentage along p24 secondary structure in HIV-1 and each HIV-1 group. Aa, amino acid. The aa for each position in groups M, N, O, and P are in reference to HIV-1 consensus. Dots represent the same aa as HIV-1 consensus for that position. The HXB2 reference sequence is described for further guidance below the groups. Colors represent the conservation percentage. P24 domains and secondary structure regions: H, helix; IDR, interdomain linker region; MHR, major homology region; in light blue, N-terminal domain; in light yellow, C-terminal domain; in dark orange, MHR. Aa code: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine. Residues P1 and H12 in the β-hairpin and R18 and D51 in the α-helixes 1 (H1) and 3 (H3) are involved in the hexameric pore function.

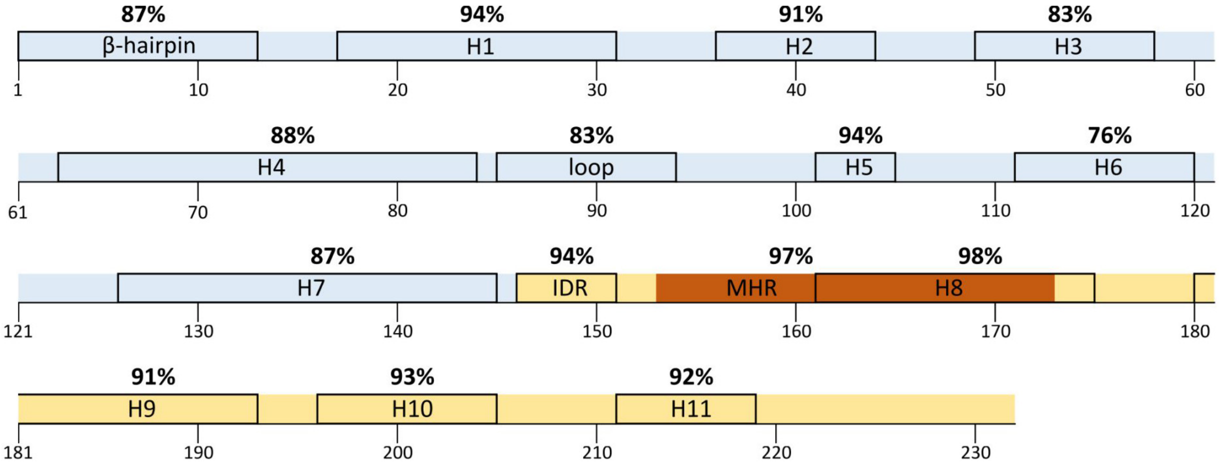


**FIGURE 2 |** HIV-1 consensus mean conservation of each element in p24 secondary structure. H, helix; IDR, interdomain linker region; MHR, major homology region; in light blue, N-terminal domain; in light yellow, C-terminal domain. The boxes indicate the residues included in each structure except for the MHR, indicated in dark orange.

**TABLE 2 |** Conservation of each element of p24 secondary structure in HIV-1, the four HIV-1 groups, and group M subtype B consensus.

| p24 structure | aa sites | HIV-1 | Group M | Group N | Group O | Group P | Group M subtype B |
|---|---|---|---|---|---|---|---|
| β-Hairpin | 1–12 | 86.7 | 93.0 | 90.9 | 84.4 | 100.0 | 96.9 |
| H1 | 17–30 | 93.8 | 96.5 | 100.0 | 99.8 | 100.0 | 97.9 |
| H2 | 36–43 | 91.2 | 96.4 | 97.7 | 99.8 | 100.0 | 96.4 |
| H3 | 49–57 | 83.2 | 95.2 | 91.9 | 95.7 | 100.0 | 99.3 |
| H4 | 63–83 | 87.8 | 95.0 | 98.3 | 95.4 | 100.0 | 98.5 |
| Cyp A loop | 85–93 | 82.8 | 86.3 | 95.8 | 86.4 | 97.2 | 90.8 |
| H5 | 101–104 | 93.6 | 99.7 | 100.0 | 100.0 | 100.0 | 99.8 |
| H6 | 111–119 | 75.9 | 91.9 | 89.6 | 86.2 | 94.4 | 95.1 |
| H7 | 126–144 | 87.1 | 96.8 | 97.6 | 99.3 | 100.0 | 98.1 |
| IDR | 146–150 | 94.3 | 94.1 | 100.0 | 99.8 | 83.3 | 91.6 |
| MHR | 153–172 | 97.1 | 95.5 | 98.6 | 97.2 | 100.0 | 97.9 |
| H8 | 161–174 | 97.9 | 96.5 | 98.7 | 99.4 | 100.0 | 99.2 |
| H9 | 180–192 | 90.9 | 91.0 | 93.4 | 98.7 | 96.1 | 96.5 |
| H10 | 196–204 | 92.8 | 90.2 | 94.5 | 99.8 | 100.0 | 98.9 |
| H11 | 211–217 | 92.3 | 97.6 | 100.0 | 96.8 | 100.0 | 99.4 |

*H, helix; IDR, interdomain linker region; MHR, major homology region; Cyp A loop, Cyp A-binding loop; aa, amino acid.*

the HIV-1 consensus and many sites with complete conservation in groups P (20/20 sites), N (17/20 sites), and O (12/20) (**Supplementary Table 1** and **Figure 1**). The most variable sites in the MHR, with conservation below 75%, were site 154 in groups O and M and site 169 in group M. However, only group M's aa differed from the general HIV-1 consensus (K154R, and Y169F). Regardless of the level of aa conservation, all the other sites presented the same aa in the consensus sequences (**Figure 3B**). Sites Q155, G156, E159, and R167, previously described as highly conserved in all retroviruses (Gamble et al., 1997), had ≥99% conservation in all the consensuses.

The main four p24 residues involved in the hexameric pore function (P1 and H12 in the β-hairpin and R18 and D51 in the α-helixes 1 and 3) were highly conserved in HIV-1 consensus (>98%), presenting the same aa in the HIV-1 groups (**Figure 1**). Groups N and P had complete conservation of these 4 sites. Group O consensus showed 100% conservation in all, except H12 (96% conservation). Group M had conservation above 99% in P1, R18, and D51, whereas H12 had conservation of 97%. In a deeper analysis, the most prevalent mutation in site 12 was tyrosine instead of histidine, found in 0.03% of the group M consensus (683 sequences) and 4 out of 104 group O sequences.

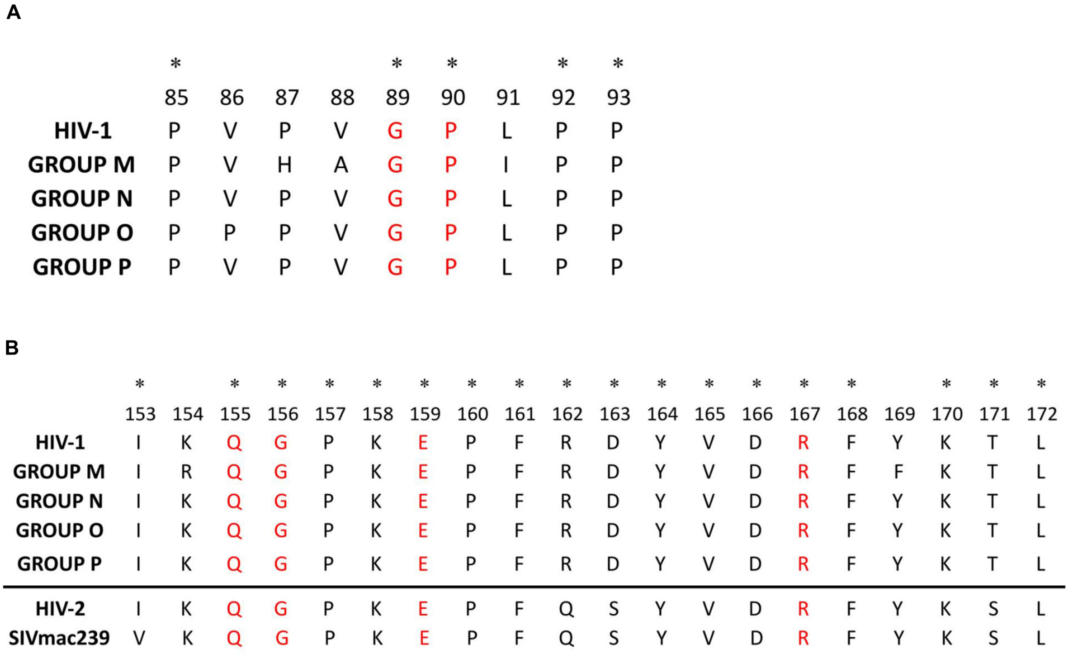
## Capsid aa Conservation Across HIV-1 Variants and V-Markers

To analyze the p24 conservation across each group M variant and identify the specific V-markers, we used the 38 variants with at least 10 p24 available LANL sequences, corresponding to 7 subtypes (A-H), 6 sub-sub types (A1–A3, A6, F1–F2), and 25 CRF. The F1 was the least conserved subtype

**TABLE 3 |** Range of conservation of each element of p24 secondary structure in the four HIV-1 groups and group M subtype B consensus.

| Conservation (%) | Group M | Group N | Group O | Group P | HXB2 subtype B |
|---|---|---|---|---|---|
| 100 | | IDR, H1, H11, H5 | H5 | β-Hairpin, H4, H3, H11, MHR, H7, H8, H2, H10, H1, H5 | |
| 98–99.9 | H5 | H4, MHR, H8 | H9, H7, H8, H2, H10, H1, IDR | | H5, H11, H3, H8, H10, H4, H7 |
| 96–97.9 | H11, H7, H8, H1, H2 | H7, H2 | H11, MHR | H9, Loop | MHR, H1, β-hairpin, H9, H2 |
| 94–95.9 | MHR, H3, H4, IDR | H10, Loop | H4, H3 | H6 | H6 |
| 92–93.9 | β-Hairpin | H9 | | | |
| 90–91.9 | H10, H9, H6 | β-Hairpin, H3 | | | IDR, Loop |
| 88–89.9 | | H6 | | | |
| 86–87.9 | Loop | | H6, Loop | | |
| 84–85.9 | | | β-Hairpin | | |
| 82–83.9 | | | | IDR | |

*H, helix; IDR, interdomain linker region; MHR, major homology region; Loop, Cyp A-binding loop.*



A

| | * | | | | * | * | | * | * |
|---|---|---|---|---|---|---|---|---|---|
| | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 |
| **HIV-1** | P | V | P | V | G | P | L | P | P |
| **GROUP M** | P | V | H | A | G | P | I | P | P |
| **GROUP N** | P | V | P | V | G | P | L | P | P |
| **GROUP O** | P | P | P | V | G | P | L | P | P |
| **GROUP P** | P | V | P | V | G | P | L | P | P |

B

| | * | | * | * | * | * | * | * | * | * | * | * | * | * | * | * | | * | * | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 |
| **HIV-1** | I | K | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | Y | K | T | L |
| **GROUP M** | I | R | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | F | K | T | L |
| **GROUP N** | I | K | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | Y | K | T | L |
| **GROUP O** | I | K | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | Y | K | T | L |
| **GROUP P** | I | K | Q | G | P | K | E | P | F | R | D | Y | V | D | R | F | Y | K | T | L |
| **HIV-2** | I | K | Q | G | P | K | E | P | F | Q | S | Y | V | D | R | F | Y | K | S | L |
| **SIVmac239** | V | K | Q | G | P | K | E | P | F | Q | S | Y | V | D | R | F | Y | K | S | L |

**FIGURE 3 |** HIV-1 and HIV-1 groups alignment of p24 Cyp A-binding loop **(A)** and major homology region **(B)**. **(A)** Cyp A-binding loop aa alignment (aa 85–93) considering the HIV-1 and the 4 HIV-1 groups consensus sequences. In red, main sites involved in the Cyp A-binding loop interaction; with an asterisk, sites with the same amino acid in all the consensus sequences. **(B)** Mayor homology region aa alignment (aa 153–172) considering the HIV-1 and the 4 HIV-1 groups consensus sequences. The HIV-2 p26 MHR consensus inferred from LANL sequences and SIVmac239 MHR sequence downloaded from LANL have been added below the line for further information. In red, invariable amino acids in retroviruses MHR; with an asterisk, sites with the same amino acid in all the consensus sequences.

(69.3%), and CRF85_BC was the least conserved CRF (82.2%) (**Figure 4**). All the other subtypes and CRF showed a p24 conservation above 89%.

We identified the variant-specific single V-markers present in >75% of the p24 sequences for a given position. Among the 38 HIV-1 group M variants with ≥10 sequences, 9 variants carried 14 single V-markers. **Figure 5** shows the 14 V-markers location in p24, the variant they belong to, and their conservation percentage. The legend describes the main residue found in the reference sequence in the corresponding sites.

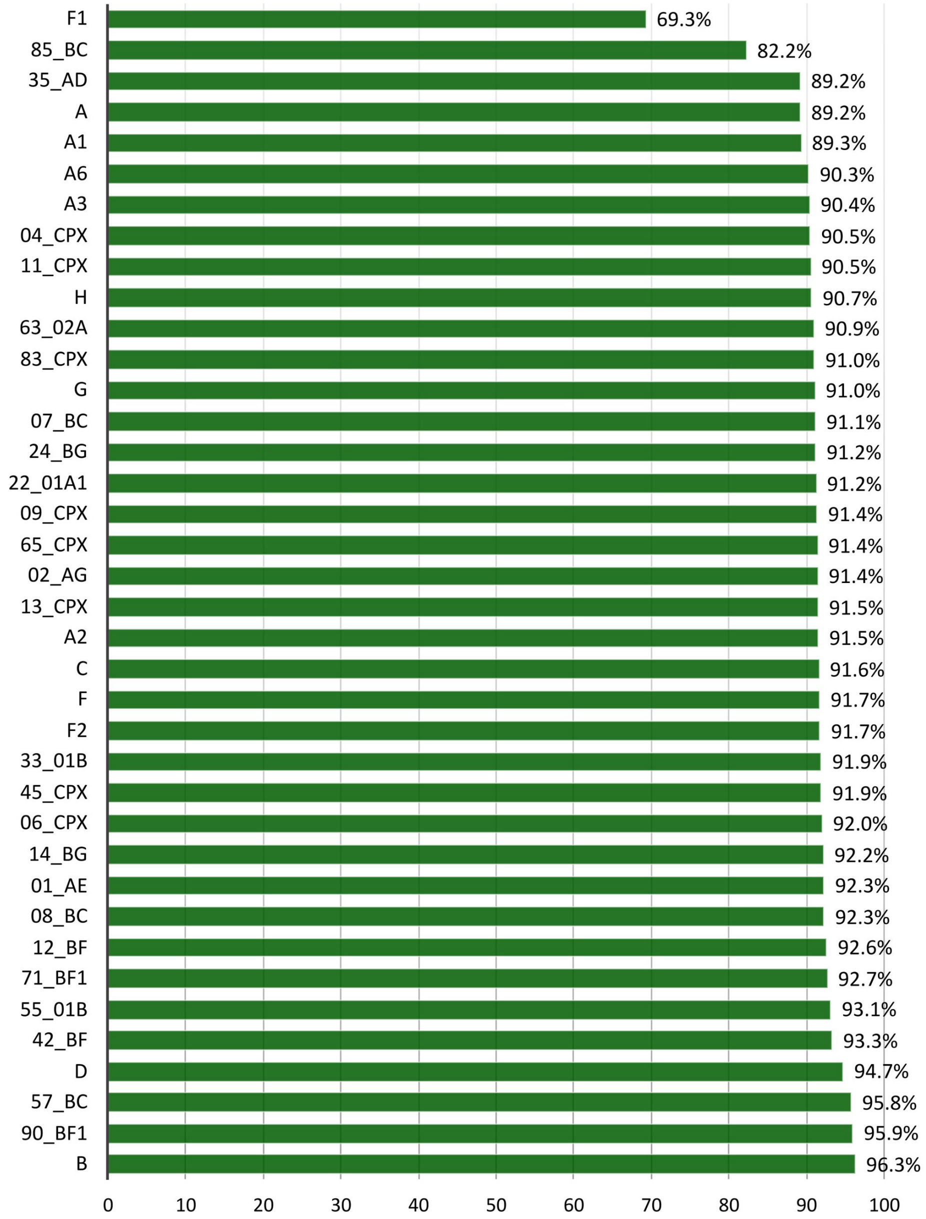


**FIGURE 4 |** Percentage of aa conservation of p24 across the 38 HIV-1 group M variants with ≥10 p24 sequences at LANL. *X*-axis: HIV-1 group M variants with at least 10 available sequences. *Y*-axis: mean p24 conservation percentage for each variant included in this analysis.

Subtype F (30 sequences) had two V-markers, Y12 and T14, located in the NTD, both present in all 30 sequences. Sub-subtypes A3 (17 sequences) and A6 (191 sequences) presented one V-marker each: C171 (94.1%) in A3 located in the MHR and the only V-marker in p24 CTD, and F91 (82.7%) in A6 located in the Cyp A-binding loop. Six CRF presented V-markers: CRF63_02A (13 sequences) presented three markers in the NTD: T11 (92.3%), M15 (84.6%), and V135 (100%). The CRF24_BG and CRF83_cpx, both with 11 sequences, presented two V-markers each. In CRF24_BG, we found the V-markers P10 (100%) and I11 (81.8%) located in the NTD β-hairpin. In CRF83_cpx, the V-markers were H91 (90.9%), located in the Cyp A-binding loop, and T116 (100%), located in helix 6, both in p24 NTD. The other three CRF presented one V-marker each. The CRF42_BF (17 sequences) had A44 (100%), CRF65_cpx (17 sequences) had M136 (94.1%), and CRF90_BF1 (11 sequences) had H120 (81.8%) all located in the NTD (**Figure 5**).

## Wu–Kabat p24 Variability Coefficient in Group M

The median variability coefficient along p24 in group M sequences was 10, being 9 in the NTD and 10.3 in the CTD. The P24 site 116 in H6 presented the maximum coefficient (WK

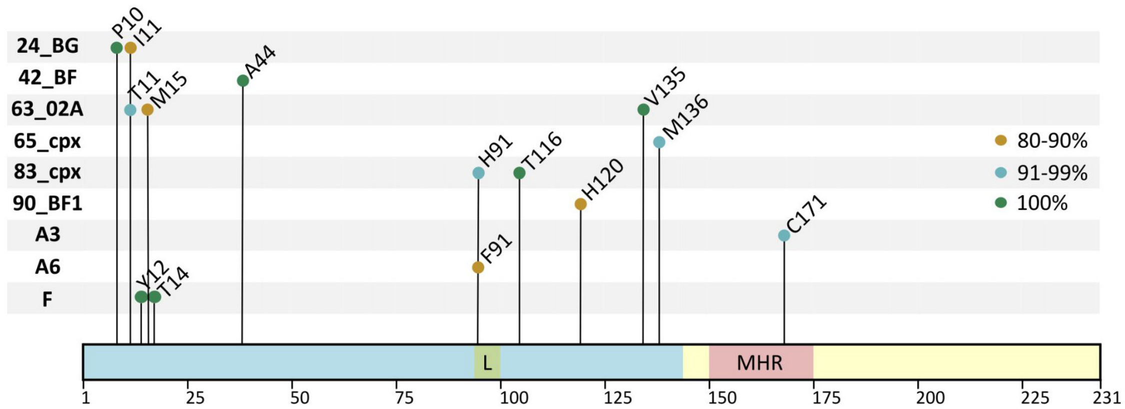


**FIGURE 5 |** Single V-markers across HIV-1 group M variants. *X*-axis, p24 positions and relevant domains; *Y*-axis, HIV-1 variants carrying single V-markers. In light blue, N-terminal domain, including the Cyp A-binding loop (L) in green. In light yellow, C-terminal domain, including the major homology region (MHR) in red; the conservation percentage of each V-marker is represented in the figure with colored circles (in green, 100%; in blue, 91–99%; in orange, 80–90%). In HXB2 HIV-1 reference sequence the residues present in each site were: M10, V11, H12, A14, I15, S44, I91, G116, N120, I135, L136, and T171. In group M consensus the same residues were present except for S120.

31.8), followed by site 120 in NTD after H6 (WK 31.4), site 6 in β-hairpin (WK 30), site 225 in CTD end (WK 28.9), and site 15 in NTD before H1 (WK 27.8) (**Figure 6**). The smallest WK coefficient was 5 found in sites 46 (NTD after H2), 60 (NTD after H3), 64 (H4), 97 (NTD after the Cyp A-binding loop), 113 (H6), 145 (last NTD aa), and 155 (MHR). The median WK in the Cyp A-binding loop was 8 (mean WK 11), with the larger variability in site 91 (WK 22), followed by sites 92, 86, and 87 (WK 14–11). The median variability coefficient of the MHR was the same as in the Cyp A-binding loop, WK 8 (mean WK 9.4), with 22 of its 27 sites (81%) below this value. The most variable sites in MHR were 148 (WK 19.5), followed by 154 (WK 18), and the most conserved was 155 (WK 5). **Supplementary Table 2** describes the WK values for each of the 231 p24 residues after the alignment of 23,552 p24 LANL sequences ascribed to HIV-1 group M variants (9 subtypes, 7 sub-sub types, and 109 CRF) under study.

## Capsid Polymorphisms Affecting HIV Phenotypic Properties

Since p24 genetic variability can impact on viral replication cycle at different stages, we analyzed the conservation percentage in these p24 positions affecting the capsid assembly (Lingappa et al., 2014), HIV uncoating (Hulme et al., 2015), infectivity (Yamashita et al., 2007; Koh et al., 2013), and molecule binding (Qi et al., 2008; Ylinen et al., 2009; Saito and Yamashita, 2021).

For the proper assembly of the immature capsid, the functional CA protein must present the residues V181/K182, W184/M185, and L189/L190 in the α-helix 9 (H9) that forms the CTD-end dimer interface between capsid hexamers. The presence of double mutants at these positions alters the hydrophobic bonds that stabilize the interaction of the dimer interface with the homologous partner (Lingappa et al., 2014). In our p24 sequence set, we found no double mutants within H9 in groups N, O, or P. In group M, we found double mutants in some variants (**Supplementary Table 3**), with an overall percentage below 0.01% for this group.

At the beginning of the CTD, there are 3 aa in non-contiguous regions close to each other in the tertiary structure: K158 (MHR), D197 (H10), and P224, where aa changes can affect Gag multimerization in the membrane. The three residues were well conserved in the O, P, and N groups, and only one group O p24 sequence (0.9%) harbored the aa change K158E. In group M, mutants were found in all 3 positions, again at very low frequency (**Supplementary Table 3**).

Double mutations at positions E75/E76 in H4, R100/S102 in H5, and T107/T108 (between H5 and H6) and T110/Q112 in H6 inhibit the final step of assembly, leading to the accumulation of pathway intermediates with no virion exit (Lingappa et al., 2014). In group M, double mutants were found in all these positions (**Supplementary Table 3**), being rare in positions 75–76, which is highly conserved. No changes were found in groups N, O, and P in positions 75/76 and 107/108. Threonine in site 100 was the predominant residue for both groups O and P (**Figure 1**). Double mutants were only found in group P for these sites, where T100-G102 was present in all group's P available p24 sequences. Positions 110–112 presented double mutants in all the groups. Alanine is the main aa in site 112 in group N consensus, but only one N sequence presented double mutants in these sites. Site 110 was one of the least conserved sites involving the capsid assembly, and most double mutants had asparagine (N) in this site. However, 110–112 double mutants accounted for less than 1% of the total studied sequences.

Mutants associated with alteration of p24 molecule-binding and HIV-1 infectivity (Saito and Yamashita, 2021) were very infrequent (**Supplementary Table 3**). One exception was H87Q found in all group M subtypes but J, all sub-sub types but A4 and A5, and in 62 CRF. This aa change was the main polymorphism in 22 variants, being present in 100% of the sequences of CRFs 03_AB, 14_BG, 20_BG, 23_BG, 24_BG, 25_CPX, 27_cpx, 41_CD, 43_02G, 61_BC, 68_01B, 73_BG, 76_01B, 83_cpx, and

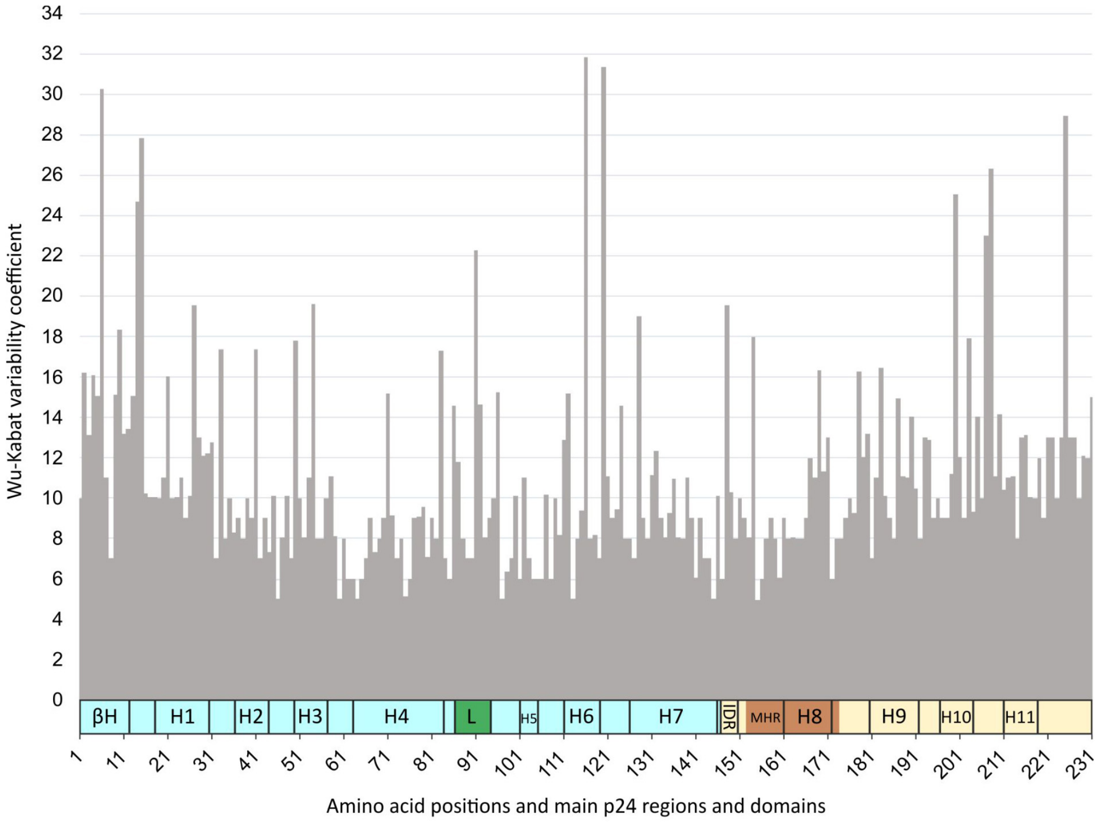


**FIGURE 6 |** Wu–Kabat (WK) p24 variability coefficient plot in HIV-1 group M sequences. *X*-axis, amino acid positions and main p24 regions and domains; *Y*-axis, Wu–Kabat variability coefficient. βH, β-hairpin; H, α-helix; L, Cyp A-binding loop (green); IDR, interdomain linker region; MHR, major homology region (dark orange); in light blue, N-terminal domain; in light yellow, C-terminal domain.

84_A1D, in sub-subtype A6, and subtype F. The H87Q was also present in ≥80% of p24 sequences from CRF13_cpx, 32_06A6, 60_BC, 82_cpx, and subtype G. It also appeared in 4% of group O sequences.

The other exception was the change G116A found in all groups but P, being present in 73% of group N sequences and in all group M subtypes except for F, in all sub-subtypes except for A4, and in 50 CRF. The G116A appeared in all (100%) p24 sequences from CRFs 21_A2D, 36_cpx, 60_BC, 62_BC, 64_BC, 65_cpx, 68_01B, 77_cpx, 86_BC, 87_cpx, 92_C2U, 93_cpx, 96_cpx, 103_01B, 105_0108, 106_cpx, 108_BC, 109_0107, and 113_01B. It was also found in ≥80% of the sequences in variants CRF05_DF, 07_BC, 19_CPX, 35_AD, 85_BC, sub-subtype F2, and subtypes C and H.

The key interactions between lenacapavir and the capsid take place in residues N57 (H3), K70 and N74 (H4), and N183 (H9) (Link et al., 2020). The first three sites were highly conserved (>99%) in the HIV-1 and groups M, N, O, and P consensus. However, site 183 showed conservation of 65.5% in the HIV-1 consensus, mainly due to residue variability in groups M and P consensus. Asparagine was the main residue found in N (90%) and O groups (94%), being less conserved in M (77%) and replaced by serine in all (100%) group P sequences. Within the group M variants, glycine was the most prevalent residue in site 183 in subtype F (100%), sub-sub types A1 (60%), F1 (68%), and F2 (69%), and in 18 CRF (**Supplementary Table 1**). In the other four CRF (36_cpx, 66_BF1, 73_BG, and 89_BF), there was not one consensus aa for this position due to the low number of sequences available in LANL (**Table 1**), sharing asparagine with glycine, serine, or alanine (**Supplementary Table 1**). The CRF113_01B had histidine in site 183 in all (100%) of its sequences.

Mutations of Q67H, L56I, M66I, K70N, and N74D/S associated with lenacapavir resistance (Bester et al., 2020; Link et al., 2020; Sun et al., 2021) were rare, found in <0.5% of the sequences in subtypes B, C, D, sub-subtype A1, CRFs 01_AE, 02_AG, and 07_BC (**Supplementary Table 3**). The exception was a change in N74S, present in 10% of CRF45_cpx sequences with only 10 available sequences in LANL. Double mutants, such as Q67H/N74D and Q67H/T107H, were not found in any of the sequences in this study. None of these lenacapavir-associated resistance mutations were found in groups N, O, or P.

## DISCUSSION

Global geographical patterns in HIV-1 variant distribution are changing due to several factors, including population movements, contributing to an unpredictable HIV-1 pandemic

(Peeters and Sharp, 2000). The HIV-1 variants have different global prevalence (Hemelaar et al., 2019) and can present distinct levels of HIV-1 genetic diversity (Abecasis et al., 2009). The HIV-1 group M subtype C is the most prevalent HIV variant in this pandemic, causing around 50% of worldwide infections (Hemelaar et al., 2019). Subtype C is also the most prevalent strain in Southern Africa and India; subtype A in parts of East Africa, Russia, and former Soviet Union countries; subtype B in Europe, Americas, and Oceania; CRF01_AE in Asia; and CRF02_AG in Western Africa (Bbosa et al., 2019). However, HIV genomic sequencing is more extended in economically developed nations, which explains that in our dataset, the most represented HIV-1 variant was subtype B (37.7%), followed by the most abundant variant subtype C (18.4%), and recombinant CRF01_AE (14.4%), according to sequence availability in LANL.

The HIV capsid protein has an essential structural and functional role in the viral replication cycle. Its genetic variability can impact the efficacy of new p24-based diagnostic, therapeutic, and vaccine strategies. Regarding HIV diagnosis, although diagnostic tests were historically developed based on HIV-1 subtype B prototype strains that showed limitations to detect some variants (Loussert-Ajaka et al., 1994), considerable efforts have been made to improve the performance of these assays. Despite this, the performance of certain serological assays is still suboptimal, mainly in countries where many variants are circulating (Aghokeng et al., 2009; Alvarez et al., 2015; Oladokun et al., 2015; Kravitz Del Solar et al., 2018). Thus, it remains vital to gain a deeper knowledge of p24 variability across HIV-1 variants, which could help to explain false-negative diagnostic results in patients with acute and established HIV infection and to develop a more rational design of new p24-based diagnostic tests. Unfortunately, manufacturers do not provide detailed information regarding which part of the viral sequence was targeted in their HIV diagnostic assays, which can differ across assays. However, the provided information in **Supplementary Table 1** can help manufacturers and researchers, working in the design of new p24-based molecular and serological diagnostic tests, to identify those HIV-1 variants whose diagnosis could be compromised by viral genetic variability.

Therefore, p24 conservation studies must consider all circulating HIV-1 variants worldwide. We present a thorough analysis of p24 variability among 23,671 HIV-1 p24 LANL sequences belonging to more than 100 different variants, including the three non-M HIV-1 groups and a large number of group M subtypes, sub-sub types, and CRF. The sequences were processed by an in-house bioinformatics tool (EpiMolBio) developed for HIV and SARS-CoV-2 variability analysis. Results were analyzed in the context of the secondary structure of p24, focusing on those residues with the most significant functional or therapeutical relevance. At the same time, we report, for the first time to our knowledge, the single natural polymorphisms in p24 that can be considered as genetic markers of each HIV-1 variant or V-markers. We also provide the consensus p24 sequences for HIV-1, HIV-1 groups, and its variants, revealing differences across p24 residues and structural regions. The consensus sequences of HIV proteins and their conservation studies allow a better understanding of structural, functional, and immunogenic potential differences across HIV-1 groups, subtypes, sub-sub types, and recombinants, and have been previously analyzed in other HIV-1 proteins (Sliepen et al., 2019; Zhang et al., 2021). A previous work by Li et al. (2013) analyzed the degree of *gag* functional conservation in 8 group M subtypes (4 subtypes, 2 sub-sub types, and 2 CRF) across 10,862 sequences. Our study updates and expands the knowledge regarding HIV capsid variability, including 23,671 p24 sequences and all the currently available HIV-1 variants, in the LANL database, including the three non-M groups and 125 group M variants (9 subtypes, 7 sub-sub types, and 109 circulating recombinants forms). Moreover, **Supplementary Table 1** summarizes the aa conservation in each capsid residue and each variant to help identify the conservation or consensus aa in any p24 residue and HIV-1 variant of interest in the largest p24 sequence set published to date.

Compared to other HIV proteins, including viral enzymes, p24 is an extremely fragile protein (Rihn et al., 2013), where non-synonymous mutations may drastically reduce its viral fitness. This fragility can be related to the need to maintain its complex structure and its interactions with host proteins. Each p24 monomer must interact with at least three others, while some must adopt slightly different structures to form pentamers that allow the capsid to close (Perilla and Gronenborn, 2016). The p24 structure and assembly mechanisms are complex (Chen, 2016), and the basic geometric principles of the capsid structure are conserved among retroviruses (Aiken and Rousso, 2021). In our study, we observed that the aa conservation in the secondary p24 structure after our sequence analysis varied between HIV-1 variants, amino acid residues, and type of secondary structure. Our results showed high mean aa conservation (>89%) for HIV-1, the four HIV-1 groups, and all group M variants consensus sequences, with only two exceptions: F1 sub-subtype (69.3%) and CRF85_BC (82.2%).

The WK protein variability coefficient was analyzed in HIV-1 group M to study the susceptibility of each aa position to evolutionary replacements. The higher WK values were located around H6 and the β-hairpin in the NTD, except for site 225 at the end of the CTD, although the median variability coefficient for the NTD was lower than for the CTD. The median WK values of the Cyp A-binding loop and the MHR were the same (WK 8), but the MHR sequence had overall fewer variable sites, as expected. A tendency for higher variability was observed at the beginning of the NTD and the end of the CTD, and for less variability between H4 and H8, where the Cyp A-binding loop, the IDR, and the MHR are located.

Previous studies have observed that the β-hairpin, Cyp A-binding loop, and IDR are fairly robust or less conserved regions while the α-helices (mainly H2, H5, H6, and H7) are less tolerant to changes and therefore highly conserved (Rihn et al., 2013). However, when studying the conservation of individual secondary structures, we found that in the HIV-1 consensus, α-helixes 6 and 3 showed the same (83%) or lower (76%) conservation than the Cyp A-binding loop, while the IDR had fairly high conservation of 94%. One of the possible reasons for this discrepancy could be the fact that most studies are centered in group M subtype B, more prevalent in West Europe and the United States (Hemelaar et al., 2019). In the specific analysis of the B subtype consensus, we observed that the IDR conservation dropped to 91.6%, being IDR and the Cyp A-binding loop the least conserved structures, whereas most α-helixes were highly conserved. However, the H6 remained the less conserved α-helix (95%).

The high conservation of p24 also makes it an attractive target for antiretrovirals (Chen, 2016; Novikova et al., 2019; McFadden et al., 2021; Saito and Yamashita, 2021; Dvory-Sobol et al., 2022; Margot et al., 2022). Many capsid-targeting molecules that aim to block HIV-1 infection through different mechanisms of action have been developed (Thenin-Houssier and Valente, 2016; Carnes et al., 2018; Cevik and Orkin, 2019; Bester et al., 2020; Dvory-Sobol et al., 2022). The p24 conservation studies across variants help identify highly conserved regions that may be useful to predict the performance of these new molecules across HIV-1 variants, determine the variability in drug CA-binding sites, and detect if reported drug resistance mutations could be naturally present in certain HIV-1 variants. Some HIV variants, such as HIV-2, present natural polymorphisms related to drug resistance fixed during viral evolution in the absence of antiretroviral therapy, which is maintained over time, providing natural resistance to specific antiretrovirals (Menéndez-Arias and Álvarez, 2014; Troyano-Hernáez et al., 2021a). The provided data in **Supplementary Table 1** will allow the analysis of the conservation percentage in specific positions and variants, including p24 residues, that may be associated with resistance to new CA-inhibitors developed in the future.

Among newly developed CA-inhibitors, one of the most promising molecules is lenacapavir (GS-6207), a selective, long-acting subcutaneous or oral p24 inhibitor with a multi-stage activity that inhibits both the early and late stages of the HIV-1 replication cycle (Link et al., 2020; Dvory-Sobol et al., 2022). This first-in-class capsid inhibitor is currently undergoing phase II/III clinical trials (Gilead, 2022). It has shown successful antiviral activity after a single subcutaneous injection (Dvory-Sobol et al., 2022), high potency, and synergy when combined with other ARV, with no cross-resistance with approved drugs (Link et al., 2020). However, phase Ib treatment led to the emergence of Q67H mutation, while other mutations have emerged in *in vitro* selection experiments, with M66I showing the higher resistance (Bester et al., 2020; Link et al., 2020; Sun et al., 2021). In our study, mutations associated with lenacapavir resistance were unfrequent, present in <0.5% of the sequences in 7 HIV-1 variants. Specifically, M66I was found in <0.2% of the sequences in subtypes B and C, and sub-subtype A1 and CRF07_BC, suggesting that natural resistance to lenacapavir, according to the mutations identified to date, is unlikely across HIV variants.

Regarding p24-lenacapavir interactions, the lenacapavir binding site is located in the CA phenylalanine–glycine (FG) binding pocket, a hydrophobic pocket formed by residues from H3 and H4 α-helixes in the NTD and H8 and H9 α-helixes in the CTD of an adjacent subunit in the hexamer (Link et al., 2020; McFadden et al., 2021). Molecules that bind at the FG binding pocket can compete with host factors that also bind in this pocket, negatively impacting nuclear import and viral infectivity (Matreyek and Engelman, 2011; Price et al., 2012). Among the four p24 residues (N57, K70, N74, and N183), with which lenacapavir establishes key interactions (Link et al., 2020), the first three were highly conserved among HIV groups and variants. On the other hand, N183 showed comparatively lower conservation in group M consensus (77%), mainly due to the presence of glycine (non-polar hydrophobic aa) instead of asparagine (polar hydrophilic aa) in some subtypes and variants. The effect of this aa change and its impact on lenacapavir-CA interaction in these variants should be determined with further structural studies.

The HIV-1 p24 interacts with host cellular proteins required for the viral replication cycle, and one of the most relevant interactions is with Cyp A (Braaten et al., 1996a; Campbell and Hope, 2015). The Cyp A is a chaperone with peptidyl isomerase activity that has a general role in the p24 tertiary structure. The CypA-capsid interactions modulate the capsid stability, affect its capacity to bind other cellular factors, and promote HIV-1 replication in human cells (Hatziioannou et al., 2004; Peng et al., 2019; Selyutina et al., 2020). The host protein Cyp A is packaged into the HIV-1 virion through the binding between Cyp A and p24 exposed loop (aa 85–93), mediating a specific interaction with a Glycine in site 89 and a Proline in site 90 (Gamble et al., 1996). Mutations that alter the aa in these sites can affect HIV-1 infectivity (Braaten et al., 1996a), although the role of Cyp A in HIV-1 infectivity varies between groups, being crucial in group M but not always required in group O viruses (Braaten et al., 1996b; Wiegers and Kräusslich, 2002). In our analysis, sites G89 and P90 were highly conserved in all the studied HIV-1 p24 sequences. In all non-M groups' consensus, site 88, also within the Cyp A active site pocket (Gamble et al., 1996), presented Valine instead of Alanine, the adjacent site 87 presented Proline instead of Histidine, and site 91 had Leucine instead of Isoleucine (**Figure 3**). In site 86, only group O consensus presented Proline instead of Valine. A previous study reported similar changes in group O isolates, with variable effects across them, suggesting that regions outside the Cyp A-binding loop may also affect the Cyp A effect in HIV-1 replication (Wiegers and Kräusslich, 2002). The most variable Cyp A-binding loop sites in group M were 87, 91, and 92 (60–68% conservation). Again, this reinforces the importance of considering all HIV-1 variants, not just group M subtype B, in such studies.

We confirmed the high level of conservation in all HIV-1 groups of the MHR, a 20 amino acid motif in the CTD region of p24 (aa 153–172), highly conserved in all orthoretroviruses. This could be explained by the fact that it is indispensable for the correct assembly of virions (Gamble et al., 1997), specifically in the stabilization step of the Gag oligomer after association with the membrane, where MHR is part of the intrahexameric interface of the immature capsid (Tanaka et al., 2016). Four MHR sites (155, 156, 159, and 167) are essential for their structural role (Gamble et al., 1997) and remain almost invariable between retroviruses, all presenting >99% conservation in the consensus sequences of HIV-1 and the four groups.

The HIV capsid protects the viral components from cytosolic sensors and nucleases by modulating the capsid hexamers' positively charged pore, while it allows access to the nucleotides required for efficient RT (Campbell and Hope, 2015; Jacques et al., 2016). Four key aa have been described for the correct functioning of the pore: P1, H12, R18, and D51, where mutations in P1 and D51 have been reported to produce non-infectious viral particles (Jacques et al., 2016). In our study, these sites were highly conserved in all groups and variants, showing relatively lower conservation in H12 in groups M and O, where the most frequent polymorphism H12Y was found in 0.03 and 0.04% of each group's sequences, respectively. This mutation has been described to favor the closed conformation of the pore, reducing the kinetics of RT (Jacques et al., 2016).

We looked for mutations in p24 related to altered assembly of the immature virus capsid, which were found anecdotally, except for some variants and specific positions, but always in a low proportion (<1% of the total sequences). However, we cannot exclude the possibility that the sequences harboring these mutations belong to non-infectious viruses. Two of the mutations that may affect the interactions between the capsid and capsid-binding molecules (H87Q, G116A) were very frequent in a large number of variants, being considered natural polymorphisms shared between several HIV-1 variants. Q87, located in the Cyp A-binding loop (main consensus aa in most HIV-1 variants), has been associated with reduced Cyp A binding (Yoo et al., 1997; Saito and Yamashita, 2021) and A116 (main consensus residue in 27 HIV-1 variants), with increased infectivity in simian cells (Hatziioannou et al., 2004; Saito and Yamashita, 2021).

We found single specific V-markers in 9 group M variants: CRFs24_BG, 42_BF, 63_02A, 65_cpx, 83_cpx, and 90_BF1, sub-sub types A3 and A6, and subtype F. These markers were conserved in >75% of sequences of the corresponding variant and were unique for that variant. Most of them were found in the capsid NTD. All V-markers found in our study were present at a high proportion of isolates of the corresponding variants, regardless of the sampling year, country, patient, or clinical outcome. Since viral evolution is the result of selective pressures for adaptation over the high diversity of HIV (Bozek and Lengauer, 2010; Peeters et al., 2020), we considered that those markers were fixed during HIV evolution due to different selective pressures when these variants were originated. However, further research is needed to elucidate if these aa changes in p24 residues found in specific variants can affect their p24 function or structure impacting viral fitness, especially those V-markers found in regions with relevant functional implications for the capsid. This would be the case of the V-markers located in the Cyp A-binding loop, H91 (CRF83_cpx) and F91 (sub-subtype A6), and the MHR C171 (sub-subtype A3).

The study's main limitation was the reduced availability of p24 LANL sequences for some HIV-1 variants. Some of them had less than 10 p24 sequences, which excluded them from the V-marker and variant conservation analysis according to our methods to avoid data overestimation. For instance, the low sequence coverage for group P could explain the absence of the aa change G116A, found in all the other HIV groups. The same could occur in the N183 lenacapavir CA-binding site, where serine replaces asparagine in all group P sequences. The HIV genomic sequencing should be strengthened, especially in developing nations, to enrich the information available on non-B variants through the development of cheaper techniques and international collaboration.

The provided information on p24 variability among all HIV-1 variants circulating to date can be helpful for a more rational design of CA-inhibitors, diagnostic tests, and future HIV-1 vaccines based on the capsid protein. Our results also suggest that natural resistance to lenacapavir (based on the mutations identified to date) is unlikely across HIV variants. Our study also provides the first identification of specific natural polymorphisms of p24 or V-markers that can be considered as specific markers in nine HIV variants. Further studies are required to evaluate the impact of the different levels of aa conservation in p24 across HIV-1 variants and of the specific V-markers found in p24 on the capsid structure, assembly, molecule-binding, and other functions in the viral replication cycle.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

PT-H downloaded and analyzed the HIV p24 LANL sequences, validated some EpiMolBio functions necessary for the sequences analyses, performed the computations, discussed the results, and wrote the first version of the manuscript. RR developed the in-house EpiMolBio bioinformatics program and validated the EpiMolBio functions necessary for the sequences analyses. AH designed and supervised the study, discussed the results, reviewed and edited the manuscript, applied for funding, and was responsible for project administration. All authors approved the submitted final version.

## FUNDING

This research was supported by the Instituto de Salud Carlos III (PI18/00904 Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016) and cofinanced by the European Regional Development Fund "A way to achieve Europe" (ERDF). This study was also included in the "Subprograma de Inmigración y Salud" from CIBERESP (Spain). PT-H was funded by ISCIII-Programa Estatal de Promoción del Talento-AES Río Hortega exte. CM19/00057. RR was funded by FONDOS FUR 2020/0285. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.854974/full#supplementary-material

## REFERENCES

Abecasis, A. B., Vandamme, A.-M., and Lemey, P. (2009). Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J. Virol.* 83, 12917–12924. doi: 10.1128/JVI.01022-09

Aghokeng, A. F., Mpoudi-Ngole, E., Dimodi, H., Atem-Tambe, A., Tongo, M., Butel, C., et al. (2009). Inaccurate diagnosis of HIV-1 group M and O is a key challenge for ongoing universal access to antiretroviral treatment and HIV prevention in Cameroon. *PLoS One* 4:e7702. doi: 10.1371/journal.pone.0007702

Aiken, C., and Rousso, I. (2021). The HIV-1 capsid and reverse transcription. *Retrovirology* 18:29. doi: 10.1186/s12977-021-00566-0

Alvarez, P., Martín, L., Prieto, L., Obiang, J., Vargas, A., Avedillo, P., et al. (2015). HIV-1 variability and viral load technique could lead to false positive HIV-1 detection and to erroneous viral quantification in infected specimens. *J. Infect.* 71, 368–376. doi: 10.1016/j.jinf.2015.05.011

Alvarez, P., Prieto, L., Martín, L., Obiang, J., Avedillo, P., Vargas, A., et al. (2017). Evaluation of four commercial virological assays for early infant HIV-1 diagnosis using dried blood specimens. *Pediatr. Res.* 81, 80–87. doi: 10.1038/pr.2016.183

Bala, J., Chinnapaiyan, S., Dutta, R. K., and Unwalla, H. (2018). Aptamers in HIV research diagnosis and therapy. *RNA Biol.* 15, 327–337. doi: 10.1080/15476286.2017.1414131

Bbosa, N., Kaleebu, P., and Ssemwanga, D. (2019). HIV subtype diversity worldwide. *Curr. Opin. HIV AIDS* 14, 153–160. doi: 10.1097/COH.0000000000000534

Beelaert, G., and Fransen, K. (2010). Evaluation of a rapid and simple fourth-generation HIV screening assay for qualitative detection of HIV p24 antigen and/or antibodies to HIV-1 and HIV-2. *J. Virol. Methods* 168, 218–222. doi: 10.1016/j.jviromet.2010.06.002

Bester, S. M., Wei, G., Zhao, H., Adu-Ampratwum, D., Iqbal, N., Courouble, V. V., et al. (2020). Structural and mechanistic bases for a potent HIV-1 capsid inhibitor. *Science* 370, 360–364. doi: 10.1126/science.abb4808

Bozek, K., and Lengauer, T. (2010). Positive selection of HIV host factors and the evolution of lentivirus genes. *BMC Evol. Biol.* 10:186. doi: 10.1186/1471-2148-10-186

Braaten, D., Franke, E. K., and Luban, J. (1996a). Cyclophilin A is required for an early step in the life cycle of human immunodeficiency virus type 1 before the initiation of reverse transcription. *J. Virol.* 70, 3551–3560. doi: 10.1128/JVI.70.6.3551-3560.1996

Braaten, D., Franke, E. K., and Luban, J. (1996b). Cyclophilin A is required for the replication of group M human immunodeficiency virus type 1 (HIV-1) and simian immunodeficiency virus SIV(CPZ)GAB but not group O HIV-1 or other primate immunodeficiency viruses. *J. Virol.* 70, 4220–4227. doi: 10.1128/JVI.70.7.4220-4227.1996

Burdick, R. C., Li, C., Munshi, M., Rawson, J. M. O., Nagashima, K., Hu, W.-S., et al. (2020). HIV-1 uncoats in the nucleus near sites of integration. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5486–5493. doi: 10.1073/pnas.1920631117

Burgos, M., Llácer, T., Reinosa, R., Rubio-Garrido, M., González, A., and Holguín, A. (2019). "Impaired genotypic resistance interpretation due to HIV-1 variant specific Markers," in *10th IAS Conference on HIV Science*, (México: Ciudad de México,).

Campbell, E. M., and Hope, T. J. (2015). HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nat. Rev. Microbiol.* 13, 471–483. doi: 10.1038/nrmicro3503

Carnes, S. K., Sheehan, J. H., and Aiken, C. (2018). Inhibitors of the HIV-1 capsid, a target of opportunity. *Curr. Opin. HIV AIDS* 13, 359–365. doi: 10.1097/COH.0000000000000472

Cevik, M., and Orkin, C. (2019). Insights into HIV-1 capsid inhibitors in preclinical and early clinical development as antiretroviral agents. *Expert Opin. Investig. Drugs* 28, 1021–1024. doi: 10.1080/13543784.2019.1692811

Chen, B. (2016). HIV capsid assembly, mechanism, and structure. *Biochemistry* 55, 2539–2552. doi: 10.1021/acs.biochem.6b00159

Christensen, D. E., Ganser-Pornillos, B. K., Johnson, J. S., Pornillos, O., and Sundquist, W. I. (2020). Reconstitution and visualization of HIV-1 capsid-dependent replication and integration in vitro. *Science* 370:eabc8420. doi: 10.1126/science.abc8420

De Leys, R., Vanderborght, B., Vanden Haesevelde, M., Heyndrickx, L., van Geel, A., Wauters, C., et al. (1990). Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J. Virol.* 64, 1207–1216. doi: 10.1128/JVI.64.3.1207-1216.1990

Dharan, A., Bachmann, N., Talley, S., Zwikelmaier, V., and Campbell, E. M. (2020). Nuclear pore blockade reveals that HIV-1 completes reverse transcription and uncoating in the nucleus. *Nat. Microbiol.* 5, 1088–1095. doi: 10.1038/s41564-020-0735-8

Dick, R. A., Zadrozny, K. K., Xu, C., Schur, F. K. M., Lyddon, T. D., Ricana, C. L., et al. (2018). Inositol phosphates are assembly co-factors for HIV-1. *Nature* 560, 509–512. doi: 10.1038/s41586-018-0396-4

Dvory-Sobol, H., Shaik, N., Callebaut, C., and Rhee, M. S. (2022). Lenacapavir: a first-in-class HIV-1 capsid inhibitor. *Curr. Opin. HIV AIDS* 17, 15–21. doi: 10.1097/COH.0000000000000713

Engelman, A. N. (2021). HIV capsid and integration targeting. *Viruses* 13:125. doi: 10.3390/v13010125

Farzin, L., Shamsipur, M., Samandari, L., and Sheibani, S. (2020). HIV biosensors for early diagnosis of infection: the intertwine of nanotechnology with sensing strategies. *Talanta* 206:120201. doi: 10.1016/j.talanta.2019.120201

Ferrari, G., Kostyu, D. D., Cox, J., Dawson, D. V., Flores, J., Weinhold, K. J., et al. (2000). Identification of highly conserved and broadly cross-reactive HIV type 1 cytotoxic T lymphocyte epitopes as candidate immunogens for inclusion in Mycobacterium bovis BCG-vectored HIV vaccines. *AIDS Res. Hum. Retroviruses* 16, 1433–1443. doi: 10.1089/08892220050140982

Fricke, T., White, T. E., Schulte, B., de Souza Aranha Vieira, D. A., Dharan, A., Campbell, E. M., et al. (2014). MxB binds to the HIV-1 core and prevents the uncoating process of HIV-1. *Retrovirology* 11:68. doi: 10.1186/s12977-014-0068-x

Gamble, T. R., Vajdos, F. F., Yoo, S., Worthylake, D. K., Houseweart, M., Sundquist, W. I., et al. (1996). Crystal structure of human cyclophilin a bound to the amino-terminal domain of HIV-1 capsid. *Cell* 87, 1285–1294. doi: 10.1016/s0092-8674(00)81823-1

Gamble, T. R., Yoo, S., Vajdos, F. F., von Schwedler, U. K., Worthylake, D. K., Wang, H., et al. (1997). Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* 278, 849–853. doi: 10.1126/science.278.5339.849

Gilead (2022). *Pipeline Gilead*. Available online at: https://www.gilead.com/science-and-medicine/pipeline (accessed February 17, 2022).

Gray, E. R., Bain, R., Varsaneux, O., Peeling, R. W., Stevens, M. M., and McKendry, R. A. (2018). P24 revisited: a landscape review of antigen detection for early HIV diagnosis. *AIDS* 32, 2089–2102. doi: 10.1097/QAD.0000000000001982

Hatziioannou, T., Cowan, S., Von Schwedler, U. K., Sundquist, W. I., and Bieniasz, P. D. (2004). Species-specific tropism determinants in the human immunodeficiency virus type 1 capsid. *J. Virol.* 78, 6005–6012. doi: 10.1128/JVI.78.11.6005-6012.2004

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* 18, 182–192. doi: 10.1016/j.molmed.2011.12.001

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

HIV sequence database (2017). *HIV Sequence Database: Nomenclature Overview. HIV Seq. Database Nomencl. Overv*. Available online at: https://www.hiv.lanl.gov/content/sequence/HelpDocs/subtypes-more.html (accessed February 9, 2022).

Hulme, A. E., Kelley, Z., Okocha, E. A., and Hope, T. J. (2015). Identification of capsid mutations that alter the rate of HIV-1 uncoating in infected cells. *J. Virol.* 89, 643–651. doi: 10.1128/JVI.03043-14

Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., and Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* 331, 280–283. doi: 10.1038/331280a0

Jacques, D. A., McEwan, W. A., Hilditch, L., Price, A. J., Towers, G. J., and James, L. C. (2016). HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. *Nature* 536, 349–353. doi: 10.1038/nature19098

Jiang, J., Ablan, S. D., Derebail, S., Hercík, K., Soheilian, F., Thomas, J. A., et al. (2011). The interdomain linker region of HIV-1 capsid protein is a critical determinant of proper core assembly and stability. *Virology* 421, 253–265. doi: 10.1016/j.virol.2011.09.012

Kabat, E. A., Wu, T. T., and Bilofsky, H. (1977). Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *J. Biol. Chem.* 252, 6609–6616. doi: 10.1016/s0021-9258(17)39891-5

Koh, Y., Wu, X., Ferris, A. L., Matreyek, K. A., Smith, S. J., Lee, K., et al. (2013). Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J. Virol.* 87, 648–658. doi: 10.1128/JVI.01148-12

Kosaka, P. M., Pini, V., Calleja, M., and Tamayo, J. (2017). Ultrasensitive detection of HIV-1 p24 antigen by a hybrid nanomechanical-optoplasmonic platform with potential for detecting HIV-1 at first week after infection. *PLoS One* 12:e0171899. doi: 10.1371/journal.pone.0171899

Kravitz Del Solar, A. S., Parekh, B., Douglas, M. O., Edgil, D., Kuritsky, J., and Nkengasong, J. (2018). A Commitment to HIV Diagnostic Accuracy - a comment on "Towards more accurate HIV testing in sub-Saharan Africa: a multi-site evaluation of HIV RDTs and risk factors for false positives 'and'. HIV misdiagnosis in sub-Saharan Africa: a performance of diagnostic algorithms at six testing sites". *J. Int. AIDS Soc.* 21:e25177. doi: 10.1002/jia2.25177

Larijani, M. S., Sadat, S. M., Bolhassani, A., Pouriayevali, M. H., Bahramali, G., and Ramezani, A. (2018). In Silico design and immunologic evaluation of HIV-1 p24-Nef fusion protein to approach a therapeutic vaccine candidate. *Curr. HIV Res.* 16, 322–337. doi: 10.2174/1570162X17666190102151717

Larijani, M. S., Ramezani, A., Mashhadi Abolghasem Shirazi, M., Bolhassani, A., Pouriayevali, M. H., Shahbazi, S., et al. (2021a). Evaluation of transduced dendritic cells expressing HIV-1 p24-Nef antigens in HIV-specific cytotoxic T cells induction as a therapeutic candidate vaccine. *Virus Res.* 298:198403. doi: 10.1016/j.virusres.2021.198403

Larijani, M. S., Sadat, S. M., Bolhassani, A., Khodaie, A., Pouriayevali, M. H., and Ramezani, A. (2021b). HIV-1 p24-nef DNA vaccine plus protein boost expands T-Cell responses in BALB/c. *Curr. Drug Deliv.* 18, 1014–1021. doi: 10.2174/1567201818666210101113601

Le Sage, V., Mouland, A. J., and Valiente-Echeverría, F. (2014). Roles of HIV-1 capsid in viral replication and immune evasion. *Virus Res.* 193, 116–129. doi: 10.1016/j.virusres.2014.07.010

Li, C., Burdick, R. C., Nagashima, K., Hu, W.-S., and Pathak, V. K. (2021). HIV-1 cores retain their integrity until minutes before uncoating in the nucleus. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2019467118. doi: 10.1073/pnas.2019467118

Li, G., Verheyen, J., Rhee, S.-Y., Voet, A., Vandamme, A.-M., and Theys, K. (2013). Functional conservation of HIV-1 Gag: implications for rational drug design. *Retrovirology* 10:126. doi: 10.1186/1742-4690-10-126

Lingappa, J. R., Reed, J. C., Tanaka, M., Chutiraka, K., and Robinson, B. A. (2014). How HIV-1 gag assembles in cells: putting together pieces of the puzzle. *Virus Res.* 193, 89–107. doi: 10.1016/j.virusres.2014.07.001

Link, J. O., Rhee, M. S., Tse, W. C., Zheng, J., Somoza, J. R., Rowe, W., et al. (2020). Clinical targeting of HIV capsid protein with a long-acting small molecule. *Nature* 584, 614–618. doi: 10.1038/s41586-020-2443-1

Los Alamos National Laboratory (LANL) (2021a). *Gag CTL/CD8$^+$ Epitope Map. HIV Molecular Immunology Database*. Available online at: https://www.hiv.lanl.gov/content/immunology/maps/ctl/Gag.html (accessed January 10, 2022).

Los Alamos National Laboratory (2021b). *HIV Circulating Recombinant Forms (CRFs).* Available online at: https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html (accessed January 10, 2022).

Los Alamos National Laboratory (2021c). *HIV Sequence Database.* Available online at: https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html (accessed January 10, 2022).

Loussert-Ajaka, I., Ly, T. D., Chaix, M. L., Ingrand, D., Saragosti, S., Couroucé, A. M., et al. (1994). HIV-1/HIV-2 seronegativity in HIV-1 subtype O infected patients. *Lancet* 343, 1393–1394. doi: 10.1016/s0140-6736(94)92524-0

Ly, T. D., Plantier, J. C., Leballais, L., Gonzalo, S., Lemée, V., and Laperche, S. (2012). The variable sensitivity of HIV Ag/Ab combination assays in the detection of p24Ag according to genotype could compromise the diagnosis of early HIV infection. *J. Clin. Virol. Off. Publ. Pan Am. Soc. Clin. Virol.* 55, 121–127. doi: 10.1016/j.jcv.2012.06.012

Mallery, D. L., Márquez, C. L., McEwan, W. A., Dickson, C. F., Jacques, D. A., Anandapadamanaban, M., et al. (2018). IP6 is an HIV pocket factor that prevents capsid collapse and promotes DNA synthesis. *Elife* 7:e35335. doi: 10.7554/eLife.35335

Mamede, J. I., Cianci, G. C., Anderson, M. R., and Hope, T. J. (2017). Early cytoplasmic uncoating is associated with infectivity of HIV-1. *Proc. Natl. Acad. Sci. U.S.A.* 114, E7169–E7178. doi: 10.1073/pnas.1706245114

Mammano, F., Ohagen, A., Höglund, S., and Göttlinger, H. G. (1994). Role of the major homology region of human immunodeficiency virus type 1 in virion morphogenesis. *J. Virol.* 68, 4927–4936. doi: 10.1128/JVI.68.8.4927-4936.1994

Margot, N., Vanderveen, L., Naik, V., Ram, R., Parvangada, P. C., Martin, R., et al. (2022). Phenotypic resistance to lenacapavir and monotherapy efficacy in a proof-of-concept clinical study. *J. Antimicrob. Chemother.* 13:dkab503. doi: 10.1093/jac/dkab503

Márquez, C. L., Lau, D., Walsh, J., Shah, V., McGuinness, C., Wong, A., et al. (2018). Kinetics of HIV-1 capsid uncoating revealed by single-molecule analysis. *Elife* 7:e34772. doi: 10.7554/eLife.34772

Matreyek, K. A., and Engelman, A. (2011). The requirement for nucleoporin NUP153 during human immunodeficiency virus type 1 infection is determined by the viral capsid. *J. Virol.* 85, 7818–7827. doi: 10.1128/JVI.00325-11

McFadden, W. M., Snyder, A. A., Kirby, K. A., Tedbury, P. R., Raj, M., Wang, Z., et al. (2021). Rotten to the core: antivirals targeting the HIV-1 capsid core. *Retrovirology* 18:41. doi: 10.1186/s12977-021-00583-z

Menéndez-Arias, L., and Álvarez, M. (2014). Antiretroviral therapy and drug resistance in human immunodeficiency virus type 2 infection. *Antiviral Res.* 102, 70–86. doi: 10.1016/j.antiviral.2013.12.001

Miller, M. D., Farnet, C. M., and Bushman, F. D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* 71, 5382–5390. doi: 10.1128/JVI.71.7.5382-5390.1997

Novikova, M., Zhang, Y., Freed, E. O., and Peng, K. (2019). Multiple Roles of HIV-1 Capsid during the Virus Replication Cycle. *Virol. Sin.* 34, 119–134. doi: 10.1007/s12250-019-00095-3

Oladokun, R., Korsman, S., Ndabambi, N., Hsiao, N., Hans, L., Williamson, C., et al. (2015). False-negative HIV-1 polymerase chain reaction in a 15-month-old boy with HIV-1 subtype C infection. *S. Afr. Med. J.* 105:877. doi: 10.7196/samjnew.8787

Parekh, B. S., Ou, C.-Y., Fonjungo, P. N., Kalou, M. B., Rottinghaus, E., Puren, A., et al. (2019). Diagnosis of human immunodeficiency virus infection. *Clin. Microbiol. Rev.* 32, e00064–e00118. doi: 10.1128/CMR.00064-18

Peeters, M., and Sharp, P. M. (2000). Genetic diversity of HIV-1: the moving target. *AIDS* 14(Suppl. 3), S129–S140.

Peeters, M., Mulanga-Kabeya, C., and Delaporte, E. (2020). La diversité génétique du VIH Type 1. *Virologie* 4, 313–320.

Peng, W., Shi, J., Márquez, C. L., Lau, D., Walsh, J., Faysal, K. M. R., et al. (2019). Functional analysis of the secondary HIV-1 capsid binding site in the host protein cyclophilin A. *Retrovirology* 16:10. doi: 10.1186/s12977-019-0471-4

Perilla, J. R., and Gronenborn, A. M. (2016). Molecular architecture of the retroviral capsid. *Trends Biochem. Sci.* 41, 410–420. doi: 10.1016/j.tibs.2016.02.009

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nat. Med.* 15, 871–872. doi: 10.1038/nm.2016

Price, A. J., Fletcher, A. J., Schaller, T., Elliott, T., Lee, K., KewalRamani, V. N., et al. (2012). CPSF6 defines a conserved capsid interface that modulates HIV-1 replication. *PLoS Pathog.* 8:e1002896. doi: 10.1371/journal.ppat.1002896

Qi, M., Yang, R., and Aiken, C. (2008). Cyclophilin A-dependent restriction of human immunodeficiency virus type 1 capsid mutants for infection of nondividing cells. *J. Virol.* 82, 12001–12008. doi: 10.1128/JVI.01518-08

Qiu, X., Sokoll, L., Yip, P., Elliott, D. J., Dua, R., Mohr, P., et al. (2017). Comparative evaluation of three FDA-approved HIV Ag/Ab combination tests using a genetically diverse HIV panel and diagnostic specimens. *J. Clin. Virol. Off. Publ. Pan Am. Soc. Clin. Virol.* 92, 62–68. doi: 10.1016/j.jcv.2017.05.005

Renner, N., Mallery, D. L., Faysal, K. M. R., Peng, W., Jacques, D. A., Böcking, T., et al. (2021). A lysine ring in HIV capsid pores coordinates IP6 to drive mature capsid assembly. *PLoS Pathog.* 17:e1009164. doi: 10.1371/journal.ppat.1009164

Rihn, S. J., Wilson, S. J., Loman, N. J., Alim, M., Bakker, S. E., Bhella, D., et al. (2013). Extreme genetic fragility of the HIV-1 capsid. *PLoS Pathog.* 9:e1003461. doi: 10.1371/journal.ppat.1003461

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Rodríguez-Galet, A., García López, , S., Kosaka, P., and Holguín, A. (2022). "New molecular assay based on nanotechnology for the early detection of HIV-1 p24," in *Proceedings of the CROI 2022*, Poster ID 1093, (San Francisco, CA: CROI).

Saito, A., and Yamashita, M. (2021). HIV-1 capsid variability: viral exploitation and evasion of capsid-binding molecules. *Retrovirology* 18:32. doi: 10.1186/s12977-021-00577-x

Salminen, M. O., Ehrenberg, P. K., Mascola, J. R., Dayhoff, D. E., Merling, R., Blake, B., et al. (2000). Construction and biological characterization of infectious molecular clones of HIV-1 subtypes B and E (CRF01_AE) generated by the polymerase chain reaction. *Virology* 278, 103–110. doi: 10.1006/viro.2000.0640

Selyutina, A., Persaud, M., Simons, L. M., Bulnes-Ramos, A., Buffone, C., Martinez-Lopez, A., et al. (2020). Cyclophilin a prevents HIV-1 restriction in lymphocytes by blocking human TRIM5α binding to the viral core. *Cell Rep.* 30, 3766.e–3777.e. doi: 10.1016/j.celrep.2020.02.100

Simon, F., Mauclère, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M. C., Saragosti, S., et al. (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med.* 4, 1032–1037. doi: 10.1038/2017

Sliepen, K., Han, B. W., Bontjer, I., Mooij, P., Garces, F., Behrens, A.-J., et al. (2019). Structure and immunogenicity of a stabilized HIV-1 envelope trimer based on a group-M consensus sequence. *Nat. Commun.* 10:2355. doi: 10.1038/s41467-019-10262-5

Stone, M., Bainbridge, J., Sanchez, A. M., Keating, S. M., Pappas, A., Rountree, W., et al. (2018). Comparison of detection limits of fourth- and fifth-generation combination HIV antigen-antibody, p24 antigen, and viral load assays on diverse HIV isolates. *J. Clin. Microbiol.* 56, e02045–e02117. doi: 10.1128/JCM.02045-17

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., et al. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5514–5519. doi: 10.1073/pnas.0509996103

Sun, Q., Levy, R. M., Kirby, K. A., Wang, Z., Sarafianos, S. G., and Deng, N. (2021). Molecular dynamics free energy simulations reveal the mechanism for the antiviral resistance of the M66I HIV-1 capsid mutation. *Viruses* 13:920. doi: 10.3390/v13050920

Sundquist, W. I., and Kräusslich, H.-G. (2012). HIV-1 assembly, budding, and maturation. *Cold Spring Harb. Perspect. Med.* 2:a006924. doi: 10.1101/cshperspect.a006924

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tanaka, M., Robinson, B. A., Chutiraka, K., Geary, C. D., Reed, J. C., and Lingappa, J. R. (2016). Mutations of conserved residues in the major homology region arrest assembling HIV-1 gag as a membrane-targeted intermediate containing genomic RNA and cellular proteins. *J. Virol.* 90, 1944–1963. doi: 10.1128/JVI.02698-15

Temple, J., Tripler, T. N., Shen, Q., and Xiong, Y. (2020). A snapshot of HIV-1 capsid-host interactions. *Curr. Res. Struct. Biol.* 2, 222–228. doi: 10.1016/j.crstbi.2020.10.002

Thenin-Houssier, S., and Valente, S. T. (2016). HIV-1 capsid inhibitors as antiretroviral agents. *Curr. HIV Res.* 14, 270–282. doi: 10.2174/1570162x14999160224103555

Toccafondi, E., Lener, D., and Negroni, M. (2021). HIV-1 capsid core: a bullet to the heart of the target cell. *Front. Microbiol.* 12:652486. doi: 10.3389/fmicb.2021.652486

Torrecilla, E., Llácer Delicado, , T., and Holguín, A. (2014). New findings in cleavage sites variability across groups, subtypes and recombinants of human immunodeficiency virus type 1. *PLoS One* 9:e88099. doi: 10.1371/journal.pone.0088099

Troyano-Hernáez, P., Reinosa, R., and Holguín, A. (2020). "Mutaciones en la proteína Spike de SARS-CoV-2 por Comunidades Autónomas en secuencias españolas recogidas hasta junio 2020," in *Proceedings of the I Congreso Nacional COVID-19*, (Spain: SETH), 76.

Troyano-Hernáez, P., Reinosa, R., Burgos, M. C., and Holguín, A. (2021a). Short communication: update in natural antiretroviral resistance-associated mutations among HIV Type 2 variants and discrepancies across HIV Type 2 resistance interpretation tools. *AIDS Res. Hum. Retroviruses* 37, 793–795. doi: 10.1089/AID.2020.0180

Troyano-Hernáez, P., Reinosa, R., and Holguín, A. (2021b). Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to september 2020: a global and regional approach by epidemiological week. *Viruses* 13:243. doi: 10.3390/v13020243

Vetter, B. N., Orlowski, V., Niederhauser, C., Walter, L., and Schüpbach, J. (2015). Impact of naturally occurring amino acid variations on the detection of HIV-1 p24 in diagnostic antigen tests. *BMC Infect. Dis.* 15:468. doi: 10.1186/s12879-015-1174-7

Wiegers, K., and Kräusslich, H.-G. (2002). Differential dependence of the infectivity of HIV-1 group O isolates on the cellular protein cyclophilin A. *Virology* 294, 289–295. doi: 10.1006/viro.2001.1347

Wilbourne, M., and Zhang, P. (2021). Visualizing HIV-1 capsid and its interactions with antivirals and host factors. *Viruses* 13:246. doi: 10.3390/v13020246

Yamaguchi, J., Vallari, A., McArthur, C., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2020). Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *J. Acquir. Immune Defic. Syndr.* 83, 319–322. doi: 10.1097/QAI.0000000000002246

Yamashita, M., and Engelman, A. N. (2017). Capsid-Dependent host factors in HIV-1 infection. *Trends Microbiol.* 25, 741–755. doi: 10.1016/j.tim.2017.04.004

Yamashita, M., Perez, O., Hope, T. J., and Emerman, M. (2007). Evidence for direct involvement of the capsid protein in HIV infection of nondividing cells. *PLoS Pathog.* 3:1502–1510. doi: 10.1371/journal.ppat.0030156

Yant, S. R., Mulato, A., Hansen, D., Tse, W. C., Niedziela-Majka, A., Zhang, J. R., et al. (2019). A highly potent long-acting small-molecule HIV-1 capsid inhibitor with efficacy in a humanized mouse model. *Nat. Med.* 25, 1377–1384. doi: 10.1038/s41591-019-0560-x

Ylinen, L. M. J., Schaller, T., Price, A., Fletcher, A. J., Noursadeghi, M., James, L. C., et al. (2009). Cyclophilin a levels dictate infection efficiency of human immunodeficiency virus type 1 capsid escape mutants A92E and G94D. *J. Virol.* 83, 2044–2047. doi: 10.1128/JVI.01876-08

Yoo, S., Myszka, D. G., Yeh, C., McMurray, M., Hill, C. P., and Sundquist, W. I. (1997). Molecular recognition in the HIV-1 capsid/cyclophilin a complex. *J. Mol. Biol.* 269, 780–795. doi: 10.1006/jmbi.1997.1051

Zhang, Y., Murakoshi, H., Chikata, T., Akahoshi, T., Tran, G., Van Nguyen, T. V., et al. (2021). Effect of difference in consensus sequence between HIV-1 Subtype A/E and Subtype B viruses on elicitation of gag-specific CD8(+) T cells and accumulation of HLA-associated escape mutations. *J. Virol* 95, e02061–e02120. doi: 10.1128/JVI.02061-20

Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., et al. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497, 643–646. doi: 10.1038/nature12162

Zhou, J., and Rossi, J. (2017). Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.* 16, 181–202. doi: 10.1038/nrd.2016.199

Zhuang, S., and Torbett, B. E. (2021). Interactions of HIV-1 capsid with host factors and their implications for developing novel therapeutics. *Viruses* 13:417. doi: 10.3390/v13030417

Zila, V., Margiotta, E., Turoòová, B., Müller, T. G., Zimmerli, C. E., Mattei, S., et al. (2021a). Cone-shaped HIV-1 capsids are transported through intact nuclear pores. *Cell* 184, 1032.e–1046.e. doi: 10.1016/j.cell.2021.01.025

Zila, V., Müller, T. G., Müller, B., and Kräusslich, H.-G. (2021b). HIV-1 capsid is the key orchestrator of early viral replication. *PLoS Pathog.* 17:e1010109. doi: 10.1371/journal.ppat.1010109

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 April 2022
doi: 10.3389/fmicb.2022.846943

# Trends of Transmitted and Acquired Drug Resistance in Europe From 1981 to 2019: A Comparison Between the Populations of Late Presenters and Non-late Presenters

**Mafalda N. S. Miranda[1]*, Marta Pingarilho[1], Victor Pimentel[1], Maria do Rosário O. Martins[1], Rolf Kaiser[2], Carole Seguin-Devaux[3], Roger Paredes[4], Maurizio Zazzi[5], Francesca Incardona[6,7] and Ana B. Abecasis[1]**

*[1] Global Health and Tropical Medicine (GHTM), Institute of Hygiene and Tropical Medicine, New University of Lisbon (IHMT/UNL), Lisbon, Portugal, [2] Institute of Virology, University of Cologne, Cologne, Germany, [3] Laboratory of Retrovirology, Department of Infection and Immunity, Luxembourg Institute of Health, Esch-sur-Alzette, Luxembourg, [4] Infectious Diseases Department and IrsiCaixa AIDS Research Institute, Hospital Universitari Germans Trias i Pujol, Badalona, Spain, [5] Department of Medical Biotechnologies, University of Siena, Siena, Italy, [6] IPRO—InformaPRO S.r.l., Rome, Italy, [7] EuResist Network, Rome, Italy*

OPEN ACCESS

**Edited by:**
Kok Keng Tee,
University of Malaya, Malaysia

**Reviewed by:**
Lingjie Liao,
National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, China
Sunee Sirivichayakul,
Chulalongkorn University, Thailand
Emmanuel Ndashimye,
Western University, Canada

***Correspondence:**
Mafalda N. S. Miranda
mafalda_nsm@hotmail.com

**Specialty section:**
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

**Received:** 31 December 2021
**Accepted:** 15 February 2022
**Published:** 13 April 2022

**Citation:**
Miranda MNS, Pingarilho M, Pimentel V, Martins MdRO, Kaiser R, Seguin-Devaux C, Paredes R, Zazzi M, Incardona F and Abecasis AB (2022) Trends of Transmitted and Acquired Drug Resistance in Europe From 1981 to 2019: A Comparison Between the Populations of Late Presenters and Non-late Presenters. Front. Microbiol. 13:846943. doi: 10.3389/fmicb.2022.846943

**Background:** The increased use of antiretroviral therapy (ART) has decreased mortality and morbidity of HIV-1 infected people but increasing levels of HIV drug resistance threatens the success of ART regimens. Conversely, late presentation can impact treatment outcomes, health costs, and potential transmission of HIV.

**Objective:** To describe the patterns of transmitted drug resistance (TDR) and acquired drug resistance (ADR) in HIV-1 infected patients followed in Europe, to compare its patterns in late presenters (LP) vs non-late presenters (NLP), and to analyze the most prevalent drug resistance mutations among HIV-1 subtypes.

**Methods:** Our study included clinical, socio-demographic, and genotypic information from 26,973 HIV-1 infected patients from the EuResist Integrated Database (EIDB) between 1981 and 2019.

**Results:** Among the 26,973 HIV-1 infected patients in the analysis, 11,581 (42.9%) were ART-naïve patients and 15,392 (57.1%) were ART-experienced. The median age was 37 (IQR: 27.0–45.0) years old and 72.6% were males. The main transmission route was through heterosexual contact (34.9%) and 81.7% of patients originated from Western Europe. 71.9% of patients were infected by subtype B and 54.8% of patients were classified as LP. The overall prevalence of TDR was 12.8% and presented an overall decreasing trend ($p$ for trend $< 0.001$), the ADR prevalence was 68.5% also with a decreasing trend ($p$ for trend $< 0.001$). For LP and NLP, the TDR prevalence was 12.3 and 12.6%, respectively, while for ADR, 69.9 and 68.2%, respectively. The most prevalent TDR drug resistance mutations, in both LP and NLP, were K103N/S, T215rev, T215FY, M184I/V, M41I/L, M46I/L, and L90M.

**Conclusion:** Our study showed that the overall TDR (12.8%) and ADR (68.5%) presented decreasing trends during the study time period. For LP, the overall TDR was slightly lower than for NLP (12.3 vs 12.6%, respectively); while this pattern was opposite for ADR (LP slightly higher than NLP). We suggest that these differences, in the case of TDR, can be related to the dynamics of fixation of drug resistance mutations; and in the case of ADR with the more frequent therapeutic failure in LPs.

**Keywords: HIV-1 infection, transmitted drug resistance, acquired drug resistance, late presenters, non-late presenters**

## INTRODUCTION

In 2014, UNAIDS implemented the Fast-Track approach driven by the 95-95-95 targets. These targets have the aim to end the pandemic by 2030 by achieving 95% of diagnosis among people living with HIV, 95% of those receiving antiretroviral treatment and 95% of those reaching viral suppression (Joint United Nations Programme on Hiv/Aids (Unaids), 2015). In the meantime, UNAIDS has developed a set of targets for 2025 to help achieve the previous goals until 2030, which are people-centered and right-based (Unaids).

At the end of 2020, there were 37.7 million people living with HIV and at least 50% of the new diagnoses were related to late HIV infection [late presenters (LP)], with regional differences. LP are patients newly diagnosed with a baseline CD4 count lower than 350 cells/mm$^3$ or with an AIDS-defining event, regardless of CD4 cell count (Miranda et al., 2021). Between 2000 and 2020 the percentage of new HIV infections dropped by 49% and HIV-related deaths dropped by 55% due to antiretroviral therapy (ART; World Health Organization).

The advent of highly active ART has greatly improved the prognosis of HIV-1 infection and reduction of the risk of HIV transmission (Cdc). Today, 73% of people living with HIV have access to ART. Drug resistance could be acquired drug resistance (ADR), due to selective pressure of antiretrovirals (ARVs) in individuals, or transmitted drug resistance (TDR) due to an infection by HIV strains that harbor drug resistance mutations (DRMs; Clutter et al., 2016; Pingarilho et al., 2020).

Drug resistance testing is recommended for individuals with HIV infection who are newly diagnosed or ART-naïve patients, individuals on ART with a viral load higher than 200 copies/mL, individuals who did not achieve viral suppression, and individuals who interrupted ART with a non-nucleoside reverse transcriptase inhibitor (NNRTI; Günthard et al., 2019). For ART-naïve patients, genotypic drug-resistance testing involved testing for mutations in the reverse transcriptase (RT), protease (PR) and integrase (IN) genes. In ART-experienced patients, genotypic and phenotypic resistance testing is recommended in individuals suspect of multi drug-resistance mutations and virological failure (Nih).

The most common DRMs among ART-naïve and ART-experienced patients for nucleoside reverse transcriptase inhibitors (NRTIs) were M41L and M184V, respectively, and K103N for NNRTIs (Rossetti et al., 2018; Zou et al., 2020).

In 2016, the World Health Organization (WHO) recommended the following guidelines as a first-line ART regimen: the combination of two NRTIs, such as tenofovir (TDF) and lamivudine (3TC) or emtricitabine (FTC), plus an integrase strand inhibitor (INSTI), such as dolutegravir (DTG), or instead of DTG the combination with the NNRTI efavirenz (EFV). The recommendations for second-line regimens included the combination of two NRTIs plus one protease inhibitor (PI), like atazanavir (ATV) or lopinavir/ritonavir (LPV/RTV) or two NRTIs and DTG. Third-line regimens included the combination of one PI, such as darunavir (DRV), DTG, and one or two NRTIs (World Health Organization).

Resistance to ART could decrease the success of first line regimens and is a major threat to halt the UNAIDS targets, as well as late presentation. Resistance to antiretrovirals and late presentation are still existing problems that could delay the success of regimens and continue the onward transmission of HIV-1 infection. In this study, we aim to describe the patterns of TDR and ADR, as well as compare them in LP and non-late presenter (NLP) populations included in this study. We also analyzed the most prevalent drug resistance mutations and their prevalence in HIV-1 subtypes among LP and NLP HIV-1 infected patients followed in Europe.

## METHODS

### Study Group

Clinical, socio-demographic, and genomic information from 26,973 HIV-1 infected patients from the EuResist Integrated Database (EIDB) between 1981 and 2019 were included in this study. The EIDB is one of the largest existing datasets which integrate clinical, socio-demographic, and viral genotypic information from HIV-1 patients. It integrates longitudinal, periodically updated data mainly from Italy (ARCA database), Germany (AREVIR database) Spain (CoRIS and IRSICAIXA), Sweden, Belgium, Portugal, and Luxembourg (EuResist; Lawyer et al., 2011; Zazzi et al., 2012).

In this study, information from the ARCA, AREVIR, Luxembourg, IRSICAIXA, and Portugal databases were used.

### Exclusion Criteria

Among the 89,851 HIV-1 infected patients included in the EuResist database, only 54,176 patients had sequence information for the RT and PR regions. Those patient sequences went through the quality control process. We calculated the ambiguity rate for each genomic sequence and only included those sequences that were larger than 500 nucleotides and with an ambiguity rate lower than 2.5%, resulting in the elimination of 4,044 sequences. Our final study population included 26,973 HIV-1 infected patients, because of the 50,132 patients, only 26,973 had information regarding their date of first ARV therapy.

## Institutional Review Board Statement

All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration. The database enrolled anonymized patients' information, including demographic, clinical, and genomic data from patients from the EuResist Integrated Database (Date of approval: January 15, 2021).

## Drug Resistance Analysis and Subtyping

HIV pol sequences were derived from existing routine clinical genotypic resistance tests (Sanger method, e.g., Viroseq, Trugene and in house genotyping). The size of RT and PR fragments used for this analysis were between 500 and 1,000 nucleotides. Only the first HIV genomic sequence per patient was analyzed. TDR was defined as the presence of one or more surveillance drug resistance mutations in a sequence, according to the WHO 2009 surveillance list (Bennett et al., 2009). The sequences were submitted to the Calibrated Population Resistance tool version 8.0. Clinical resistance to ARV drugs was calculated through the Standford HIVdb version 9.0.

We analyzed TDR and ADR overall proportions between 1981 and 2019, although we only used the years 1995–2019, divided into three time periods (1995–2002; 2003–2010, and 2011–2019), to compute TDR and ADR trends, since the absolute number before 1995 was smaller than 10 patients per year. We also analyzed TDR and ADR proportions in countries of follow-up. For this analysis, we limited the analyses to the last 10 years divided into two time periods (2008–2012 and 2013–2019).

HIV-1 subtyping was performed using the consensus of the result obtained based on three different subtyping tools: Rega HIV Subtyping Tool version 3.46[1] (Pineda-Peña et al., 2013), COMET: adaptive context-based modeling for HIV-1[2] (Struck et al., 2014) and SCUEAL.[3]

## Study Variables

New variables were created according to:

- Migrant/Native—based on country of origin and country of follow-up (if country of origin and country of follow-up is the same, then patient was classified as native; otherwise as migrant)
- Age at Drug Resistance Test—based on the difference between year of birth and date of the first drug resistance test;
- Region of Origin—based on country of origin;
- Treatment Status at Date of First Drug Resistance Test—based on the difference between sample collection date for first drug resistance test and start date of first therapy:

  ART-naïve → patients who had a sample collection date for first drug resistance test before the start date of first therapy
  ART-experienced → patients who had a sample collection date for first drug resistance test after the start date of first therapy

- Recentness of infection—based on ambiguity rate of genomic sequences. We defined as Chronic if the ambiguity rate was higher than 0.45% otherwise was defined as Recent infection, as previously described (Andersson et al., 2013).

LP vs NLP at HIV diagnosis- based on CD4 count, LP were defined as patients with a baseline CD4 count $\leq$ 350 cells/mm$^3$ and NLP were defined as patients with baseline CD4 count > 350 cells/mm$^3$ (Antinori et al., 2011).

## Statistical Analysis

The proportion and median [interquartile range (IQR)] were calculated for every categorical and continuous variable, respectively. The treatment status variable was compared with the categorical variables with the Chi-square test and continuous variables with the Mann-Whitney U test. Also, we analyzed the trends over time for the overall TDR and ADR through logistic regression models. Data was analyzed using RStudio (Version 1.2.5033).

# RESULTS

## Characteristics of European Population

Among the 26,973 HIV-1 infected patients included in the analysis, 11,581 (42.9%) were ART-naïve patients and 15,392 (57.1%) were ART-experienced patients. Other socio-demographic characteristics of the population of patients has been analyzed and described in "Determinants of Determinants of HIV-1 Late Presentation in Patients Followed in Europe" (Miranda et al., 2021).

In the total population, the median age was 37 (IQR: 27.0–45.0) years old and 72.6% of HIV-1 infected patients were males. The main transmission route was through heterosexual contact (34.9%) and 81.7% were originated from Western Europe. The most prevalent subtype observed in this population was subtype B (71.9%). Most patients included in this study were native (77.4%) and as having chronic infection (63.6%) based on the ambiguity rate of the first genomic sequence. CD4 count at diagnosis and viral load at diagnosis (log10) presented a median of 318 cells/mm$^3$ (IQR 151–513) and log10 4.3 copies/mL (IQR 3.3–5.0), respectively.

54.8% of patients were classified as LP (CD4 < 350 cells/mm$^3$). In ART-naïve patients, 52.8% were LP, meanwhile in ART-experienced patients, 56.4% were LP at time of diagnosis (**Table 1**).

[1] https://www.genomedetective.com/app/typingtool/hiv

[2] https://comet.lih.lu

[3] http://classic.datamonkey.org/dataupload_scueal.php

**TABLE 1 |** Sociodemographic and clinical patient characteristics.

| Patient characteristics | Total | ART-naive | ART-experienced | *p*-value |
|---|---|---|---|---|
| Total, *n* (%) | 26973 (100) | 11581 (42.9) | 15392 (57.1) | |
| Gender, *n* (%) | 26475 (98.2) | 11458 (43.3) | 15017 (56.7) | |
| Male | 19224 (72.6) | 8797 (76.8) | 10427 (69.4) | $p < 0.001$ |
| Female | 7251 (27.4) | 2661 (23.2) | 4590 (30.6) | |
| Median age at resistance test in years IQR, *n* (%) | 26973 (100) | 11581 (42.9) | 15392 (57.1) | $p < 0.001$ |
| | 37.0 (27.0–45.0) | 37.0 (30.0–45.0) | 37.0 (0.0–44.0) | |
| ≤18 | 5047 (18.7) | 761 (6.6) | 4286 (27.8) | $p < 0.001$ |
| 19–30 | 3423 (12.7) | 2468 (21.3) | 955 (6.2) | |
| 31–55 | 16707 (61.9) | 7472 (64.5%) | 9235 (60.0) | |
| ≥56 | 1796 (6.7) | 880 (7.6) | 916 (6.0) | |
| Transmission route, *n* (%) | 18118 (67.2) | 8336 (46.0) | 9782 (54.0) | |
| Heterosexual | 6326 (34.9) | 3130 (37.5) | 3196 (32.7) | $p < 0.001$ |
| MSM | 6124 (33.8) | 3863 (46.3) | 2261 (23.1) | |
| IDU | 4370 (24.1) | 838 (10.1) | 3532 (36.1) | |
| Other | 1298 (7.2) | 505 (6.1) | 793 (8.1) | |
| Region of origin, *n* (%) | 19881 (73.7) | 9460 (47.6) | 10421 (52.4) | |
| Western Europe | 16249 (81.7) | 7436 (78.6) | 8813 (84.6) | $p < 0.001$ |
| Eastern Europe | 554 (2.8) | 377 (4.0) | 177 (1.7) | |
| Africa | 2109 (10.6) | 1051 (11.1) | 1058 (10.2) | |
| South America | 611 (3.1) | 338 (3.6) | 273 (2.6) | |
| Other | 358 (1.8) | 258 (2.7) | 100 (1.0) | |
| Migrant status, *n* (%) | 19881 (73.7) | 9460 (47.6) | 10421 (52.4) | |
| Migrant | 4494 (22.6) | 2616 (27.7) | 1878 (18.0) | $p < 0.001$ |
| Native | 15387 (77.4) | 6844 (72.3) | 8543 (82.0) | |
| Recentness of infection, *n* (%) | 26973 (100) | 11581 (42.9) | 15392 (57.1) | |
| Chronic | 17151 (63.6) | 6915 (59.7) | 10236 (66.5) | $p < 0.001$ |
| Recent | 9822 (36.4) | 4666 (40.3) | 5156 (33.5) | |
| Subtype, *n* (%) | 26973 (100) | 11581 (42.9) | 15392 (57.1) | |
| HIV-1 Subtype B | 19387 (71.9) | 8047 (69.5) | 11340 (73.7) | $p < 0.001$ |
| HIV-1 Subtype non-B | 7586 (28.1) | 3534 (30.5) | 4052 (26.3) | |
| Median (IQR) CD4 count at diagnosis (cells/mL), *n* (%) | 24442 (90.6) | 10937 (44.7) | 13505 (55.3) | $p < 0.001$ |
| | 318.0 (151.0–513.0) | 332.0 (160.0–518.0) | 306.0 (147.0–508.5) | |
| LP | 13390 (54.8) | 5776 (52.8) | 7614 (56.4) | $p < 0.001$ |
| NLP | 11052 (45.2) | 5161 (47.2) | 5891 (43.6) | |
| Viral Load at diagnosis (log10 copies/mL), *n* (%), IQR | 14005 (51.9) | 4589 (32.8) | 9416 (67.2) | $p < 0.001$ |
| | 4.3 (3.3–5.0) | 4.6 (3.8–5.3) | 4.1 (3.2–4.9) | |
| ≤4.0 | 5814 (41.5) | 1410 (30.7) | 4404 (46.8) | $p < 0.001$ |
| 4.1–5.0 | 4573 (32.7) | 1580 (34.4) | 2993 (31.8) | |
| ≥5.1 | 3618 (25.8) | 1599 (34.8) | 2019 (21.4) | |

## Transmitted and Acquired Drug Resistance

The overall prevalence of TDR was 12.8% (95%CI: 12.2–13.4%). NRTI, NNRTI and PI TDR were detected in 8.2% (95%CI: 7.7–8.7%), 5.6% (95%CI: 5.2–6.0%) and 3.7% (95%CI: 3.4–4.1%) of ART-naïve patients, respectively. 9.1% (95%CI: 8.6–9.7%) of these patients presented single class resistance, 2.8% (95%CI: 2.5–3.1%) presented dual class resistance and 0.9% (95%CI: 0.8–1.1%) presented triple class resistance (**Table 2**).

68.5% (95%CI: 67.8–69.2%) of experienced patients presented ADR, with higher drug resistance mutations for NRTI (59.1%; 95%CI: 58.3–59.8%), followed by NNRTI (42.2%; 95%CI: 41.4–43.0%) and by PI (24.2%; 95%CI: 23.5–24.9%). 23.5% (95%CI: 22.8–24.2%) of ART-experienced patients presented single class resistance, 33.0% (95%CI: 32.3–33.8%) presented dual class resistance and 12.0% (95%CI: 11.5–12.5%) presented triple class resistance (**Table 2**).

TDR presented an overall decreasing trend between 1995 and 2019 (*p* for trend < 0.001; **Table 2** and **Supplementary Data**). The same decreasing trend for TDR was observed for NRTIs, NNRTIs and PIs drug classes (*p* for trend < 0.001; **Table 2**). TDR between three time-periods (1995–2002; 2003–2010, and 2011–2019) was analyzed and it was observed that the overall TDR decreased from 20.0% to 13.3% to 10.7%. The same happened for every drug class, PIs (8.2% to 3.8% to 2.7% for the three time-periods, respectively), NRTIs (17.0% to 8.9% to 5.4% for the three time-periods, respectively) and for the NNRTIs (8.1% to 6.0% to 4.4% for the three time-periods, respectively). Moreover, between the 2003–2010 time-period, the overall TDR had a statistically significant decreasing trend (OR = 0.87; *p* = 0.001; **Figure 1A**).

**TABLE 2 |** Proportion of transmitted drug (TDR) and acquired drug resistance (ADR) between 1991 and 2019.

| | Transmitted drug resistance (TDR) | | | Acquired drug resistance (ADR) | | |
|---|---|---|---|---|---|---|
| | ***n* (%)** | **95% CI** | ***p* for trend** | ***n* (%)** | **95% CI** | ***p* for trend** |
| Total | 11581 (100) | | | 15392 (100) | | |
| Any DRMs | 1482 (12.8) | 12.2–13.4 | <0.001 | 10543 (68.5) | 67.8–69.2 | <0.001 |
| NRTI resistance | 944 (8.2) | 7.7–8.7 | <0.001 | 9089 (59.1) | 58.3–59.8 | <0.001 |
| NNRTI resistance | 644 (5.6) | 5.2–6.0 | <0.001 | 6499 (42.2) | 41.4–43.0 | <0.001 |
| PI resistance | 427 (3.7) | 3.4–4.1 | <0.001 | 3727 (24.2) | 23.5–24.9 | <0.001 |
| Single class resistance | 1056 (9.1) | 8.6–9.7 | 0.049 | 3617 (23.5) | 22.8–24.2 | <0.001 |
| Dual class resistance | 319 (2.8) | 2.5–3.1 | <0.001 | 5080 (33.0) | 32.3–33.8 | <0.001 |
| Triple class resistance | 107 (0.9) | 0.8–1.1 | <0.001 | 1846 (12.0) | 11.5–12.5 | <0.001 |
| PI + NRTI resistance | 115 (1.0) | 0.8–1.2 | <0.001 | 1671 (10.9) | 10.4–11.4 | <0.001 |
| PI + NNRTI resistance | 13 (0.1) | 0.07–0.2 | 0.452 | 63 (0.4) | 0.3–0.5 | 0.179 |
| NRTI + NNRTI resistance | 191 (1.6) | 1.4–1.9 | <0.001 | 3346 (21.7) | 21.1–22.4 | <0.001 |

*p-value for trend of TDR and ADR between 1995 and 2019. DRM, drug resistance mutations; NRTI, nucleotide reverse transcriptase inhibitors; NNRTI, non-nucleotide reverse transcriptase inhibitors; PI, protease inhibitors; CI, confidence interval.*

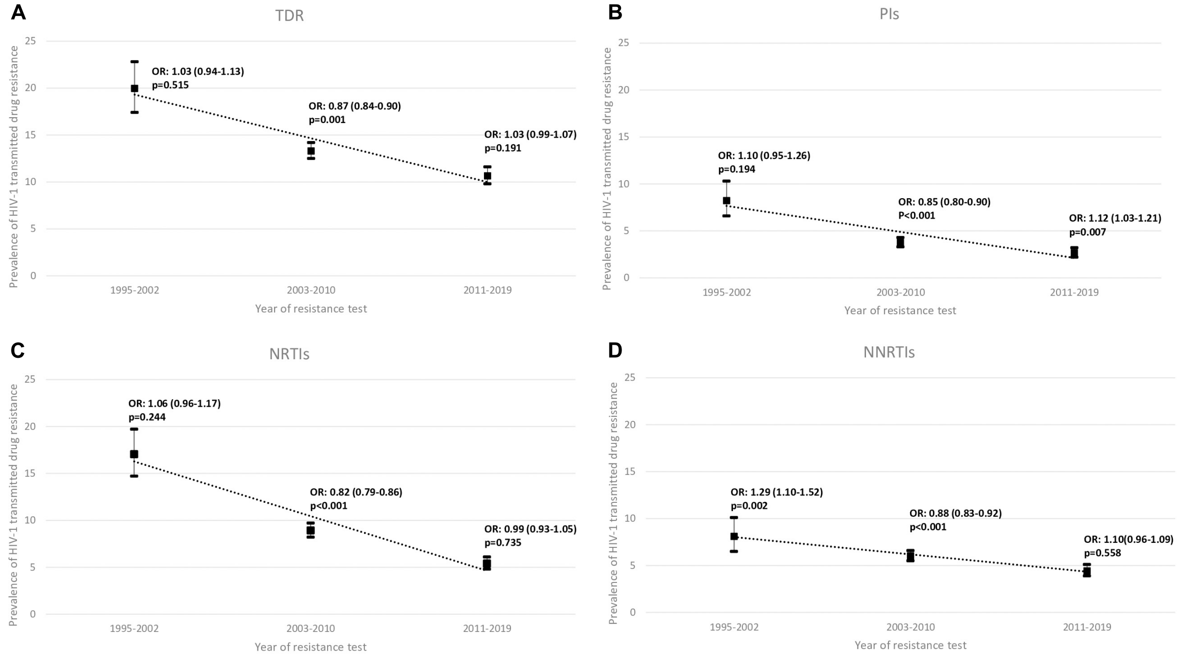


**FIGURE 1 |** Proportion of **(A)** overall transmitted drug resistance (TDR), **(B)** of protease inhibitors (PIs), **(C)** of nucleoside reverse transcriptase inhibitor (NRTIs) and **(D)** of non-nucleoside reverse transcriptase inhibitor (NNRTIs) in sequences from drug-naïve patients between three periods 1995–2002, 2003–2010, and 2011–2019. OR, Odds Ratio; *p*, *p*-value.

For the same time-period, the ARV drug classes also showed a decreasing trend, PI (OR = 0.85; $p < 0.001$), NNRTIs (OR = 0.82; $p < 0.001$) and NNRTIs (OR = 0.88; $p < 0.001$; **Figures 1A–D**).

Regarding the overall ADR trend, it has been decreasing over the three time-periods (80.0% to 70.7% to 44.5%) as well as in all drug classes studied except for NNRTIs (**Figure 1A**). PIs decreased from 36.3% to 24.8% to 5.9% and NRTIs decreased from 74.3% to 61.4% to 29.8%. Conversely, NNRTIs increased from 36.9% to 47.0% and then decreased to 31.4%. In the last time-period, 2011–2019, the overall ADR showed a decreasing trend (OR = 0.96; $p = 0.018$). The drug classes, in the same time-period, also showed a decreasing trend, but without being statistically significant PIs (OR = 0.94; $p = 0.092$), NRTIs (OR = 0.97; $p = 0.163$) and NNRTIs (OR = 0.98; $p = 366$; **Figure 2A–D**).

Differences in TDR and ADR prevalence between different countries included in this study were also analyzed between two time-periods (2008–2012 and 2013–2018). In our study population, in the first time-period (2008–2012), Luxembourg had the higher rate of TDR (16.8%). This scenario changed for TDR when the last time-period (2013–2018) was analyzed, since Germany (13.9%) presented the highest TDR rate. Comparing

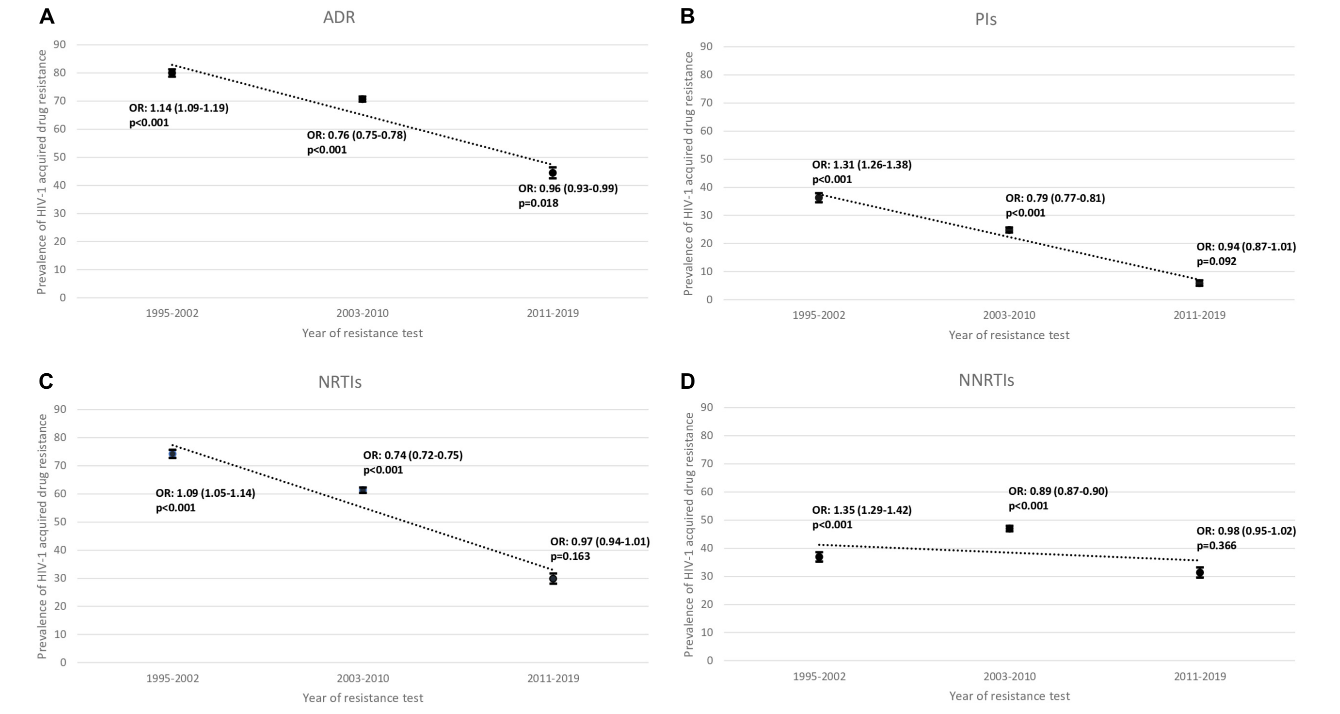


**FIGURE 2 |** Proportion of **(A)** overall acquired drug resistance (ADR), **(B)** of protease inhibitors (PIs), **(C)** of nucleoside reverse transcriptase inhibitor (NRTIs) and **(D)** of non-nucleoside reverse transcriptase inhibitor (NNRTIs) in sequences from drug-experienced patients between three periods 1995–2002, 2003–2010 and 2011–2019. OR, Odds Ratio; *p*, *p*-value.

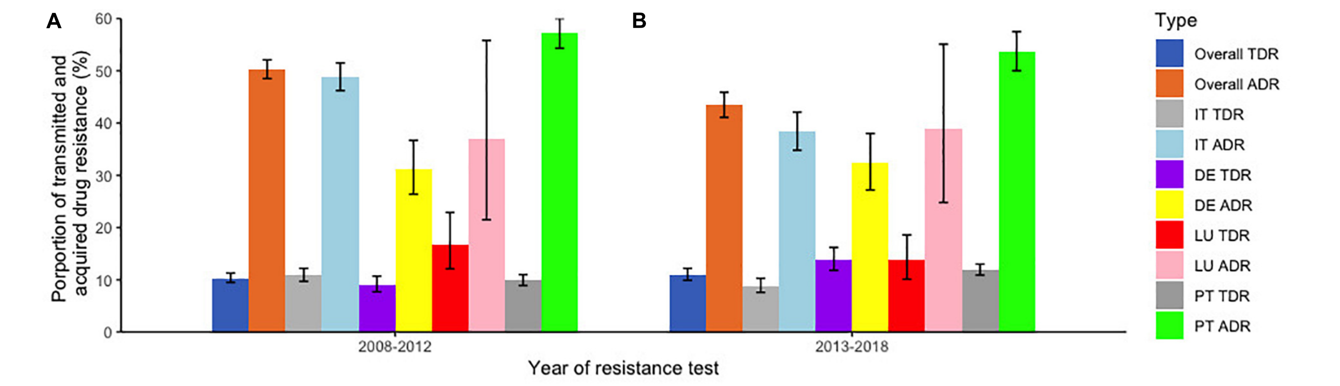


**FIGURE 3 |** Proportion of transmitted and acquired drug resistance per country of follow-up in two different time periods. **(A)** Between 2008 and 2012; **(B)** between 2013 and 2018. TDR, Transmitted drug resistance; ADR, Acquired drug resistance; IT, Italy; DE, Germany; LU, Luxembourg; PT, Portugal.

each country in those two time-periods, the TDR rate of Italy and Luxembourg decreased from one period to another (10.9% to 8.8%; 16.8% to 13.8%, respectively), while the rates of Germany and Portugal increased (9.9% to 11.9%; 9.1% to 13.9%, respectively).The ADR rates for the first time-period, indicated that all the countries, with the exception of Portugal (57.2%), presented a ADR lower than 50% (**Figure 3A**) and for the last time-period Portugal maintained the highest rate (53.7%; **Figure 3B**). Comparing the ADR rates between the same time-periods, the rate of Italy and Portugal decreased from one period to another (48.9% to 38.4%; 57.2% to 53.7%, respectively), while the rates of Germany and Luxembourg increased (31.3% to 32.4%; 37% to 38.9%, respectively; **Figure 3**).

## Transmitted and Acquired Drug Resistance Among Late Presenters and Non-late Presenters

Focusing now on the LP and NLP population, we observed a TDR of 12.3% (95%CI: 11.5–13.2) for LP population and 12.6% (95%CI: 11.8–13.6) for NLP population. In relation to drug resistance classes, the rates of resistance were higher in

**TABLE 3 |** Proportion of transmitted drug (TDR) and acquired drug resistance (ADR) in Late Presenters (LP) and Non-Late Presenters (NLP) between 1991 and 2019.

| Transmitted drug resistance (TDR) | Late presenters (LP) | | Non-late presenters (NLP) | |
|---|---|---|---|---|
| | ***n* (%)** | **95% CI** | ***n* (%)** | **95% CI** |
| Total | 5776 (100) | | 5161 (100) | |
| Any DRMs | 710 (12.3) | 11.5–13.2 | 652 (12.6) | 11.8–13.6 |
| NRTI resistance | 446 (7.7) | 7.1–8.4 | 428 (8.3) | 7.6–9.1 |
| NNRTI resistance | 317 (5.5) | 4.9–6.1 | 269 (5.2) | 4.6–5.9 |
| PI resistance | 202 (3.5) | 3.1–4.0 | 191 (3.7) | 3.2–4.3 |
| **Acquired drug resistance (ADR)** | | | | |
| Total | 7614 (100) | | 5891 (100) | |
| Any DRMs | 5319 (69.9) | 68.8–70.9 | 4016 (68.2) | 67.0–69.3 |
| NRTI resistance | 4588 (60.3) | 59.2–61.4 | 3538 (60.1) | 58.6–61.1 |
| NNRTI resistance | 3354 (44.1) | 42.9–45.2 | 2327 (39.5) | 38.3–40.8 |
| PI resistance | 2047 (26.9) | 25.9–27.9 | 1328 (22.5) | 21.5–23.6 |

*DRM, drug resistance mutations; NRTI, nucleotide reverse transcriptase inhibitors; NNRTI, non-nucleotide reverse transcriptase inhibitors; PI, protease inhibitors; CI, confidence interval.*

the NLP when compared to LPs, except for the NNRTIs class. LP presented higher rates of ADR—69.9% (95%CI: 68.8–70.9)—when compared to NLP: 68.2% (95%CI: 67.0–69.3). Contrary to TDR, the rates of ADR were higher in LP when compared to NLP (**Table 3**).

In both LP and NLP populations, the NNRTIs class K103N/S mutation presented the highest prevalence (3.1%; **Figure 4**). For PIs, M46I/L was more prevalent (1.5% for both LP and NLP) followed by L90M (1.4% for LP and 1.2% for NLP). Futhermore, in the PIs class there were two mutations present in LP (I47VA and V32I, respectively), that were not present in NLP (**Figure 4**). In the NLP, for NRTIs, we observed that M41I/L (3.2%) was the mutation with highest prevalence, followed by T215 revertants (3.0%) and by D67N/G/E and M184I/V (2.5%). Conversely, in the LP population, T215 revertants were more prevalent (3.2%), followed by M41I/L (2.4%) and M184I/V (2.3%).

Drug resistance mutations in ART-experienced patients in both LP and NLP populations were also analysed and compared (**Figure 5**). The more prevalent mutations consistently presented higher prevalences in LPs than in NLPs. Similarly to ART-naïve patients, for NNRTIs drug class, K103N/S mutation presented the highest prevalence (21.0% in LP and 19.0%, in NLP; **Figure 5**). For NRTIs, M184I/V had the highest prevalence (42.5% for LP and 41.7% for NLP). In the PIs class, the mutations with higher prevalence were L90M (11.8% NLP and 14.3% LP) and M46I/L (9.4% for NLP and 12.4% for LP). Also, K238TN mutation from the NNRTIs class was present only in the LP population. The presence of these mutations could lead to reduced susceptibility to some specific ARV.

## Analysis of Mutations Per Subtype Among Late Presenters and Non-late Presenters Patients

Finally, we compared mutations in LP and NLP, according to subtype B and non-B subtypes. As we can see in **Figure 6**, for subtype B ART-naïve patients, for both NRTIs and NNRTIs, most mutations—except T215rev—were more prevalent in NLP when compared to LP. K103N/S mutation was the one

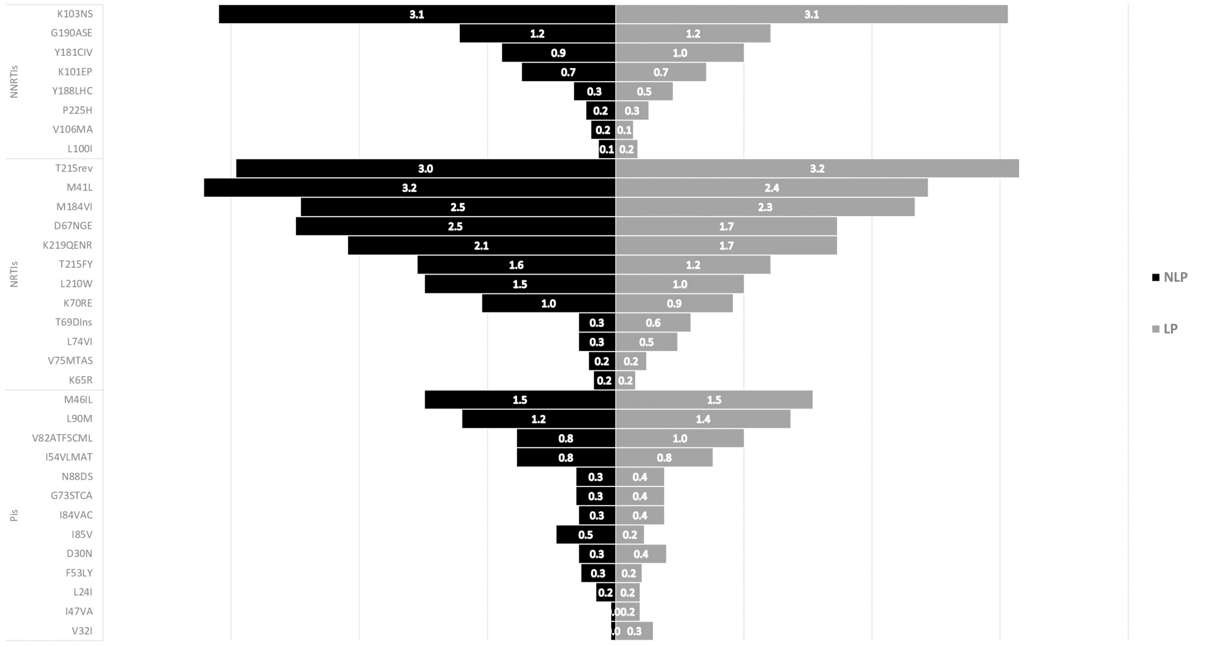


**FIGURE 4 |** ART-naïve mutations in Non-Late-Presenters (NLP) vs Late Presenters (LP). PIs, protease inhibitors; NRTIs, nucleoside reverse transcriptase inhibitor; NNRTIs, non-nucleoside reverse transcriptase inhibitor.

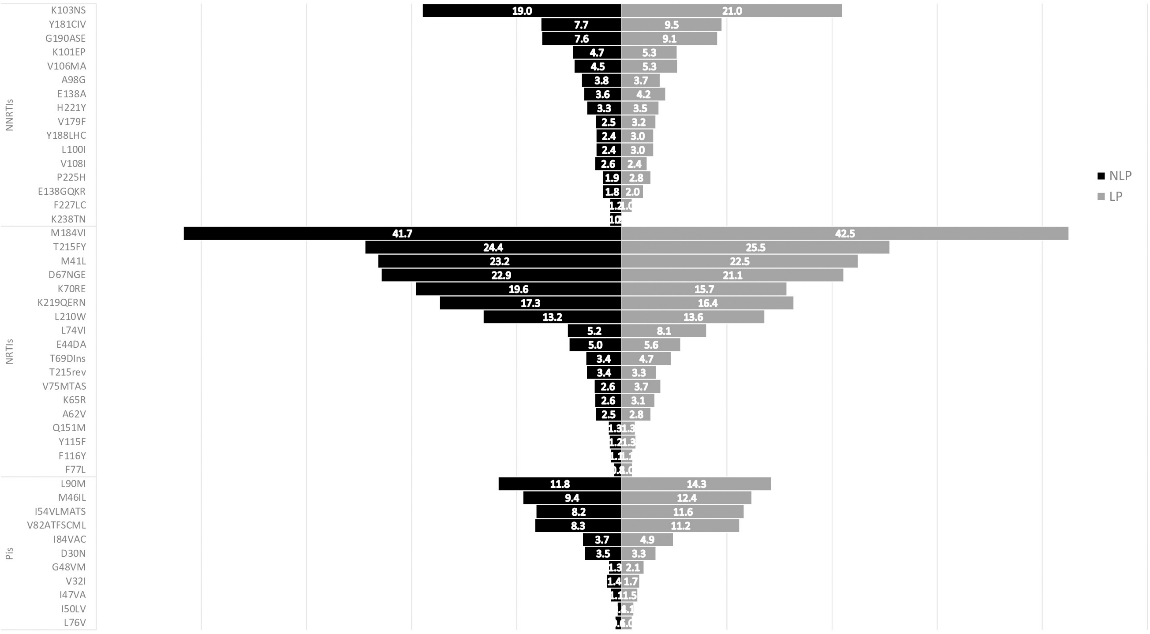


**FIGURE 5 |** ART-experienced mutations in Non-Late-Presenters (NLP) vs Late Presenters (LP). PIs, protease inhibitors; NRTIs, nucleoside reverse transcriptase inhibitor; NNRTIs, non-nucleoside reverse transcriptase inhibitor.

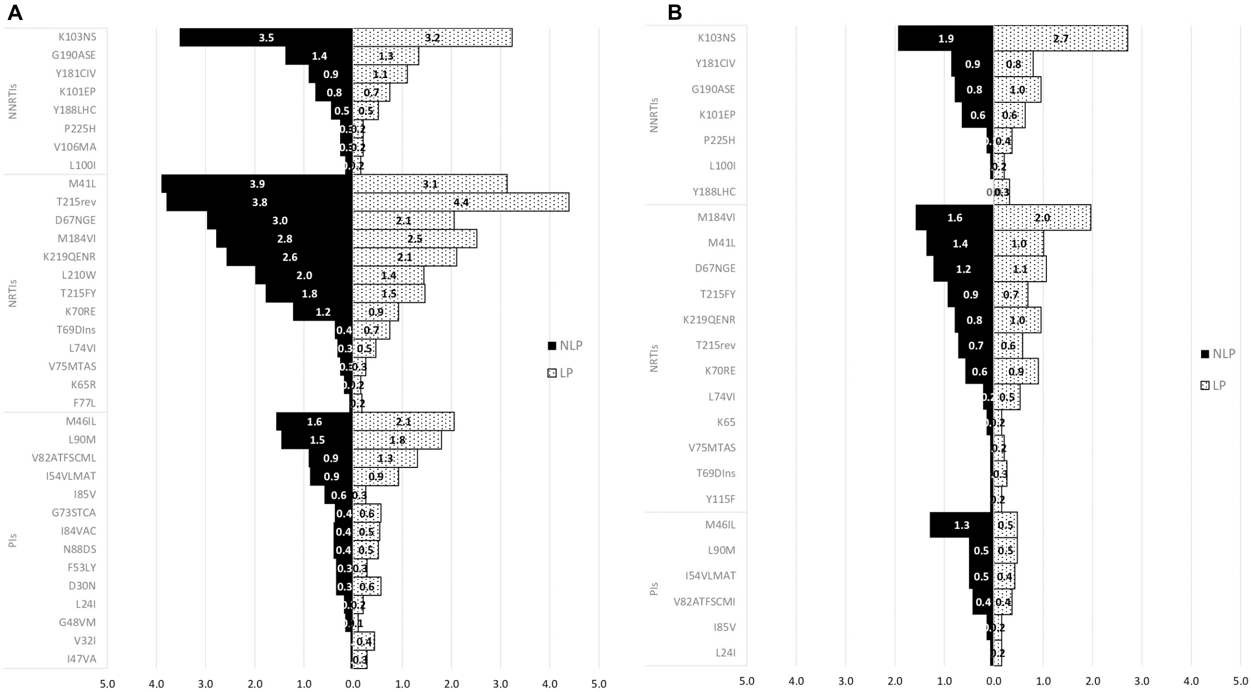


**FIGURE 6 |** Mutations in Non-Late presenters (NLP) and Late presenters (LP) in subtype B **(A)** and subtype non-B **(B)** for ART-naïve patients. PIs, protease inhibitors; NRTIs, nucleoside reverse transcriptase inhibitor; NNRTIs, non-nucleoside reverse transcriptase inhibitor.

with higher prevalence for NNRTIs (3.5% for NLP and 3.2% for LP). For NRTIs, M41L was the mutation with highest prevalence (3.9% for NLP vs 3.1% for LP), while for LP it was T215rev mutation (4.4% LP vs 3.8% NLP). For the PIs class, conversely, M46I/L and L90M were the mutations with the highest prevalence with higher prevalence in LP compared to

NLP (1.6 and 1.5% for NLP and 2.1 and 1.8% for LP, respectively; **Figures 6A,B**).

Regarding the non-B subtypes, K103N/S mutation was more prevalent in LP compared to NLP (2.7 vs 1.9%, respectively) which was the one with the highest prevalence. For NRTIs, M184V/I, M41L and D67NGE mutations (1.6, 1.4, and 1.2% for NLP and 2.0, 1.0, and 1.1 for LP, respectively) were the ones with higher prevalence. For PIs, M46I/L (1.3% for NLP and 0.5% for LP) was the one with the higher prevalence (**Figures 6A,B**). Comparing both populations regarding subtype non-B, opposite to what happens in subtype B, we observed that the LP population carried higher a prevalence of the most prevalent mutations (**Figures 6A,B**). Also, the K103N/S and the M184V/I were the mutations that were present in more non-B subtypes in the LP population, while the M46I/L was the one for the NLP populations. The most prevalent non-B subtype was subtype C (data not shown).

In ART-experienced patients, both in subtype B and in non-B subtypes, the most prevalent mutations occurred more frequently in LP than in NLP. For NNRTIs class K103N/S mutation had the highest prevalence in both NLP and LP (18.8 and 20.5%, respectively). For NRTIs the mutation with the highest prevalence was M184V/I mutation (43.6% for NLP and 43.9% for LP), and for PIs L90M and M46I/L were the mutations with the highest prevalence (12.7 and 10.4% for NLP and 16.7 and 14.2% for LP, respectively; **Figures 7A,B**).

Regarding the non-B subtypes, similiar to subtype B, K103N/S mutation (19.7% for NLP and 22.3% for LP) for NNRTIs, and M184I/V (33.1% for NLP and 38.8% LP) for NRTIs, were the ones with the highest prevalence. While in the PIs class, I54VLMATS (7.5% for NLP and 8.8% for LP) and L90M mutations (7.3% for NLP and 7.4% for LP) were the ones with the higher prevalence (**Figures 7A,B**). Also, M184V/I was the mutation that was present in the most diversity and proportion of non-B subtypes in both NLP and LP populations. The most prevalent non-B subtype was subtype G (data not shown).

## DISCUSSION

There are no recent studies with updated information regarding TDR and ADR prevalence in Europe and the most recent study about this topic only includes TDR and is based on the median overall values from different studies (Rhee et al., 2020). In our study, we presented updated information of the prevalence of TDR and ADR in the overall population and compared its patterns between LP and NLP. Overall, TDR had a prevalence of 12.8% and ADR of 68.5%. The TDR and ADR prevalence from our study was slightly higher when compared to other studies and this could be explained by the fact that our timeline includes patients diagnosed between 1981 and 2019 (Tostevin et al., 2017; Zazzi et al., 2018). Regarding the overall trends, both TDR and ADR presented a decreasing trend, consistently with other studies in and outside of Europe (Schmidt et al., 2014; Rocheleau et al., 2018).

We also compared TDR and ADR for the countries of follow-up included in the database divided into two time periods (2008–2012 and 2013–2018). For Italy, TDR prevalence decreased within time-periods (2008–2012:10.9% and 2013–2018: 8.8%), which is in accordance with studies from that country and around the same timeline (Franzetti et al., 2018; Rossetti et al., 2018). The prevalence of ADR also decreased

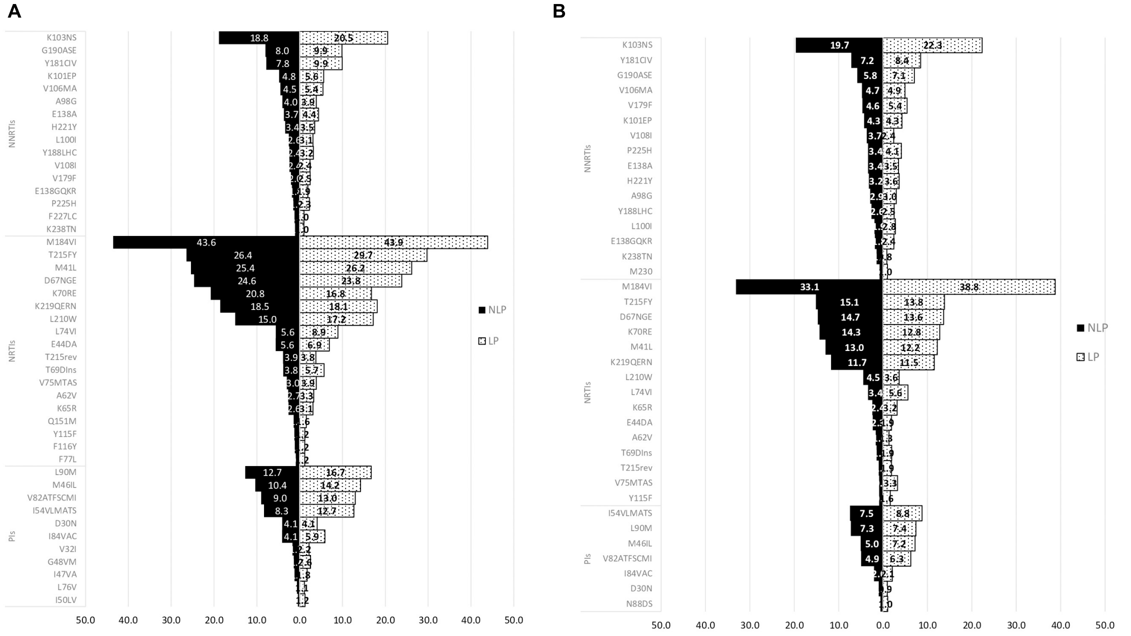


**FIGURE 7 |** Mutations in Non-Late presenters (NLP) and Late presenters (LP) in subtype B **(A)** and subtype non-B **(B)** for ART-experienced patients. PIs, protease inhibitors; NRTIs, nucleoside reverse transcriptase inhibitor; NNRTIs, non-nucleoside reverse transcriptase inhibitor.

in Italy (2008–2012: 48.9% and 2013–2018; 38.4%), and these results are slightly lower than those from a study from the Italian ARCA database. Moreover, the decrease in the last 5 years is in accordance with that study (Lombardi et al., 2021). For Germany, TDR prevalence was 9.1% and ADR prevalence was 31.3% between 2008 and 2012, and for a similar time-period, the TDR rate was around the same, but our ADR rate was lower than in another study reported in this country (Schmidt et al., 2014). For Luxembourg, the TDR prevalence was 16.8% and the ADR prevalence was 37% between 2008 and 2012, which is higher when compared to the values in Europe (Hofstra et al., 2016). For Portugal, TDR prevalence increased between time-periods (2008–2012:9.9% and 2013–2018: 11.9%), while ADR prevalence decreased between the same time-periods (2008–2012: 57.2% and 2013–2018: 53.7%). The TDR prevalence in the first time-period was closer to the one from a study conducted in Portugal between 2001 and 2017 and that same study indicated an increase trend for TDR. Our ADR prevalence for Portugal in the first time-period, had a lower value than the overall ADR prevalence from that study, although the decreasing trend was concordant (Pingarilho et al., 2020).

We also compared drug resistance in LP vs NLP, both in ART-naïve or ART-experienced patients. There were no major differences in the prevalence of drug resistance mutations in both LP and NLP from the ART-naïve population. However, LPs presented a lower prevalence of TDR than NLP, potentially suggesting a reversion of these mutations when patients are diagnosed late. The most prevalent mutations were the K103N/S, T215 revertants, the M184V/I, the M41I/L, the M46I/L and the L90M. However, in the LP, there were two mutations—I47V/A and V32I—that were not present in the NLP. Despite the lack of significance of these findings, we were not expecting to find mutations occurring specifically in late presenters, that could eventually indicate the irreversible fixation of these mutations in some cases, where they are not associated with a fitness cost (Winand et al., 2015; Nagaraja et al., 2016). In the ART-experienced population, there were also no significant differences between the LP and NLP populations, however, LPs presented a higher prevalence of ADR compared to NLP. The most prevalent mutations among LP and NLP were the K103N/S, the M184IV/I, the L90M and M46I/L. The K103N/S mutation presented similar prevalence in LP and NLP in ART-naïve, while ART-experienced LP had higher prevalence compared to NLP (Hiv Drug Resistance Database). T215rev in drug naïve patients was more prevalent in LP compared to NLP.The NRTIs T215rev mutants is associated with risk of virological failure to zidovidine (AZT) or stavudine (d4T). M41I/L impacts negatively virological response to regimens with abacavir (ABC), didanosine (ddl) or tenofovir (TDF). Together, these mutations confer high-level resistance to AZT and d4T. For the same drug class, M184V/I mutation reduces susceptibility to lamivudine (3TC) and emtricitabine (FTC; Hiv Drug Resistance Database). PI mutations were consistently more prevalent in LP compared to NLP, both in experienced and naïve patients, indicating a potential irreversible fixation of these mutations when they occur. The most prevalent were M46I/L which is associated with a reduction in the susceptibility to atazanavir (ATV), fosamprenavir (FPV), indinavir (IDV), lopinavir (LPV) and NFV, and L90M which is associated to reduced susceptibility to almost all PIs, except for tipranavir (TPV) and darunavir (DRV; Hiv Drug Resistance Database).

It is known that some mutations are closely related to specific subtypes and recombinant forms. As such, we conducted a final analysis distinguishing the patterns found in subtype B when compared to non-B subtypes. The most prevalent subtype was subtype B and the mutation with the highest prevalence in NLP ART-naïve patients was M41L from the NRTIs drug class. This result is in accordance with a study of mutations according to subtypes in Brazil (Westin et al., 2011).

In the LP and NLP patients, in the ART-experienced population, for both subtypes B and non-B, M184V/I mutation was the one with the higher prevalence.

This study was the first to analyze and compare transmitted and ADR in LP and NLP populations. Despite the lack of significant differences, we consistently found higher levels of TDR in NLP and higher levels of ADR in LP. We find this pattern consistent, except for non-B subtypes and the PIs class. This suggests different dynamics of reversion and irreversible fixation of mutations that should be further investigated in future studies.

## Limitations

Our study had some limitations. For example, concerning the analysis time-period, the first years and the more recent ones can be a bias in the analysis, since the number of individuals of those years is low compared to other years of resistance test collection date. Also, our population is mainly from Western Europe, providing a certain imbalance when characterizing the population and the TDR and ADR origins regarding geographical distribution. Another limitation of our study is the definition of LP as there is lack of consensus as to whether this definition ("baseline CD4 count in newly diagnosed patient is lower than 350 cells/mm$^3$ or has an AIDS-defining event, regardless of CD4 cell count") is the correct one to characterize those who present late to diagnosis. Some discuss that the threshold should be CD4 count lower than 200 cells/mm$^3$, i.e., those characterized in LP with advanced disease.

## CONCLUSION

In conclusion, our study showed that the overall TDR and ADR had a decreasing trend and the prevalence has been steady through the years. There were no significant differences in the TDR rate between the LP and NLP (around 12% in both), with slightly higher levels in the NLP. The mutation profile was also similar, again with most mutations presenting a higher prevalence of TDR in NLP and higher prevalence of ADR in LP. Late presentation for HIV remains a key unresolved challenge in HIV/AIDS with serious adverse consequences at the individual and societal levels. Our study highlights ADR and TDR patterns and drug resistance mutations, alone and according to subtypes in the LP population, when compared to NLP.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

MNSM, MP, and AA: conceptualization. MNSM, MP, VP, MdROM, and AA: methodology. MNSM, VP: software. MNSM, MP, FI, and AA: validation. MNSM, VP, MP, and AA: formal analysis. MNSM, MP, VP, and MdROM: investigation. CS-D, RP, RK, MZ, and FI: resources. CS-D, RP, RK, MZ, and FI: data curation. MNSM, MP, and AA: writing—original draft preparation. MNSM, MP, FI, and AA: writing—review, and editing. MNSM, MP, VP, MdROM, and AA: visualization. AA: supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was financially supported by FCT through the following projects: GHTM (UID/04413/2020), INTEGRIV (PTDC/SAU-INF/31990/2017) and the scholarship (PD/BD/135714/2018), and Gilead Génese HIVLatePresenters.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.846943/full#supplementary-material

## REFERENCES

Andersson, E., Shao, W., Bontell, I., Cham, F., Cuong, D. D., Wondwossen, A., et al. (2013). Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. *Infect. Genet. Evol.* 18, 125–131. doi: 10.1016/j.meegid.2013.03.050

Antinori, A., Coenen, T., Costagiola, D., Dedes, N., Ellefson, M., Gatell, J., et al. (2011). Late presentation of HIV infection: a consensus definition. *HIV Med.* 12, 61–64. doi: 10.1111/j.1468-1293.2010.00857.x

Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., et al. (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* 4:e4724. doi: 10.1371/journal.pone.0004724

Cdc *Treatment | Living with HIV | HIV Basics | HIV/AIDS | CDC*. Available online at: https://www.cdc.gov/hiv/basics/livingwithhiv/treatment.html. [accessed on May 20, 2021].

Clutter, D. S., Jordan, M. R., Bertagnolio, S., and Shafer, R. W. (2016). HIV-1 drug resistance and resistance testing. *Infect. Genet. Evol.* 46, 292–307. doi: 10.1016/j.meegid.2016.08.031

EuResist *Euresist Data Analysis - database*. Available online at: http://engine.euresist.org/database/. [accessed on Jan 04, 2021].

Franzetti, M., De Luca, A., Ceccherini-Silberstein, F., Spagnuolo, V., Nicastri, E., Mussini, C., et al. (2018). Evolution of HIV-1 transmitted drug resistance in Italy in the 2007–2014 period: A weighted analysis. *J. Clin. Virol.* 106, 49–52. doi: 10.1016/j.jcv.2018.07.009

Günthard, H. F., Calvez, V., Paredes, R., Pillay, D., Shafer, R. W., Wensing, A. M., et al. (2019). Human Immunodeficiency Virus Drug Resistance: 2018 Recommendations of the International Antiviral Society–USA Panel. *Clin. Infect. Dis.* 68, 177–187. doi: 10.1093/cid/ciy463

Hiv Drug Resistance Database *NNRTI Resistance Comments - HIV Drug Resistance Database*. Available online at: https://hivdb.stanford.edu/dr-summary/comments/NNRTI/. [accessed on February 22, 2021].

Hofstra, L. M., Sauvageot, N., Albert, J., Alexiev, I., Garcia, F., Struck, D., et al. (2016). Transmission of HIV drug resistance and the predicted effect on current first-line regimens in Europe. *Clin. Infect. Dis.* 62, 655–663. doi: 10.1093/cid/civ963

Joint United Nations Programme on Hiv/Aids (Unaids) (2015). *Understanding Fast-Track Targets: accelerating action to end the AIDS epidemic by 2030*. (Geneva: UNAIDS).

Lawyer, G., Altmann, A., Thielen, A., Zazzi, M., Sönnerborg, A., and Lengauer, T. (2011). HIV-1 mutational pathways under multidrug therapy. *AIDS Res. Ther.* 8:26. doi: 10.1186/1742-6405-8-26

Lombardi, F., Giacomelli, A., Armenia, D., Lai, A., Dusina, A., Bezenchek, A., et al. (2021). Prevalence and factors associated with HIV-1 multi-drug resistance over the past two decades in the Italian ARCA database. *Int. J. Antimicrob. Agents* 57:106252. doi: 10.1016/j.ijantimicag.2020.106252

Miranda, M. N. S., Pingarilho, M., Pimentel, V., Martins, M. D. R. O., Vandamme, A.-M., Bobkova, M., et al. (2021). Determinants of HIV-1 Late Presentation in Patients Followed in Europe. *Pathogens* 10:835. doi: 10.3390/pathogens10070835

Nagaraja, P., Alexander, H. K., Bonhoeffer, S., and Dixit, N. M. (2016). Influence of recombination on acquisition and reversion of immune escape and compensatory mutations in HIV-1. *Epidemics* 14, 11–25. doi: 10.1016/j.epidem.2015.09.001

Nih *Drug-Resistance Testing | NIH*. Available online at: https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv/drug-resistance-testing. [accessed on October 25, 2018].

Pineda-Peña, A. C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., et al. (2013). Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* 19, 337–348. doi: 10.1016/j.meegid.2013.04.032

Pingarilho, M., Pimentel, V., Diogo, I., Fernandes, S., Miranda, M., Pineda-Pena, A., et al. (2020). Increasing Prevalence of HIV-1 Transmitted Drug Resistance in Portugal: Implications for First Line Treatment Recommendations. *Viruses* 12:1238. doi: 10.3390/v12111238

Rhee, S., Kassaye, S. G., Barrow, G., Sundaramurthi, J. C., Jordan, M. R., and Shafer, R. W. (2020). HIV-1 transmitted drug resistance surveillance: shifting trends in study design and prevalence estimates. *J. Int. AIDS Soc.* 23:e25611. doi: 10.1002/jia2.25611

Rocheleau, G., Brumme, C. J., Shoveller, J., Lima, V. D., and Harrigan, P. R. (2018). Longitudinal trends of HIV drug resistance in a large Canadian cohort, 1996–2016. *Clin. Microbiol. Infect.* 24, 185–191. doi: 10.1016/j.cmi.2017.06.014

Rossetti, B., Di Giambenedetto, S., Torti, C., Postorino, M. C., Punzi, G., Saladini, F., et al. (2018). Evolution of transmitted HIV-1 drug resistance and viral subtypes circulation in Italy from 2006 to 2016. *HIV Med.* 19, 619–628. doi: 10.1111/hiv.12640

Schmidt, D., Kollan, C., Fätkenheuer, G., Schülter, E., Stellbrink, H. J., Noah, C., et al. (2014). Estimating trends in the proportion of transmitted and acquired HIV drug resistance in a long term observational cohort in Germany. *PLoS One* 9:e104474. doi: 10.1371/journal.pone.0104474

Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C., and Bercoff, D. P. (2014). COMET: Adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res.* 42:e144. doi: 10.1093/nar/gku739

Tostevin, A., White, E., Dunn, D., Croxford, S., Delpech, V., Williams, I., et al. (2017). Recent trends and patterns in HIV-1 transmitted drug resistance in the United Kingdom. *HIV Med.* 18, 204–213. doi: 10.1111/hiv.12414

Unaids *2025 AIDS TARGETS - UNAIDS*. Available online at: https://aidstargets2025.unaids.org/#section-targets. [accessed on Aug 03, 2021].

Westin, M. R., Biscione, F. M., Fonseca, M., Ordones, M., Rodrigues, M., Greco, D. B., et al. (2011). Resistance-Associated Mutation Prevalence According to Subtypes B and Non-B of HIV Type 1 in Antiretroviral-Experienced Patients in Minas Gerais, Brazil. *AIDS Res. Hum. Retroviruses* 27, 981–987. doi: 10.1089/aid.2010.0260

Winand, R., Theys, K., Eusébio, M., Aerts, J., Camacho, R. J., Gomes, P., et al. (2015). Assessing transmissibility of HIV-1 drug resistance mutations from treated and from drug-naive individuals. *AIDS* 29, 2045–2052. doi: 10.1097/QAD.0000000000000811

World Health Organization *Updated recommendations on first-line and second-line antiretroviral regimens and post-exposure prophylaxis and recommendations on early infant diagnosis of HIV*. (Geneva: World Health Organization).

World Health Organization *Global HIV Programme*. Available online at: https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics. [accessed on Aug 03, 2021].

Zazzi, M., Hu, H., and Prosperi, M. (2018). The global burden of HIV-1 drug resistance in the past 20 years,". *PeerJ* 6:e4848. doi: 10.7717/peerj.4848

Zazzi, M., Incardona, F., Rosen-Zvi, M., Prosperi, M., Lengauer, T., Altmann, A., et al. (2012). Predicting response to antiretroviral treatment by machine learning: The euresist project. *Intervirology* 55, 123–127. doi: 10.1159/000332008

Zou, X., He, J., Zheng, J., Malmgren, R., Li, W., Wei, X., et al. (2020). Prevalence of acquired drug resistance mutations in antiretroviral- experiencing subjects from 2012 to 2017 in Hunan Province of central South China. *Virol. J.* 17:38. doi: 10.1186/s12985-020-01311-3

**Conflict of Interest:** FI was employed by IPRO—InformaPRO S.r.l.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from Gilead Sciences. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fmicb.2022.823208

# HIV-1-Transmitted Drug Resistance and Transmission Clusters in Newly Diagnosed Patients in Portugal Between 2014 and 2019

**Marta Pingarilho[1]*, Victor Pimentel[1], Mafalda N. S. Miranda[1], Ana Rita Silva[2], António Diniz[3], Bianca Branco Ascenção[4], Carmela Piñeiro[5], Carmo Koch[6], Catarina Rodrigues[7], Cátia Caldas[5], Célia Morais[8], Domitília Faria[9], Elisabete Gomes da Silva[10], Eugénio Teófilo[11], Fátima Monteiro[6], Fausto Roxo[12], Fernando Maltez[13], Fernando Rodrigues[8], Guilhermina Gaião[14], Helena Ramos[15], Inês Costa[16], Isabel Germano[7], Joana Simões[7], Joaquim Oliveira[17], José Ferreira[18], José Poças[4], José Saraiva da Cunha[17], Jorge Soares[5], Júlia Henriques[16], Kamal Mansinho[19], Liliana Pedro[9], Maria João Aleixo[20], Maria João Gonçalves[21], Maria José Manata[13], Margarida Mouro[22], Margarida Serrado[3], Micaela Caixeiro[23], Nuno Marques[20], Olga Costa[24], Patrícia Pacheco[23], Paula Proença[25], Paulo Rodrigues[2], Raquel Pinho[9], Raquel Tavares[2], Ricardo Correia de Abreu[26], Rita Côrte-Real[24], Rosário Serrão[5], Rui Sarmento e Castro[21], Sofia Nunes[22], Telo Faria[10], Teresa Baptista[19], Maria Rosário O. Martins[1], Perpétua Gomes[16,28], Luís Mendão[27], Daniel Simões[27] and Ana Abecasis[1] on behalf of the BESTHOPE Study Group**

[1] Global Health and Tropical Medicine (GHTM), Instituto de Higiene e Medicina Tropical (IHMT), Universidade Nova de Lisboa (UNL), Lisbon, Portugal, [2] Serviço de Infeciologia, Hospital Beatriz Ângelo, Loures, Portugal, [3] Unidade de Imunodeficiência, Centro Hospitalar Universitário Lisboa Norte - HPV, Lisbon, Portugal, [4] Serviço de Infeciologia, Centro Hospitalar de Setúbal, Setúbal, Portugal, [5] Serviço de Doenças Infeciosas, Centro Hospitalar Universitário de São João, Porto, Portugal, [6] Centro de Biologia Molecular, Serviço de Imunohemoterapia do Centro Hospitalar Universitário de São João, Porto, Portugal, [7] Serviço de Medicina, Hospital de São José, Centro Hospitalar Universitário de Lisboa Central, Lisbon, Portugal, [8] Serviço de Patologia Clínica, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal, [9] Serviço de Medicina, Hospital de Portimão, Centro Hospitalar Universitário do Algarve, Portimão, Portugal, [10] Unidade Local de Saúde do Baixo Alentejo, Hospital José Joaquim Fernandes, Beja, Portugal, [11] Serviço de Medicina, Hospital de Santo António dos Capuchos, Centro Hospitalar Universitário de Lisboa Central, Lisbon, Portugal, [12] Hospital de Dia de Doenças Infeciosas, Hospital Distrital de Santarém, Santarém, Portugal, [13] Serviço de Doenças Infeciosas, Hospital de Curry Cabral, Centro Hospitalar Universitário de Lisboa Central, Lisbon, Portugal, [14] Serviço de Patologia Clínica, Hospital de Santa Maria, Centro Hospitalar Universitário de Lisboa Norte, Lisbon, Portugal, [15] Serviço de Patologia Clínica, Centro Hospitalar do Porto, Porto, Portugal, [16] Laboratório de Biologia Molecular (LMCBM, SPC, CHLO-HEM), Lisbon, Portugal, [17] Serviço de Doenças, Centro Hospitalar e Universitário de Coimbra, Coimbra, Portugal, [18] Serviço de Medicina, Hospital de Faro, Centro Hospitalar Universitário do Algarve, Faro, Portugal, [19] Serviço de Doenças Infeciosas, Hospital de Egas Moniz, Centro Hospitalar de Lisboa Ocidental, Lisbon, Portugal, [20] Serviço de Infeciologia, Hospital Garcia da Orta, Almada, Portugal, [21] Serviço de Infeciologia, Centro Hospitalar do Porto, Porto, Portugal, [22] Serviço de Infeciologia, Hospital de Aveiro, Centro Hospitalar Baixo Vouga, Aveiro, Portugal, [23] Serviço de Infeciologia, Hospital Professor Doutor Fernando da Fonseca, Amadora, Portugal, [24] Serviço de Patologia Clínica, Biologia Molecular, Centro Hospitalar Universitário de Lisboa Central, Lisbon, Portugal, [25] Serviço de Infeciologia, Hospital de Faro, Centro Hospitalar Universitário do Algarve, Faro, Portugal, [26] Serviço de Infeciologia, Unidade de Local de Saúde de Matosinhos, Hospital Pedro Hispano, Matosinhos, Portugal, [27] Grupo de Ativistas em Tratamentos (GAT), Lisbon, Portugal, [28] Centro de Investigação Interdisciplinar Egas Moniz (CiiEM), Instituto Universitário Egas Moniz, Costa da Caparica, Portugal

**Objective:** To describe and analyze transmitted drug resistance (TDR) between 2014 and 2019 in newly infected patients with HIV-1 in Portugal and to characterize its transmission networks.

OPEN ACCESS

***Edited by:***
*Joris Hemelaar,*
*University of Oxford, United Kingdom*

***Reviewed by:***
*Marta Giovanetti,*
*Oswaldo Cruz Foundation (Fiocruz), Brazil*
*Philippe Colson,*
*IHU Mediterranee Infection, France*
*Orna Mor,*
*Ministry of Health, Israel*

****Correspondence:***
*Marta Pingarilho*
*martapingarilho@ihmt.unl.pt*

***Specialty section:***
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *26 November 2021*
***Accepted:*** *21 March 2022*
***Published:*** *25 April 2022*

***Citation:***
*Pingarilho M, Pimentel V, Miranda MNS, Silva AR, Diniz A, Ascenção BB, Piñeiro C, Koch C, Rodrigues C, Caldas C, Morais C, Faria D, da Silva EG, Teófilo E, Monteiro F, Roxo F, Maltez F, Rodrigues F, Gaião G, Ramos H, Costa I, Germano I, Simões J, Oliveira J, Ferreira J, Poças J, da Cunha JS, Soares J, Henriques J, Mansinho K, Pedro L, Aleixo MJ, Gonçalves MJ, Manata MJ, Mouro M, Serrado M, Caixeiro M, Marques N, Costa O, Pacheco P, Proença P, Rodrigues P, Pinho R, Tavares R, de Abreu RC, Côrte-Real R, Serrão R, Castro RS, Nunes S, Faria T, Baptista T, Martins MRO, Gomes P, Mendão L, Simões D and Abecasis A (2022) HIV-1-Transmitted Drug Resistance and Transmission Clusters in Newly Diagnosed Patients in Portugal Between 2014 and 2019. Front. Microbiol. 13:823208. doi: 10.3389/fmicb.2022.823208*

**Methods:** Clinical, socioepidemiological, and risk behavior data were collected from 820 newly diagnosed patients in Portugal between September 2014 and December 2019. The sequences obtained from drug resistance testing were used for subtyping, TDR determination, and transmission cluster (TC) analyses.

**Results:** In Portugal, the overall prevalence of TDR between 2014 and 2019 was 11.0%. TDR presented a decreasing trend from 16.7% in 2014 to 9.2% in 2016 ($p_{for-trend}$ = 0.114). Multivariate analysis indicated that TDR was significantly associated with transmission route (MSM presented a lower probability of presenting TDR when compared to heterosexual contact) and with subtype (subtype C presented significantly more TDR when compared to subtype B). TC analysis corroborated that the heterosexual risk group presented a higher proportion of TDR in TCs when compared to MSMs. Among subtype A1, TDR reached 16.6% in heterosexuals, followed by 14.2% in patients infected with subtype B and 9.4% in patients infected with subtype G.

**Conclusion:** Our molecular epidemiology approach indicates that the HIV-1 epidemic in Portugal is changing among risk group populations, with heterosexuals showing increasing levels of HIV-1 transmission and TDR. Prevention measures for this subpopulation should be reinforced.

**Keywords: HIV-1, TDR, transmission clusters, Portugal, newly infected patients**

## INTRODUCTION

The "Treatment for All" program was implemented in many countries with an aim to offer treatment and care to anyone diagnosed with HIV, regardless of the stage of infection (CD4 cell count). In Portugal, this program was implemented in 2015 [World Health Organiztion (WHO), 2021]. The widespread use and increased coverage of antiretroviral therapy (ART) have reduced the risk of HIV transmission, decreased HIV-related morbidity and mortality, and improved life quality. However, treatment scale-up can potentiate the risk for the development of antiretroviral (ARV) drug resistance, which can be transmitted to newly infected individuals (Palella et al., 1998; Lima et al., 1999; Clavel and Hance, 2004; Cohen et al., 2011). TDR in HIV-1-infected patients has become a major concern as it may lead to the failure of first-line ART. There are several studies indicating that the prevalence of TDR is largely variable in different settings and risk groups and that this could be related to the differences in the availability of treatment and levels of socioeconomic development (Pennings, 2013; Frentz et al., 2014; Yang et al., 2015). For example, TDR levels are highly discrepant when we compare Germany (18.4%) (van de Laar et al., 2019)[9], Belgium (15.7%) (van de Laar et al., 2019)[9], Hungary (7.1%) (van de Laar et al., 2019)[9], Netherlands (12.3%) (van de Laar et al., 2019)[9], Mozambique (14.0% in women) (Pina-Araujo et al., 2014)[10], Latin America (7.7%) (Avila-Rios et al., 2016)[11], and Washington DC (20.0%) (Aldous et al., 2017)[12]. Portugal, on the other hand, presented an overall TDR of 9.4% between 2001 and 2017, with a significantly increasing trend from 7.9% in 2003 to 13.1% in 2017 (Pingarilho et al., 2020). These results were obtained in a retrospective study of our study group and were based only on RegaDB, a laboratory database including clinical, demographic, and genomic data of patients followed up in hospitals located in the southern region of Portugal.

As the HIV epidemic continues to spread, it is very important to investigate the prevalence and transmission of TDR over the years in individual settings/locations. Moreover, the phylogenetic analysis provides insight into the HIV transmission clusters (TCs). The characterization of HIV-1 transmission clusters and associated TDR allows for targeted interventions to individuals at higher risk.

In this study, we aim to describe TDR between 2014 and 2019 in newly diagnosed patients with HIV-1 in Portugal, characterize the most prevalent drug resistance mutations, and identify predictors of TDR in Portugal. Furthermore, we aim to characterize HIV-1 transmission clusters involving these patients.

## MATERIALS AND METHODS

### Study Population and Data Collection

The protocol was in accordance with the Declaration of Helsinki and approved by the Ethical Committee of all hospitals involved in the study.

Clinical, socioepidemiological, and risk behavior data were collected prospectively from 820 newly diagnosed patients from 17 hospitals located across the whole country from north to south of Portugal between September 2014 and December 2019. This sampling corresponds to a sampling rate of 17% of the total newly diagnosed cases in Portugal within these 5 years. The BEST HOPE database contains anonymized patients' information, including demographic, clinical, behavioral, and genotype resistance data. The data of all the patients were generated in the context of routine clinical care.

## Drug Resistance Analyses and Subtyping

The genomic data included protease and reverse transcriptase sequences obtained through population sequencing performed at the molecular biology laboratories of different hospitals during daily care routine analysis. The genomic sequences were obtained for all the patients at the time of diagnosis, before starting ARV therapy. TDR was defined as the presence of one or more surveillance drug resistance mutations (SDRMs) according to the WHO 2009 surveillance list (Bennett et al., 2009)[14]. Nucleotide sequences were submitted to the Calibrated Population Resistance tool version 8.0. Clinical resistance to ARV drugs was inferred using the Stanford HIVdb v8.4. HIV-1 subtypes and circulating recombinant forms (CRFs) were determined as previously described (Pineda-Peña et al., 2013; Struck et al., 2014)[15,16]. Sequence alignments and associated metadata are available from the authors upon request.

## Late Presenters and Late Presenters With Advanced Disease

According to the European Late Presenter Consensus working group, late presenters (LP) were defined as a CD4 count lower than 350 cells/$\mu$L at the time of diagnosis or present with an AIDS-defining event at diagnosis, regardless of the CD4 cell count. A subgroup of late presenters, called late presenters with advanced disease (LPAD), were characterized by presenting a CD4 count lower than 200 cells/$\mu$L or an AIDS-defining event, regardless of the CD4 cell count (Antinori et al., 2011). The groups of patients considered as LP or LPAD were stratified and analyzed according to this definition.

## Genetic Ambiguities

HIV-1 protease and reverse transcriptase sequences derived from standard genotyping methods were used to determine the recentness of infection, which was calculated based on the ambiguity rate of the genomic sequences. Chronic infection was defined as an ambiguity rate with a cut-off value higher than 0.45% and recent infection as an ambiguity rate with a cut-off value equal to or below 0.45% (Andersson et al., 2013).

## Transmission Cluster Identification

For the TC analysis, the dataset was divided into three separate datasets: subtypes B, A, and G. Control sequences were collected from the Los Alamos database and included all HIV-1 pol subtype (B, A, and G) sequences from Europe, South America, and Africa[1] (Kuiken et al., 2003). Three reference sequences (from subtypes B and C) were used as outgroup. The resulting dataset was aligned against the global background dataset selected as control using VIRULIGN (Libin et al., 2019). Sequences with low quality, duplicates, and clones were deleted. The sequence dataset for transmission cluster analysis included the sequences of 500 patients from subtypes B, A, and G (obtained from the BEST HOPE dataset) and the sequences included in the control dataset. The total number of sequences used in this analysis was 37,822 (333 seqs were from BEST HOPE and 37,489 seqs were from controls), 7,853 (78 seqs were from BEST HOPE and 7,775 seqs were from controls), and 2,254 (89 seqs were from BEST HOPE and 2,165 seqs were from controls) from subtypes B, A, and G, respectively, with a length of 947 nucleotides. Codon positions associated with drug resistance were removed from the alignment. Maximum likelihood (ML) phylogenies were constructed using FastTree with the generalized time-reversible model. Statistical support of clades was assessed using the Shimodaira-Hasegawa-like test (SH-test). Putative transmission clusters were identified using ClusterPicker v1.332 (Ragonnet-Cronin et al., 2013)[21] and were defined as clades with branch support $\geq$ 0.99 in the likelihood ratio test (aLRT), as implemented in ClusterPicker v1.332. The clusters were categorized based on the size as large clusters (comprising eight patients or more) and small clusters (comprising less than eight patients). The origin of transmission clusters was assigned if at least 66% of the sequences in the cluster corresponded to the same sampling country. If there was no consistent sampling country for at least 66% of the sequences in the cluster, no origin was assigned.

## Statistical Analysis

Proportions and confidence intervals for proportions were calculated using a 95% Wilson confidence interval for binomially distributed data. The differences between the prevalence of resistance in naïve patients were analyzed using the Mann–Whitney $U$ test and the $X^2$ tests. Logistic regression was used to examine the association between demographic and clinical factors and the occurrence of SDRMs, and to analyze the trends over time. For all the statistical analyses, we considered a 5% significance level. All the analyses were conducted in SPSS Statistic version 25 software and R3.5.1.

# RESULTS

## Epidemiological and Clinical Data

The characteristics of the study population are presented in **Table 1**. More than half (77.3%) of the patients ($n$ = 820) included in the database were men. The median age at diagnosis was 37.0 (IQR: 29.0–47.0) years. During this period (2014–2019), new diagnoses occurred predominantly (56.0%) between the ages of 22 and 40 years. The main modes of transmission were heterosexual and homosexual contact (49.8% and 47.3%, respectively), followed by intravenous drug use (1.8%). Most patients were born in Portugal (71.9%), and patients born abroad (13.9%) originated from Portuguese-Speaking African Countries (PSAC) and Brazil (11.2%). Individuals with a higher level of education had a higher prevalence when compared to the individuals with other levels of education (35.1%). Most patients (74.4%) were employed; however, 44.2% of them considered the current income as insufficient. More than half (68.5%) were single, 20.9% were married, and 8.3% were divorced and widowers (0.5%). About 75.5% of men reported having

[1] http://www.hiv.lanl.gov

**TABLE 1 |** Demographic and patients characteristics.

| Patient characteristics | Total | With TDR | Without TDR | *p*-value |
|---|---|---|---|---|
| Total, *n* (%) | 820 (100%) | 89 (10.9%) | 729 (88.9%) | |
| Gender, *n* (%) | 816 (99.5%) | 90 (98.9%) | 726 (99.6%) | |
| Male | 631 (77.3%) | 75 (83.3%) | 556 (76.6%) | 0.149 |
| Female | 185 (22.7%) | 15 (16.7%) | 170 (23.4%) | |
| Median age at diagnosis in years IQR, *n* (%) | 811 (98.9%) | 89 (97.8%) | 722 (99.0%) | |
| | 37.0 (29.0–47.0) | 36.0 (29.5–46.5) | 37.0 (29.0–47.0) | 0.651 |
| 15–21 | 34 (4.2%) | 4 (4.4%) | 30 (4.1%) | 0.668 |
| 22–40 | 454 (56.0%) | 54 (60.7%) | 400 (55.4%) | |
| 41–55 | 231 (28.5%) | 24 (27.0%) | 207 (28.7%) | |
| ≥ 56 | 92 (11.3%) | 7 (7.9%) | 85 (11.8%) | |
| Transmission route, *n* (%) | 811 (98.9%) | 90 (98.9%) | 721 (98.9%) | |
| Heterosexual | 404 (49.8%) | 54 (60.0%) | 350 (48.7%) | 0.093 |
| MSM** | 384 (47.3%) | 33 (36.7%) | 351 (48.7%) | |
| IDU | 15 (1.8%) | 1 (1.1%) | 14 (1.9%) | |
| Other | 8 (1.0%) | 2 (2.2%) | 6 (0.8%) | |
| Country of origin, *n* (%) | 807 (98.4%) | 91 (100%) | 716 (99.2%) | |
| Portugal | 580 (71.9%) | 67 (73.6%) | 513 (71.6%) | 0.370 |
| Brazil | 90 (11.2%) | 11 (12.1%) | 79 (11.0%) | |
| Guinea-Bissau | 46 (5.7%) | 2 (2.2%) | 44 (6.1%) | |
| Angola | 34 (4.2%) | 6 (6.6%) | 28 (3.9%) | |
| Cabo-Verde | 24 (3.0%) | 1 (1.1%) | 23 (3.2%) | |
| Mozambique | 8 (1.0%) | 0 (0.0%) | 8 (1.1%) | |
| Others | 25 (3.1%) | 4 (4.4%) | 21 (2.9%) | |
| Region of origin, *n* (%) | 815 (99.4%) | 91 (100%) | 724 (99.3%) | |
| Europe | 598 (73.4%) | 70 (76.9%) | 528 (72.9%) | 0.578 |
| Africa | 120 (14.7%) | 9 (9.9%) | 112 (15.3%) | |
| South America | 90 (11.0%) | 11 (12.1%) | 79 (10.9%) | |
| Other | 7 (0.9%) | 1 (1.1%) | 6 (0.8%) | |
| District of residence, *n* (%) | 716 (87.3%) | 81 (89.0%) | 635 (87.1%) | |
| Lisboa | 287 (40.1%) | 30 (37.0%) | 257 (40.5%) | 0.326 |
| Porto | 155 (21.6%) | 17 (21.0%) | 138 (21.7%) | |
| Faro | 82 (11.5%) | 9 (11.1%) | 73 (11.5%) | |
| Setúbal | 79 (11.0%) | 6 (7.4%) | 73 (11.5%) | |
| Aveiro | 40 (5.6%) | 8 (9.9%) | 32 (5.0%) | |
| Beja | 13 (1.6%) | 2 (2.5%) | 11 (1.7%) | |
| Coimbra | 15 (2.1%) | 4 (4.9%) | 11 (1.7%) | |
| Outro | 45 (6.3%) | 5 (6.2%) | 40 (6.3%) | |
| Migrant status, *n* (%) | 815 (99.4%) | 91 (100%) | 724 (99.3%) | |
| Migrant | 235 (28.8%) | 24 (26.4%) | 211 (29.1%) | 0.582 |
| Native | 580 (71.2%) | 67 (73.6%) | 513 (70.9%) | |
| School level, *n* (%) | 444 (54.1%) | 39 (42.9%) | 405 (55.6%) | |
| Third level (9th degree) | 98 (22.1%) | 12 (30.8%) | 86 (21.2%) | |
| Secondary (12th degree) | 143 (32.2%) | 12 (30.8%) | 131 (32.3%) | 0.576 |

*(Continued)*

**TABLE 1 |** (Continued)

| Patient characteristics | Total | With TDR | Without TDR | *p*-value |
|---|---|---|---|---|
| Advanced Technical Specialization | 41 (9.2%) | 2 (5.1%) | 39 (9.6%) | |
| Higher education (bachelor, master, PhD) | 156 (35.1%) | 13 (33.3%) | 143 (35.3%) | |
| None | 6 (1.4%) | 0 | 6 (1.5%) | |
| Current occupation, *n* (%) | 433 (52.8%) | 39 (42.9%) | 394 (54.0%) | |
| Employed | 322 (74.4%) | 27 (69.2%) | 295 (74.9%) | |
| Retired | 11 (2.5%) | 1 (2.6%) | 10 (2.5%) | 0.033 |
| Sex-worker | 1 (0.2%) | 1 (2.6%) | 0 | |
| Student | 24 (5.5%) | 2 (5.1%) | 22 (5.6%) | |
| Unemployed | 75 (17.3%) | 8 (20.5%) | 67 (17.0%) | |
| Current income, *n* (%) | 405 (49.4%) | 34 (37.4%) | 371 (50.9%) | |
| Very Insufficient | 68 (16.8%) | 7 (20.6%) | 61 (16.4%) | |
| Insufficient | 179 (44.2%) | 15 (44.1%) | 164 (44.2%) | 0.914 |
| Sufficient | 141 (34.8%) | 11 (32.4%) | 130 (35.0%) | |
| More than sufficient | 17 (4.2%) | 1 (2.9%) | 16 (4.3%) | |
| Civil status, *n* (%) | 444 (54.1%) | 40 (44.0%) | 404 (55.4%) | |
| Single | 304 (68.5%) | 22 (55.0%) | 282 (69.8%) | |
| Married | 93 (20.9%) | 11 (27.5%) | 82 (20.3%) | 0.066 |
| Divorced | 37 (8.3%) | 4 (10.0%) | 33 (8.2%) | |
| Widower | 2 (0.5%) | 1 (2.5%) | 1 (0.2%) | |
| Other | 8 (1.8%) | 2 (5.0%) | 6 (1.5%) | |
| Men sexual partners, *n* (%) | 379 (60.1%) | 35 (46.7%) | 344 (61.9%) | |
| Men*** | 286 (75.5%) | 26 (74.3%) | 260 (75.6%) | 0.452 |
| Women | 44 (11.6%) | 6 (17.1%) | 38 (11.0%) | |
| Men and women | 49 (12.9%) | 3 (8.6%) | 46 (13.4%) | |
| Women sexual partners, *n* (%) | 56 (100.0%) | 5 (100.0%) | 51 (100.0%) | |
| Men | 56 (100%) | 5 (100.0%) | 51 (100.0%) | - |
| Women | 0 | 0 | 0 | |
| Clinical* | | | | |
| Type of infection, *n* (%) | 820 (100%) | 91 (100%) | 729 (100%) | |
| Chronic | 421 (51.3%) | 54 (59.3%) | 367 (50.3%) | 0.105 |
| Recent | 399 (48.7%) | 37 (40.7%) | 362 (49.7%) | |
| Infection stage, *n* (%) | 784 (95.6%) | 86 (94.5%) | 698 (95.7%) | |
| A | 549 (70.0%) | 51 (59.3%) | 498 (71.3%) | 0.034 |
| B | 93 (11.9%) | 11 (12.8%) | 82 (11.7%) | |
| C | 142 (18.1%) | 24 (27.9%) | 118 (16.9%) | |
| Aids-defining event, *n* (%) | 788 (96.1%) | 86 (94.5%) | 702 (96.3%) | |
| Yes | 140 (17.8%) | 23 (26.7%) | 117 (16.7%) | 0.021 |
| No | 648 (82.2%) | 63 (73.3%) | 585 (83.3%) | |
| ISTs, *n* (%) | 789 (96.2%) | 87 (95.6%) | 702 (96.3%) | |
| Yes | 235 (29.8%) | 27 (31.0%) | 208 (29.6%) | 0.787 |
| No | 554 (70.2%) | 60 (69.0%) | 494 (70.4%) | |
| Subtype, *n* (%) | 820 (100%) | 91 (100%) | 729 (100%) | |
| HIV-1 Subtype B | 333 (40.6%) | 39 (42.9%) | 294 (40.3%) | 0.643 |

*(Continued)*

**TABLE 1 |** (Continued)

| Patient characteristics | Total | With TDR | Without TDR | *p*-value |
|---|---|---|---|---|
| HIV-1 Subtype non-B | 487 (59.4%) | 52 (57.1%) | 435 (59.7%) | |
| Distribution of subtypes | | | | |
| HIV-1 Subtype B | 333 (40.6%) | 39 (42.9%) | 294 (40.3%) | 0.005 |
| HIV-1 Subtype C | 67 (8.2%) | 18 (19.8%) | 49 (6.7%) | |
| HIV-1 Subtype G | 89 (10.9%) | 7 (7.7%) | 82 (11.2%) | |
| HIV-1 Subtype A1 | 78 (9.5%) | 6 (6.6%) | 72 (9.9%) | |
| HIV-1 Subtype D | 2 (0.2%) | 0 (0.0%) | 2 (0.3%) | |
| HIV-1 Subtype F1 | 57 (7.0%) | 4 (4.4%) | 53 (7.3%) | |
| HIV-1 Subtype H | 4 (0.5%) | 0 (0.0%) | 4 (0.5%) | |
| HIV-1 CRF-14BG | 48 (5.9%) | 7 (7.7%) | 41 (5.6%) | |
| HIV-1 CRF-02AG | 61 (7.4%) | 5 (5.5%) | 56 (7.7%) | |
| HIV-1 recombinants | 81 (9.9%) | 5 (5.5%) | 76 (10.4%) | |
| Median CD4 count at diagnosis (cells/μL) IQR, *n* (%) | 803 (97.9%) | 90 (98.9%) | 713 (97.8%) | |
| | 339.0 (140.0 -519.0) | 318.0 (73.3 -513.5) | 366.3 (152.0 -522.0) | 0.061 |
| < 200 | 253 (31.5%) | 36 (40.0%) | 217 (30.4%) | |
| 201-349 | 168 (20.9%) | 15 (16.7%) | 153 (21.5%) | 0.272 |
| 350-500 | 164 (20.4%) | 15 (16.7%) | 149 (20.9%) | |
| > 501 | 218 (27.1%) | 24 (26.7%) | 194 (27.2%) | |
| Late Presentation, *n* (%) | 803 (97.9%) | 90 (98.9%) | 713 (97.8%) | |
| LP | 419 (52.2%) | 51 (56.7%) | 368 (51.6%) | 0.373 |
| NLP | 384 (47.8%) | 39 (43.3%) | 345 (48.4%) | |
| Viral Load at diagnosis (log10 copies/mL) IQR, n (%) | 746 (91.0%) | 84 (92.3%) | 662 (90.8%) | |
| | 4.95 (4.4–5.5) | 5.0 (4.4–5.6) | 4.94 (4.4–5.5) | 0.310 |
| ≤ 4.0 | 122 (16.4%) | 13 (15.5%) | 109 (16.5%) | |
| 4.1-5.0 | 291 (39.0%) | 29 (34.5%) | 262 (39.6%) | 0.565 |
| ≥ 5.1 | 333 (44.6%) | 42 (50.0%) | 291 (44.0%) | |

**Clinical type of infection was determined based on the ambiguity rate of genomic sequences. Chronic infection was defined as an ambiguity value > 0.45% and recent infection as an ambiguity value ≤ 0.45% (Andersson et al., 2013). **MSM refers to the transmission route variable, obtained from a closed question in the clinical questionnaires, filled in by the patients' clinicians (options: "Heterosexual," "MSM," "IDU," and "other"). ***Men sexual partners refer to a question in the sociobehavioral questionnaire, filled in by the patients belonging to the vulnerable groups of migrants, and MSM refers to with whom they usually have sex (options: "I have sex with men," " I have sex with women," and "I have sex with men and women").*

sex with men, and 12.9% of men reported having sex with both men and women. However, 100% of women included in the study reported only having sex with men. Based on the percentage of genomic ambiguities, approximately half of the patients (51.3%) presented chronic disease, while 70.0% presented infection stage A. About 82.2% presented no AIDS-defining events, and 70.2% had no sexually transmitted diseases other than HIV. Patients who reported STIs other than HIV were heterosexual (15.3%) and MSMs (43.4%) (data not shown). Patients were predominantly infected with subtype B (40.6%), followed by subtype G (10.9%). At diagnosis, the median viral load (VL) was 4.9 (IQR: 4.4–5.5) Log10 copies/ml, and the CD4 cell count was 339.0 (IQR: 140.0–519.0) cells/μL. About 52.4% of the study population were late presenters. Moreover, the study population presented 31.3% of patients diagnosed as late presenters with advanced disease, with a CD4 count lower than 200 cells/μL. About 77.4% of the LPAD cases were men, and 93.0% were heterosexual. None were MSM. Among these LPAD cases, more than half were born in Portugal (52.8%), and 45.3% were migrants. Among the migrants, 62.5% were born in Africa (data not shown).

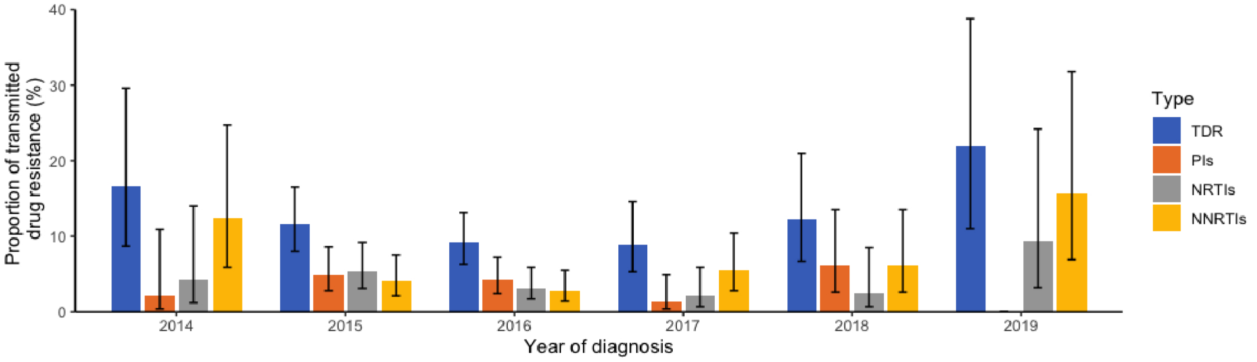
## Transmitted HIV Drug Resistance

The overall prevalence of TDR between 2014 and 2019 was 11.0% (95%CI: 9.0–13.3%). Nucleoside reverse transcriptase inhibitor (NRTI) mutations were detected in 3.9% (95%CI: 2.8–5.5%), non-nucleoside reverse transcriptase inhibitor (NNRTI) mutations in 5.0% (95%CI: 3.6–6.6%), and protease inhibitors (PI) in 3.9% (95%CI: 2.8–5.5%) of the HIV cases. In total, 9.6% (95%CI: 7.8–11.9%) presented single-class resistance, 1.2% (95%CI: 0.7–2.2%) dual-class resistance, and 0.2% triple-class resistance (**Figure 1**). Trends for TDR were determined for the period 2014–2016, where it was decreasing, and for the period 2017–2019, where it was increasing (**Figure 1**).

Overall TDR presented a decreasing trend from 16.7% in 2014 to 9.2% in 2016 ($p_{for-trend} = 0.114$), and TDR to NRTIs also showed a declining trend (4.17% in 2014 to 3.17% in 2016; $p_{for-trend} = 0.361$). TDR to NNRTIs presented a significantly decreasing trend from 12.5% in 2014 to 2.82% in 2016 ($p_{for-trend} < 0.05$). TDR to PIs, on the other hand, presented an increasing trend (2.08% in 2014 to 4.23% in 2016; $p_{for-trend} = 0.797$). Between 2017 and 2019, TDR presented an increasing trend from 8.9% in 2017 to 21.9% in 2019 ($p_{for-trend} < 0.05$) and also showed an increasing trend for all the drug classes, however, without statistical significance (**Figure 1**).

According to the HIVdb Stanford database algorithm, NNRTIs presented the highest level of high-level resistance (9.7%) among the drug classes. Nevirapine (NVP) showed the highest proportion of high-level resistance (4.7%), followed by efavirenz (EFV) with 3.7%, and both are related to the most frequently detected mutation K103NS (3.1%). High-level resistance to NRTIs occurred in 2.7% of the patients, with TDR to emtricitabine (FTC) and lamivudine (3TC) presenting the highest levels (1.0%), which is related to M41L (1.4%) mutation. High-level resistance to PIs was found in 0.4% of the patients, with atazanavir (ATV) presenting 0.4% of high-level resistance, which is related to the presence of L90M mutation (2.2%) (**Figures 2A,B**).

We also analyzed the association between HIV drug resistance mutations and subtypes of infection. We observed that individuals infected with subtype B were more likely to develop mutations associated with resistance to all the antiretroviral classes (NRTIs, NNRTIs, and PIs). In the individuals infected with subtype B, resistance to NRTIs can be attributed mainly to the M41L (0.8%), K219QR, and T215rev (0.4%) mutations. For



|  |  | Trend (2014–2016) |  |  | Trend (2017–2019) |  |
|---|---|---|---|---|---|---|
|  | n (%) | 95% CI | OR (95% CI) | p for Trend | OR (95% CI) | p for Trend |
| Total | 820 (100.0) |  |  |  |  |  |
| Any DRMs | 90 (11.0%) | 9.0-13.3 | 0.726 (0.488-1.080) | 0.114 | 1.651 (0.997-2.732) | 0.051 |
| NRTI resistance | 32 (3.9%) | 2.8–5.5 | 0.751 (0.406-1.389) | 0.361 | 2.205 (0.896–5.425) | 0.085 |
| NNRTI resistance | 40 (4.9%) | 3.6–6.6 | 0.468 (0.258–0.849) | 0.012 | 1.704 (0.913–3.183) | 0.094 |
| PI resistance | 32 (3.9%) | 2.8–5.5 | 1.088 (0.571-2.073) | 0.797 | 1.344 (0.495-3.651) | 0.562 |
| Single class resistance | 79 (9.6%) | 7.8-11.9 | 0.743 (0.487-1.135) | 0.170 | 1.458 (0.857-2.481) | 0.165 |
| Dual class resistance | 10 (1.2%) | 0.7-2.2 | 0.808 (0.245–2.670) | 0.727 | 3.822 (0.812–17.994) | 0.090 |
| Triple class resistance | 2 (0.2%) | 0.07-0.8 | 0.409 (0.059-2.862) | 0.368 | —– | — |

**FIGURE 1 |** Proportion of transmitted drug resistance (TDR) in sequences obtained from newly diagnosed patients between 2014 and 2019. NRTI, Nucleoside reverse transcriptase inhibitor; NNRTI, Non-nucleoside reverse transcriptase inhibitor; PI, Protease inhibitor; CI, confidence interval; OR, odds ratio.

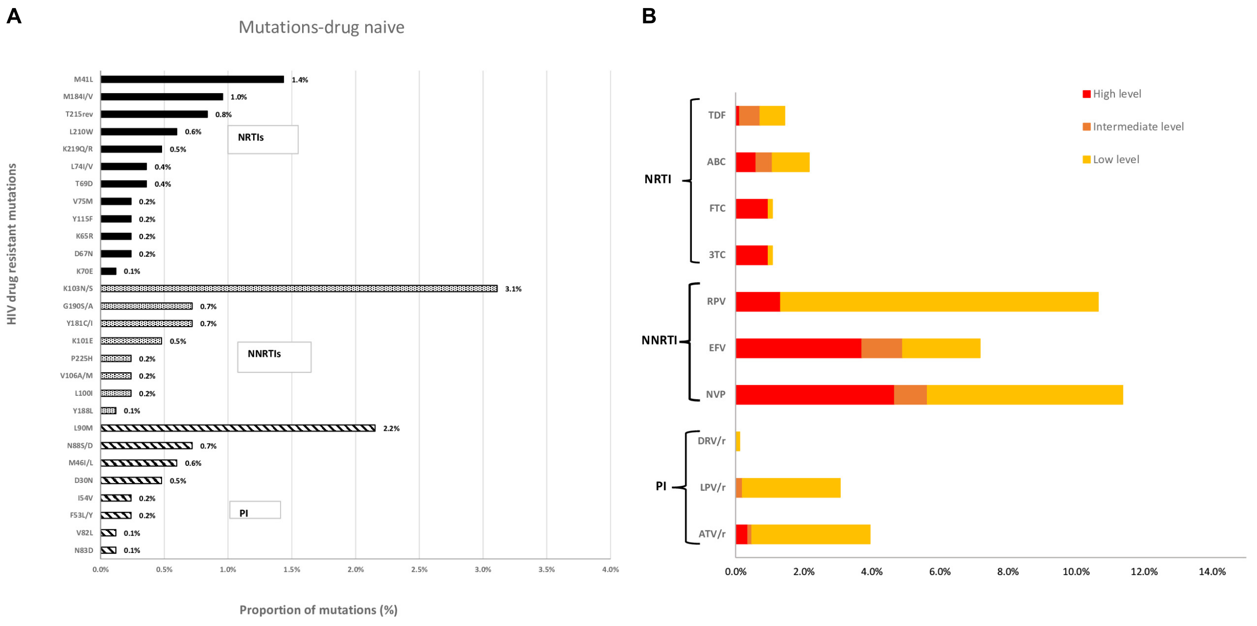


**FIGURE 2 | (A)** Proportion of resistance mutations in the sequences of newly diagnosed patients and **(B)** Predicted phenotypic resistance (Stanford scores) to antiretroviral drugs currently recommended as first-line therapy in Portugal for newly diagnosed patients (2014–2019). NRTI, Nucleoside reverse transcriptase inhibitor; NNRTI, Non-nucleoside reverse transcriptase inhibitor; PI, Protease inhibitor; FTC, Emtricitabine; TDF, Tenofovir; 3TC, Lamivudine; ABC, Abacavir; EFV, Efavirenz; RPV, Rilpivirine; DRV/r, Darunavir; LPV/r, Lopinavir; ATV/r, Atazanavir. Scores of low-level (score 2 and 3), intermediate-level (score 4), or high-level (score 5) resistance were used to predict phenotypic resistance.

**TABLE 2 |** Unadjusted and adjusted regression analysis of factors associated with HIV-transmitted drug resistance.

| **Any TDR** | | **Unadjusted** | | **Final model** | |
|---|---|---|---|---|---|
| | | **OR (95%CI)** | ***p*-value** | **aOR (95%CI)** | ***p*-value** |
| Sex, *n* (%) | Male | 1 | 1 | 1 | 1 |
| | Female | 1.53 (0.86–2.74) | 0.149 | 1.75 (0.89–3.44) | 0.10 |
| Age groups | 15–21 | 1 | 1 | | |
| | 25–40 | 1.06 (0.36–3.11) | 0.92 | | |
| | 41–55 | 0.91 (0.30–2.81) | 0.88 | | |
| | ≥56 | 0.64 (0.18–2.32) | 0.49 | | |
| Transmission route | Heterosexual | 1 | 1 | 1 | 1 |
| | MSM | 0.62 (0.40–0.98) | 0.04 | 0.55 (0.30–1.02) | 0.06 |
| | IDU | 0.48 (0.06–3.72) | 0.48 | 0.00 | 1 |
| | Other | 2.24 (0.44–11.36) | 0.33 | 1.51 (0.22–10.5) | 0.68 |
| Country of origin | Portugal | 1 | 1 | | |
| | Brazil | 1.08 (0.55–2.13) | 0.82 | | |
| | Guinea-Bissau | 0,36 (0,09–1.51) | 0.16 | | |
| | Angola | 1.69 (0.67–4.21) | 0.26 | | |
| | Cabo-Verde | 0.33 (0.04–2.46) | 0.28 | | |
| | Mozambique | 0.00 | 1.00 | | |
| | Others | 1.50 (0.50–4.50) | 0.47 | | |
| Region of origin | Europe | 1 | | | |
| | Africa | 0.06 (0.54–2.10) | 0.86 | | |
| | South America | 0.62 (0.30–1.28) | 0.20 | | |
| | Other | 1.29 (0.15–10.88) | 0.82 | | |
| District of residence | Lisboa | 1 | | | |
| | Porto | 1.04 (0.56–1.96) | 0.89 | | |
| | Faro | 1.04 (0.47–2.29) | 0.92 | | |
| | Setúbal | 0.69 (0.28–1.73) | 0.43 | | |
| | Aveiro | 2.10 (0.89–4.96) | 0.09 | | |
| | Beja | 1.58 (0.33–7.45) | 0.57 | | |
| | Coimbra | 1.06 (0.40–2.88) | 0.91 | | |
| | Outro | 2.89 (0.88–9.52) | 0.08 | | |
| Migrant status | Migrant | 1 | | | |
| | Native | 0.89 (0.54–1.45) | 0.63 | | |
| School level | Third level (9th degree) | 1 | | | |
| | Secondary (12th degree) | 0.65 (0.28–1.52) | 0.32 | | |
| | Advanced Technical Specialization | 0.37 (0.08–1.48) | 0.21 | | |
| | Higher education (bachelor, master, PhD) | 0.65 (0.28–1.48) | 0.30 | | |
| | None | 0 | 0.99 | | |
| Current occupation | Employed | 1 | | | |
| | Retired | 1.10 (0.14–9.0) | 0.93 | | |
| | Sex-worker | | 1 | | |
| | Student | 0.96 (0.21–4.29) | 0.96 | | |
| | Unemployed | 1.30 (0.57–2.98) | 0.54 | | |
| Current income | Very Insufficient | 1 | | | |
| | Insufficient | 0.81 (0.32–2.09) | 0.67 | | |
| | Sufficient | 0.76 (0.28–2.04) | 0.58 | | |
| | More than sufficient | 0.56 (0.06–4.91) | 0.60 | | |
| Civil status | Single | 1 | | | |
| | Married | 1.73 (0.80–3.71) | 0.16 | | |
| | Divorced | 1.58 (0.51–4.87) | 0.42 | | |
| | Widower | 13.04 (0.79–215.7) | 0.07 | | |
| | Other | 4.35 (0.83–22.8) | 0.08 | | |
| Men sexual partners, *n* (%) | Men | 1 | | | |

*(Conitnued)*

**TABLE 2 |** (Conitnued)

| **Any TDR** | | **Unadjusted** | | **Final model** | |
|---|---|---|---|---|---|
| | | **OR (95%CI)** | ***p*-value** | **aOR (95%CI)** | ***p*-value** |
| | Women | 1.58 (0.61–4.08) | 0.35 | | |
| | Men and women | 0.65 (0.19–2.24) | 0.50 | | |
| Type of infection | Chronic | 1 | | | |
| | Recent | 0.68 (0.44–1.06) | 0.092 | | |
| Infection stage | A | 1 | | 1 | |
| | B | 1.31 (0.66–2.62) | 0.44 | 1.16 (0.51–2.65) | 0.72 |
| | C | 2.04 (1.21–3.45) | 0.01 | 1.67 (0.38–7.40) | 0.50 |
| Aids-defining event | Yes | 1 | | 1 | 1 |
| | No | 0.53 (0.32–0.89) | 0.02 | 0.76 (0.17–3.38) | 0.72 |
| ISTs | Yes | 1 | | | |
| | No | 1.07 (0.66–1.73) | 0.79 | | |
| Subtype | HIV-1 Subtype B | 1 | | | |
| | HIV-1 Subtype non-B | 1.09 (0.70–1.69) | 0.71 | | |
| Distribution of Subtypes | HIV-1 Subtype B | 1 | | 1 | |
| | HIV-1 Subtype C | 2.86 (1.52–5.40) | 0.001 | 3.10 (1.50–6.44) | 0.002 |
| | HIV-1 Subtype G | 0.63 (0.27–1.47) | 0.29 | 0.62 (0.25–1.55) | 0.31 |
| | HIV-1 Subtype A1 | 0.63 (0.26–1.55) | 0.32 | 0.77 (0.30–1.98) | 0.59 |
| | HIV-1 Subtype D | 0.00 | 1.0 | 0.00 | 1 |
| | HIV-1 Subtype F1 | 0.59 (0.20–1.71) | 0.33 | 0.59 (0.19–1.85) | 0.29 |
| | HIV-1 Subtype H | 0.00 | 1.0 | 1.06 (0.38–2.94) | 0.91 |
| | HIV-1 CRF-14BG | 0.70 (0.26–1.84) | 0.47 | 0.58 (0.19–1.85) | 0.36 |
| | HIV-1 CRF-02AG | 1.33 (0.56–3.17) | 0.52 | 1.06 (0.38–2.94) | 0.91 |
| | HIV-1 recombinants | 0.51 (0.20–1.34) | 0.18 | 0.39 (0.13–1.20) | 0.10 |
| CD4 count at diagnosis (cells/μL) | < 200 | 1 | | | |
| | 201–350 | 0.59 (0.31–1.12) | 0.11 | | |
| | 351–500 | 0.61 (0.32–1.15) | 0.13 | | |
| | > 501 | 0.75 (0.43–1.30) | 0.30 | | |
| Late presentation | LP | 1 | | | |
| | NLP | 0.82 (0.52–1.27) | 0.37 | | |
| Viral load at diagnosis (log10 copies/mL) | ≤4.0 | 1 | | | |
| | 4.1–5.0 | 0.93 (0.47–1.86) | 0.84 | | |
| | ≥5.1 | 1.19 (0.62–2.31) | 0.60 | | |

NNRTIs, the most common mutations were K103NS (1.20%) and G190SA (0.4%), and for PIs, the main mutation found was M46IL (0.6%) (data not shown).

## Predictors of Transmitted Drug Resistance

The clinical and sociodemographic factors significantly associated with TDR in the univariate model were transmission route (sex between men, OR = 0.62), stage of infection (stage C, OR = 2.04), having an AIDS-defining event (patients without AIDS-defining events, OR = 0.52), and being infected with subtype C (OR = 2.86). The multivariate analysis indicated that TDR was significantly associated with the transmission route (MSM presented a lower probability of having TDR when compared to the heterosexual contact) and infection with subtype C (compared to subtype B) (**Table 2**).

## Inference of Transmission Clusters

Based on the PR+RT phylogenetic analysis, transmission clusters (TCs) were defined as clades with a branch support value ≥ 99% for subtypes B, A, and G.

We identified 87 transmission clusters comprising 273 of the 500 patients (54.6%). The average cluster size was 17.4, with a minimum of 2 (31 clusters) and a maximum of 118 (1 cluster).

When the proportion of transmission clusters between the subtypes was compared, patients carrying subtype G strains were more likely to be inside the clusters (67/89; 75.3%), followed by subtypes B (180/333; 54.1%) and A (22/78; 28.2%) ($p < 0.001$). Subtype G and A sequences in clusters presented a high proportion of heterosexual patients (84.1% and 57.1%, respectively). On the other hand, among subtype B sequences inside the clusters, 72.7% belonged to the MSM population.

Among the subtypes B, A, and G, 52 out of 500 (10.6%) patients presented HIV drug resistance, while almost the same proportion was observed in transmission clusters (9.5%). The

**TABLE 3 |** Subtypes, TDR, and risk factor of patients associated with HIV-1 molecular transmission clusters.

| | Cluster *N* = 273 | Non-cluster *N* = 227 | *p* value |
|---|---|---|---|
| Subtypes | | | < 0.0001 |
| A | 28.2 (22/78) | 71.8 (56/78) | |
| B | 55.3 (180/333) | 44.7 (149/333) | |
| G | 75.3 (67/89) | 24.7 (22/89) | |
| **TDR** | | | |
| Overall | 9.5 (26/273) | 11.5 (26/227) | 0.730 |
| A | 13.6 (3/22) | 5.4 (3/56) | 0.192 |
| B | 9.8 (18/184) | 14.0 (21/149) | 0.235 |
| G | 7.5 (5/67) | 9 (2/22) | 0.821 |
| **Risk factor** | | | 0.219 |
| MSM | 53.0 (150/283) | 47.0 (133/283) | |
| Heterosexual | 57.0 (114/200) | 43.0 (86/200) | |

**TABLE 4 |** Characteristic of patients in HIV-1 molecular transmission clusters according to the transmission route.

| | MSM | Heterosexual | *p* value |
|---|---|---|---|
| Subtypes | | | < 0.0001 |
| A (*n* = 21) | 42.9 (9/21) | 57.1 (12/21) | |
| B (*n* = 180) | 72.7 (131/180) | 27.3 (49/180) | |
| G (*n* = 63) | 15.9 (10/63) | 84.1 (53/63) | |
| TDR | | | |
| Overall (264) | 9.3 (14/150) | 9.6 (11/114) | 0.900 |
| A (*n* = 21) | 11.1 (1/9) | 16.7 (2/12) | 0.699 |
| B (*n* = 180) | 7.6 (10/131) | 14.3 (7/49) | 0.189 |
| G (*n* = 63) | 0 (0/53) | 50.0 (5/10) | – |

highest prevalence of TDR (73%) was observed among small clusters (cluster size lower than 10 sequences). When we analyzed TDR inside the clusters between the subtypes, we observed different proportions of TDR. Subtypes A1 (13.6%) and B (9.8%) were more likely to carry SDRMs inside the clusters when compared to subtype G (7.5%). Regardless of the subtypes, the heterosexual population presented the highest proportion of TDR compared to MSMs. Among subtype A1, heterosexuals presented 16.7% of TDR, followed by 14.3% in patients infected with subtype B and in lower proportion (9.4%) in patients with subtype G (**Tables 3**, **4**).

## DISCUSSION

This study aimed to understand HIV-1-transmitted drug resistance in newly diagnosed patients by providing a current picture of transmitted drug resistance patterns in these populations. This study is particularly important, as it provides information that can guide the development of preventive measures directed at specific risk populations.

The study population was mostly composed of men (77.3%). The most prevalent age group was 22–40 years, and most patients originated from Portugal and were included in the heterosexual and MSM risk groups. The characteristics of this population are consistent with the patterns reported in the latest Portuguese health authorities report. Clinically, it is worth emphasizing the high prevalence of very late presenters (LPAD; 31.5%) identified in this study, which is also consistent with the prevalence reported in the Portuguese health authorities report (Direção Geral da Saúde and Instituto Nacional de Saúde Doutor Ricardo Jorge, 2020).

Our study showed that the estimated prevalence of TDR in Portugal was 11.0% (IC95%: 9.0–13.3) in patients diagnosed between 2014 and 2019. A decreasing trend for TDR was also observed between 2014 and 2016. A similar trend was observed for NRTIs and NNRTIs, but not for PIs, which showed increasing TDR in this period. Between 2017 and 2019, an increasing trend was observed for TDR and all drug classes. This increasing trend in TDR had already been observed by our study group in a longitudinal study that included patients between 2001 and 2017 (Pingarilho et al., 2020), as well as in a study published by a Canadian group (Rocheleau et al., 2018). This increase could be related to population mobility and an increase in the number of migrants from Portuguese-speaking Sub-Saharan African countries, where TDR has been increasing in the last few years (Rhee et al., 2015; Pingarilho et al., 2018; Sebastião et al., 2019; Pimentel et al., 2020).

The most prevalent mutation detected was K103N, which confers high-level resistance to NVP and EFV, followed by M41L, which reduces susceptibility to TDF and ABC when in combination with other NRTI mutations, and M184VI, which causes high-level resistance to 3TC and FTC. L90M resistance mutation presented the highest prevalence for PIs, and it causes reduced susceptibility to ATV and LPV. L90M mutation (1.8%) was widely observed in patients infected with subtype C, and this could be the reason for the increased resistance to PIs noticed in our study (3.9%), compared to 2.8% obtained in our previously published study (Pingarilho et al., 2020).

We also observed that the risk of TDR was significantly higher in patients infected with subtype C when compared to those infected with subtype B. We hypothesize that this finding could be explained by the higher prevalence of L90M mutations among MSMs infected with subtype C, eventually caused by its forward transmission in MSM transmission clusters. However, in this study, the transmission clusters of subtype C were not reconstructed, since the prevalence of this subtype was lower than 10%. Future studies will address this problem.

Although the L90M mutation was more frequent among the MSM group, in fact, the overall TDR was 1.7 times ($p = 0.027$) higher within the heterosexual population than that observed in the MSM group. This result is discordant from previous studies that have shown faster onward transmission of HIV infection with less reversion of DRM in the MSM transmission clusters and therefore a potentization of the transmission of TDR (Vercauteren et al., 2009). However, this finding agrees with the results obtained through transmission cluster analyses, which showed that subtype G heterosexuals (Pineda-Peña et al., 2019) were more frequently inside clusters when compared to subtypes A and B, indicating that transmission is more active in this subgroup. This is in contrast to what other European studies showed, where it was concluded that non-B subtypes are associated with the heterosexual population and are less frequently found in the transmission clusters (Lorenzin et al., 2019; Paraskevis et al., 2019; Pimentel et al., 2022). Since our study is based on newly diagnosed patients, we believe that our results present a more recent view of the epidemic in the country that could already reflect a successful impact of pre-exposure prophylaxis (PrEP) among the MSM patients, with a slowdown of HIV transmission clusters in this risk group. We hypothesize that MSM patients, in addition to auto-testing more frequently, have a higher risk perception that plays an important role in the acceptance of PrEP (Plotzker et al., 2017). Consistently with our hypothesis, some other studies have already reported that there is a high level of willingness and acceptance of PrEP use among MSM (Frankis et al., 2016; Phan and Vu, 2017; Spinner et al., 2018; Nguyen et al., 2021). Moreover, in Portugal, it is known that almost the entire population interested in using PrEP is MSM; however, no such study has been published until now. Nevertheless, a study conducted in 2016 reported that in France, more than 95% of the people interested in the use of PrEP were MSM (Loos et al., 2016).

The overall rate of TDR was 9.5% inside the clusters compared to 11.5% outside the clusters, and the highest proportion of TDR was observed in small clusters (73%), mostly composed of heterosexuals. Subtype A presented the highest prevalence of TDR in clusters, followed by subtypes B and G. However, consistent with the results of the analyses of TCs, we observed that heterosexual individuals from subtype G presented higher levels of TDR in clusters, compared to MSMs of the same subtype. The same was observed for subtypes A and B, where heterosexuals inside the clusters presented higher levels of TDR compared to MSM, despite the fact that subtype B presented a higher rate of MSM. Heterosexual transmission accounts for approximately 57.8% of the new HIV infections in Portugal in the last few years, with migrants contributing to approximately 43.1% of these new infections, mostly from Sub-Saharan African countries (51.2%), where heterosexual transmission is predominant (Direção Geral da Saúde and Instituto Nacional de Saúde Doutor Ricardo Jorge, 2020). Most of these heterosexual contacts present small clusters of two individuals (men–women), indicating low levels of forward transmission that still represent a high proportion of TDR transmission in Portugal.

Given the new trends presented in this manuscript, our results seem to indicate that the successful HIV prevention measures implemented in the MSM populations, which include the use of PreP (pre-exposure prophylaxis), frequent medical appointments and testing, and earlier diagnosis in community-based centers, seem to have been successful in decreasing the HIV transmission and consequently the TDR among the MSM group. Since, in Portugal, PreP is mostly used by the MSM population and not by heterosexuals, this can explain the higher proportion of heterosexuals in TCs and hence higher TDR transmission in heterosexuals. Also, it is important to note that the MSM population is more frequently tested and has an opportunity for earlier diagnosis, which may imply less transmission of HIV and thus TDR. HIV-1 prevention measures should now be strengthened for heterosexual risk groups.

## CONCLUSION

The transmission patterns of HIV-1 are changing in Portugal, probably due to the new prevention measures introduced in the country and which are mostly accepted by the MSM groups. However, it is very important to address the heterosexual group where TDR is increasing and which presents high levels of late diagnosis. For this reason, it is important to develop preventive measures for HIV-1 transmission addressing the specificities of this group.

## DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The dataset would be available through requisition and explanation of the study purposes. Requests to access these datasets should be directed to AS, ana.abecasis@ihmt.unl.

## OTHER MEMBERS OF THE BESTHOPE STUDY GROUP

Ana Bandeiras, Ana Pimenta, Anabela Granado, André Gomes, António Maio, Catarina Messias, Celina Bredes, Diana Seixas, Diva Trigo, Edite Mateus, Fátima Gonçalves, Filipa Azevedo, Francisco Vale, Henriqueta Pereira, Inês Siva, Isabel Casella, Isabel Diogo, Isabel Neves, Joana Sá, Joana Simões, Joana Granado, Joana Vasconcelos, João Cabo, João Pereira-Vaz, João Domingos, João Torres, Joaquim Cabanas, Johana Jesus, José Melo Cristino, Karen Pereira, Luís Caldeira†, Luísa Sêco, Lurdes Correia, Manuela Simão, Maria Saudade Ivo, Mariana Pessanha, Marta Feijó, Margarida Cardoso, Nildelema Malaba, Nádia Gomes, Natália Patrício, Nuno Luís, Nuno Janeiro, Patrícia Carvalho, Paula Brito, Pedro Simões, Rosário Prazos, Sara Lino, Sara Casanova, Sofia Pinheiro, Sónia Marques, Sofia Jordão, Sueila Martins, Telma Azevedo, Teresa Meira, Vanda Mota, and Vanda Silva.

## AUTHOR CONTRIBUTIONS

MP and AA: conceptualization and writing—original draft and writing—reviewing and editing. MP, MNM, and VP: data curation and validation. MP, MNM, VP, and AA: formal analysis. AA: funding acquisition, project administration, and supervision. MP, VP, PG, and MRM: investigation. MP, VP, MRM, and AA: methodology. PG, AS, AD, BA, CP, CK, CR, CC, CM, DF, ES, ET, FMo, FaR, FMa, FeR, GG, HR, IC, IG, JSi, JO, JF, JP, JC, JSo, JH, KM, LP, MA, MG, MJM, MGM, MS, MC, NM, OC, PPa, PPr, PR, RP, RT, RA, RR, RS, RC, SN, TF, and TB: resources, data collection, and approval of the publication contents. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was supported by the European funds through grant Bio-Molecular and Epidemiological Surveillance of HIV Transmitted Drug Resistance, Hepatitis Co-Infections and Ongoing Transmission Patterns in Europe (BEST HOPE) (project funded through HIVERA: Harmonizing Integrating Vitalizing European Research on HIV/Aids, grant no: 249697); by FCT for funds to GHTM-UID/Multi/04413/2013 and GHTM-UID/04413/2020; by the MigrantHIV project (financed by FCT: PTDC/DTP-EPI/7066/2014); by Characterization of drug-resistance TB and HIV, and associated sociobehavioral factors among migrants in Lisbon, Portugal project financed by GHTM-UID/Multi/04413/2013; and by Integriv project (financed by FCT: PTDC/SAU-INF/31990/2017).

## ACKNOWLEDGMENTS

We would like to thank Dr. Ricardo Camacho, who passed away in 2018, who designed and triggered this study. Without his deep involvement, this study would not have been possible. We would also like to thank the patients included in this study.

## REFERENCES

Aldous, A. M., Castel, A. D., Parenti, D. M., Greenberg, A. E., Benator, D., Kumar, P., et al. (2017). Prevalence and trends in transmitted and acquired antiretroviral drug resistance, Washington, DC, 1999–2014. *BMC Res. Notes* 10:474. doi: 10.1186/s13104-017-2764-9

Andersson, E., Shao, W., Bontell, I., Cham, F., Cuong, D. D., Wondwossen, A., et al. (2013). Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. *Infect Genet Evol* 18, 125–131. doi: 10.1016/j.meegid.2013.03.050

Antinori, A., Coenen, T., Costagiola, D., Dedes, N., Ellefson, M., Gatell, J., et al. (2011). Late presentation of HIV infection: a consensus definition. *HIV Med.* 12, 61–64. doi: 10.1111/j.1468-1293.2010.00857.x

Avila-Rios, S., Sued, O., Rhee, S.-Y., Shafer, R. W., Reyes-Teran, G., and Ravasi, G. (2016). Surveillance of HIV transmitted drug resistance in latin america and the caribbean: a systematic review and meta-analysis. *PLoS One* 11:e0158560. doi: 10.1371/journal.pone.0158560

Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., et al. (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* 4:e4724. doi: 10.1371/journal.pone.0004724

Clavel, F., and Hance, A. J. (2004). HIV drug resistance. *New Eng. J. Med.* 350, 1023–1035. doi: 10.1056/NEJMra025195

Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N., et al. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New Eng. J. Med.* 365, 493–505. doi: 10.1056/NEJMoa1105243

Direção Geral da Saúde, and Instituto Nacional de Saúde Doutor Ricardo Jorge (2020). *Infeção VIH e SIDA em Portugal- 2020*. Portugal: DGS.

Frankis, J. S., Young, I., Lorimer, K., Davis, M., and Flowers, P. (2016). Towards preparedness for PrEP: PrEP awareness and acceptability among MSM at high risk of HIV transmission who use sociosexual media in four celtic nations: scotland, wales, northern ireland and the republic of ireland: an online survey. *Sex Transm Infect.* 92, 279–285. doi: 10.1136/sextrans-2015-052101

Frentz, D., Van de Vijver, D. A., Abecasis, A. B., Albert, J., Hamouda, O., Jørgensen, L. B., et al. (2014). Increase in transmitted resistance to non-nucleoside reverse transcriptase inhibitors among newly diagnosed HIV-1 infections in Europe. *BMC Infect Dis.* 14:407. doi: 10.1186/1471-2334-14-407

Kuiken, C., Korber, B., and Shafer, R. W. (2003). HIV sequence databases. *AIDS Rev.* 5, 52–61.

Libin, P. J. K., Deforche, K., Abecasis, A. B., and Theys, K. (2019). VIRULIGN: fast codon-correct alignment and annotation of viral genomes. *Bio. Oxf. Eng.* 35, 1763–1765. doi: 10.1093/bioinformatics/bty851

Lima, V. D., Harrigan, R., Bangsberg, D. R., Hogg, R. S., Gross, R., Yip, B., et al. (1999). The combined effect of modern highly active antiretroviral therapy regimens and adherence on mortality over time. *J. Acquir Immune Defic Syndr* 2009, 529–536. doi: 10.1097/QAI.0b013e31819675e9

Loos, J., Nöstlinger, C., Reyniers, T., Colebunders, R., Jespers, V., Manirankunda, L., et al. (2016). PrEP for african migrants in Europe? A research agenda. *Lancet HIV* 3, e505–e507. doi: 10.1016/S2352-3018(16)30173-4

Lorenzin, G., Gargiulo, F., Caruso, A., Caccuri, F., Focà, E., Celotti, A., et al. (2019). Prevalence of non-B HIV-1 subtypes in north italy and analysis of transmission clusters based on sequence data analysis. *Microorganisms* 8:36. doi: 10.3390/microorganisms8010036

Nguyen, L. H., Nguyen, H. L. T., Tran, B. X., Larsson, M., Rocha, L. E. C., Thorson, A., et al. (2021). A qualitative assessment in acceptability and barriers to use pre-exposure prophylaxis (PrEP) among men who have sex with men: implications for service delivery in Vietnam. *BMC Infect. Dis.* 21:472. doi: 10.1186/s12879-021-06178-5

Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., et al. (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV outpatient study investigators. *New Eng. J. Med.* 338, 853–860. doi: 10.1056/NEJM199803263381301

Paraskevis, D., Beloukas, A., Stasinos, K., Pantazis, N., de Mendoza, C., Bannert, N., et al. (2019). HIV-1 molecular transmission clusters in nine european countries and canada: association with demographic and clinical factors. *BMC Med.* 17:4. doi: 10.1186/s12916-018-1241-1

Pennings, P. S. (2013). HIV drug resistance: problems and perspectives. *Infect Dis. Rep.* 2013:e5. doi: 10.4081/idr.2013.s1.e5

Phan, H. T. T., and Vu, N. T. T. (2017). Acceptance to use daily oral pre-exposure prophylaxis (PrEP) as an HIV prevention method and ability to pay for PrEP among men who have sex with men in ho chi minh city. *Vietnam. Health (N Y)* 9, 1326–1336. doi: 10.4236/health.2017.99096

Pimentel, V., Giordano, S., Marta, P., Daniela, A., Mafalda, M., Isabel, D., et al. (2022). Differential patterns of post-migration HIV-1 infection acquisition among portuguese immigrants of different geographical origin. *AID.* [Online ahead of print] doi: 10.1097/QAD.0000000000003203

Pimentel, V., Pingarilho, M., Alves, D., Diogo, I., Fernandes, S., Miranda, M., et al. (2020). Molecular epidemiology of HIV-1 infected migrants followed up in portugal: trends between 2001-2017. *Viruses* 12, 268. doi: 10.3390/v12030268

Pina-Araujo, M., Guimarães, M. L., Bello, G., Vicente, A. C. P., and Morgado, M. G. (2014). Profile of the HIV epidemic in cape verde: molecular epidemiology and drug resistance mutations among HIV-1 and HIV-2 infected patients from distinct islands of the archipelago. *PLoS One* 9:e96201. doi: 10.1371/journal.pone.0096201

Pineda-Peña, A.-C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., et al. (2013). Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 19, 337–348. doi: 10.1016/j.meegid.2013.04.032

Pineda-Peña, A.-C., Pingarilho, M., Li, G., Vrancken, B., Libin, P., Gomes, P., et al. (2019). Drivers of HIV-1 transmission: the portuguese case. *PLoS One* 14:e0218226. doi: 10.1371/journal.pone.0218226

Pingarilho, M., Pimentel, V., Diogo, I., Fernandes, S., Miranda, M., Pineda-Pena, A., et al. (2020). Increasing prevalence of HIV-1 transmitted drug resistance in portugal: implications for first line treatment recommendations. *Viruses* 12:1238. doi: 10.3390/v12111238

Pingarilho, M., Pineda-Peña, A., Gomes, P., Libin, P., Theys, K., and Abecasis, A. (2018). Molecular epidemiology of hiv infection in portuguese migrant population. *Viruse* 12:268.

Plotzker, R., Seekaew, P., Jantarapakde, J., Pengnonyang, S., Trachunthong, D., Linjongrat, D., et al. (2017). Importance of risk perception: predictors of PrEP acceptance among thai MSM and TG women at a community-based health service. *JAIDS J. Acquir Immune Defic Syndr* 76, 473–481. doi: 10.1097/QAI.0000000000001536

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., et al. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinform.* 14:317. doi: 10.1186/1471-2105-14-317

Rhee, S.-Y., Blanco, J. L., Jordan, M. R., Taylor, J., Lemey, P., Varghese, V., et al. (2015). Geographic and temporal trends in the molecular epidemiology and genetic mechanisms of transmitted HIV-1 drug resistance: an individual-patient- and sequence-level meta-analysis. *PLoS Med.* 12:1001810. doi: 10.1371/journal.pmed.1001810

Rocheleau, G., Brumme, C. J., Shoveller, J., Lima, V. D., and Harrigan, P. R. (2018). Longitudinal trends of HIV drug resistance in a large canadian cohort, 1996-2016. *Clin. Microbiol. Infect. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* 24, 185–191. doi: 10.1016/j.cmi.2017.06.014

Sebastião, C. S., Neto, Z., de Jesus, C. S., Mirandela, M., Jandondo, D., Couto-Fernandez, J. C., et al. (2019). Genetic diversity and drug resistance of HIV-1 among infected pregnant women newly diagnosed in luanda, angola. *PLoS One* 14:e0225251. doi: 10.1371/journal.pone.0225251

Spinner, C. D., Hanhoff, N., Krznaric, I., Knecht, G., Kuemmerle, T., Ruesenberg, R., et al. (2018). 2016 PREP attitudes in germany: high awareness and acceptance in MSM at risk of HIV. *Infection* 46, 405–408. doi: 10.1007/s15010-018-1127-3

Struck, D., Lawyer, G., Ternes, A.-M., Schmit, J.-C., and Bercoff, D. P. (2014). COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res.* 42:e144. doi: 10.1093/nar/gku739

van de Laar, M. J., Bosman, A., Pharris, A., Andersson, E., Assoumou, L., Ay, E., et al. (2019). Piloting a surveillance system for HIV drug resistance in the European Union. *Eurosurveillance* 24:1800390. doi: 10.2807/1560-7917.ES.2019.24.19.1800390

Vercauteren, J., Wensing, A. M. J., van de Vijver, D. A., Albert, J., Balotta, C., Hamouda, O., et al. (2009). Transmission of drug-resistant HIV-1 is stabilizing in europe. *J. Infect. Dis.* 200, 1503–1508. doi: 10.1086/644505

World Health Organiztion (WHO) (2021). *Portugal on Fast Track to Achieve HIV Targets Ahead of 2020 Deadline*. Available online at: https://www.euro.who.int/en/countries/portugal/news/news/2018/7/portugal-on-fast-track-to-achieve-hiv-targets-ahead-of-2020-deadline (accessed May 6, 2021)

Yang, W.-L., Kouyos, R. D., Böni, J., Yerly, S., Klimkait, T., Aubert, V., et al. (2015). Persistence of transmitted HIV-1 drug resistance mutations associated with fitness costs and viral genetic backgrounds. *PLoS Pathog* 11:e1004722. doi: 10.1371/journal.ppat.1004722

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 12 May 2022
doi: 10.3389/fmicb.2022.855049

# Multiple CRF01_AE/CRF07_BC Recombinants Enhanced the HIV-1 Epidemic Complexity Among MSM in Shenyang City, Northeast China

Shan He[1,2,3], Wei Song[4], Gang Guo[5], Qiang Li[5], Minghui An[1,2,3], Bin Zhao[1,2,3], Yang Gao[1,2,3], Wen Tian[1,2,3], Lin Wang[1,2,3], Hong Shang[1,2,3]* and Xiaoxu Han[1,2,3]*

[1]NHC Key Laboratory of AIDS Immunology (China Medical University), National Clinical Research Center for Laboratory Medicine, The First Affiliated Hospital of China Medical University, Shenyang, China, [2]Chinese Academy of Medical Sciences Research Unit (No. 2019RU017), China Medical University, Shenyang, China, [3]Key Laboratory of AIDS Immunology of Liaoning Province, Shenyang, China, [4]Department of Food Safety and Nutrition, Shenyang Center for Health Service and Administrative Law Enforcement (Shenyang Center for Disease Control and Prevention), Shenyang, China, [5]Department of Clinical Laboratory, The Sixth People's Hospital of Shenyang, Shenyang, China

OPEN ACCESS

**Edited by:**
Kok Keng Tee,
University of Malaya, Malaysia

**Reviewed by:**
Weiping Cai,
Guangzhou Eighth People's Hospital, China
Marcel Tongo,
Center for Research on Emerging and Re-Emerging Diseases (CREMER), Cameroon

***Correspondence:**
Hong Shang
hongshang100@hotmail.com
Xiaoxu Han
hanxiaoxu@cmu.edu.cn

**Specialty section:**
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

**Received:** 14 January 2022
**Accepted:** 11 April 2022
**Published:** 12 May 2022

**Citation:**
He S, Song W, Guo G, Li Q, An M, Zhao B, Gao Y, Tian W, Wang L, Shang H and Han X (2022) Multiple CRF01_AE/CRF07_BC Recombinants Enhanced the HIV-1 Epidemic Complexity Among MSM in Shenyang City, Northeast China. Front. Microbiol. 13:855049. doi: 10.3389/fmicb.2022.855049

The transmission of Unique Recombinant Forms (URFs) has complicated the molecular epidemic of HIV-1. This increasing genetic diversity has implications for prevention surveillance, diagnosis, and vaccine design. In this study, we characterized the HIV-1 URFs from 135 newly diagnosed HIV-1 infected cases between 2016 and 2020 in Shenyang, northeast China and analyzed the evolutionary relationship of them by phylogenetic and recombination approaches. Among 135 URFs, we found that the CRF01_AE/CRF07_BC recombinants were the most common (81.5%, 110/135), followed by CRF01_AE/B (11.9%, 16/135), B/C (3.7%, 5/135), and others (3.0%, 4/135). 94.8% (128/135) of patients infected by URFs were through homosexual contact. Among 110 URFs_0107, 60 (54.5%) formed 11 subclusters (branch support value = 1) and shared the consistent recombination structure, respectively. Four subclusters have caused small-scale spread among different high-risk populations. Although the recombination structures of URFs_0107 are various, the hotspots of recombinants gathered between position 2,508 and 2,627 (relative to the HXB2 position). Moreover, the CRF07_BC and CRF01AE fragments of URFs_0107 were mainly derived from the MSM population. In brief, our results reveal the complex recombinant modes and the high transmission risk of URFs_0107, which calls for more attention on the new URFs_0107 monitoring and strict control in the areas led by homosexual transmission route.

**Keywords: HIV-1, men who have sex with men, unique recombinant forms, transmission, phylogenetics**

## INTRODUCTION

The occurrence of human immunodeficiency virus type 1 (HIV-1) recombination significantly increases the genetic complexity and enhances viral evolution and adaptation (Zhang et al., 2010). By far, more than 118 circulating recombinant forms (CRFs) and increasing unique recombinant forms (URFs), the latter was isolated only from single individuals (Hemelaar, 2013), have been reported in the Los Alamos National Laboratory HIV database.[1] A systematic literature search and global survey study on HIV-1 diversity from 1990 to 2015 demonstrated the proportion of HIV-1 infections due to recombinants is highest in Southeast Asia, followed in China, and West and Central Africa, mainly by CRF01_AE and CRF02_AG, respectively (Hemelaar et al., 2020b). Meanwhile, URFs have also contributed greatly to the HIV-1 epidemic in Central, West, and East Africa, as well as Latin America (Hemelaar et al., 2020a), and led to up to 30% of infections in regions where multiple subtypes coexist (Hemelaar et al., 2020b). Therefore, the genetic diversity of HIV-1 has posed huge challenges to prevention, surveillance, drug resistance, treatments, and the broad vaccines development (Nájera et al., 2002; Nora et al., 2007).

China is one of the countries with the most subtypes and complexity of HIV-1 epidemic in the world (Hemelaar et al., 2020b). The co-circulation of various subtypes of strains and the occurrence of multiple infections have resulted in the continuous emergence of new recombinants (Vuilleumier and Bonhoeffer, 2015). In the 1990s, the generation of CRF07_BC and CRF08_BC was due to the high incidence of B/C recombination events in Yunnan, the gateway of HIV-1 in China (McClutchan et al., 2002; Feng et al., 2016; Chen et al., 2019), which had a profound impact on the epidemic of HIV-1 (Yuan et al., 2016; Zhang et al., 2017). Another circulating subtype CRF55_01B is composed of CRF01_AE and B strains, which can be traced back to MSM in Shenzhen, Guangdong in the 2000s, and have been currently found throughout the country (Han et al., 2013a; Zai et al., 2020). The latest national HIV molecular epidemiological survey shows that novel CRFs and URFs account for as high as 3.5% and 5.0%, respectively, which basically cover all provinces and cities in China. Among them, recombinants with CRF01_AE and CRF07_BC as parents are the most common, mainly circulating among men who have sex with men (MSM).

The incidence of HIV-1 among MSM in China continues to rise, especially in large cities such as Beijing, Shanghai, Chongqing, Kunming, Shenzhen, and Shenyang (Wu et al., 2013; Qi et al., 2015). Multiple sexual partners and unprotected high-risk behaviors among MSM lead to the persistent emergence of URFs (Xu et al., 2010; Yin et al., 2019). HIV-1 recombination has been reported to enhance biological fitness (Njai et al., 2006), escape the host immune response (Streeck et al., 2008), and generate variants of dual or multi-drug resistant (Moutouh et al., 1996; Nikolaitchik et al., 2015). Although the increasing number of distinct CRFs and URFs sequences is submitted to the HIV-1 database, most of them are used for sequence annotation and reporting of characteristic samples (Hao et al., 2019; Li et al., 2019; Chang et al., 2020; Jiang et al., 2020), and hence may not accurately represent the current epidemic of HIV-1 URFs, which brings difficulties to guide the precise prevention and control.

In the present study, we collected all newly diagnosed infections with URFs from 2016 to 2020 in Shenyang, a transportation hub in northeast China, where HIV-1 transmission is predominantly through the homosexual route, with multiple subtypes co-existing and an increasing proportion of recombinants (Jin-ping et al., 2018; Liu et al., 2020). It aimed to capture the full profile of the ongoing epidemic of URFs in Shenyang through characterizing the phylogeny and recombination patterns, prove critical of tracing the source to cut the chain of transmission, as well as have great significance for public intervention to control the HIV-1 URFs epidemic among Chinese MSM.

# MATERIALS AND METHODS

## Data Collection and Study Samples

A total of 135 HIV-1 URFs pol sequences (HXB2: 2268-3278) from 2016 to 2020 were screened out by phylogenetic and RIP (recombinant identification program) from a local HIV-1 drug resistance database built by the First Affiliated Hospital of China Medical University (Zhao et al., 2021). The database was established by all 3,934 newly diagnosed HIV-1 infections in Shenyang between 2016 and 2020. Social-demographic information including date of diagnosis, gender, age, ethnic group, marriage status, education, and transmission routes was collected. Plasma samples were collected at the time of diagnosis and stored at −80°C. All participants signed informed consent forms, and the Ethical approval was obtained from the ethics committee of the First Affiliated Hospital of China Medical University.

## Extraction, Amplification, and Sequencing

5′ half-genome (HXB2: 790-5056) amplifications of partial patients were performed. The RNA was transcribed into cDNA using Superscript III Reverse Transcriptase (Life Technologies, Carlsbad, CA, United States), then amplified by nested PCR using Platinum Taq DNA Polymerase High Fidelity (Life Technologies, Carlsbad, CA, United States). The primers of reverse transcription and PCR amplification, reaction system, and condition were reported previously (Zhao et al., 2011; Sanchez et al., 2014). The amplified DNA fragments were purified and sequenced by a commercial company (Huada, Beijing).

## Phylogenetic Analysis

The raw sequences were cleaned and assembled by Sequencer software v.5.4 (Gene Codes, Ann Arbor, MI, United States), and then were compared using the BLAST online tool (see footnote 1) to avoid potential experiment cross-contamination. The final sequences were aligned using HIV Align online tool (see footnote 1) and conducted manually using BioEdit 7.0 (Tippmann, 2004).[2] Phylogenetic trees of pol region were constructed using the Maximum-Likelihood (ML) method under the General Time Reversible (GTR) + I + G nucleotide substitution model with 1,000 replicates using IQ-Tree v2.0.5 (Nguyen et al., 2015), subsequently visualized by FigTree v1.4.2 (Price et al., 2010). The CRF01_AE, CRF07_BC, CRF_01B, and pure genotypes

[1] https://www.hiv.lanl.gov

[2] https://bioedit.software.informer.com/7.0/

of M group were used as reference sequences and downloaded from the Los Alamos HIV Database.

## Recombination Analysis

Recombination structures were preliminary analyzed using two online software tools (recombinant identification program and jumping profile Hidden Markov Model; see Footnote 1). Further, the recombination breakpoints were confirmed using SimPlot (Lole et al., 1999; version 3.5.1) with the following parameters: 200 nucleotides (nt) window, 20 nt step size, and 100 bootstrap replicates. Five subtypes were selected as reference sequences in Simplot analyses including CRF01_AE (JX960615), CRF07_BC (JX960601), B (U71182), C (AF067155), and subtype J (AF082395). And the recombination mosaic map was generated using the online Recombinant HIV-1 Drawing Tool.[3]

Recombination fragments (as determined using jpHMM and Simplot bootscanning) were phylogenetically studied using IQ-Tree v2.0.5. For some URFs with more than one recombination breakpoints, the pure and same subtype fragments were generally connected together for analysis to retain more sequence information. The subregion trees were generated by URFs recombination fragments using ML method under (GTR) + I + G model, including seven CRF01_AE lineage in China (Feng et al., 2013), CRF01_AE lineage in Thailand and Africa, as well as pure genotypes of HIV-1 group A as reference sequences.

## The Identification of HIV Infection Status by Limiting-Antigen Avidity Enzyme Immunoassay

Plasma specimens were tested with the Lag-Avidity EIA according to the manufacturer's instructions (Maxim Biomedical, Rockville, MD, United States). Specimens with initial ODn > 2.0 are considered long-term HIV-1 infection but if ODn ≤ 2.0, the specimens are tested against in triplicate to confirm their ODn values. In confirmatory testing, if the ODn of the specimen is >1.5, the specimen is considered a long-term infection. If ODn >0.4, but ≤1.5, the specimen is considered a recent HIV infection (RHI). The specimens which ODn values lower than 0.4 need further confirm HIV seropositivity *via* Western blot assay.

# RESULTS

## The Recombinant Profile and Social-Demographic Characteristics of 135 URFs Subjects

From 2016 to 2020, a total of 135 cases were identified infection with HIV-1 URFs in Shenyang. Among these cases, diverse recombinant combinations were identified, including the URFs_0107 (81.5%, 110/135), followed by URFs_01B (11.9%, 16/135), URFs_BC (3.7%, 5/135), and other unique recombinants (3.0%, 4/135). Overall, the proportion of URFs_0107 increased from 70.8% (17/24) in 2016 to 88.9% (16/18) in 2020, while the URFs_01B gradually declined from 20.8% (5/24) to 5.6% (1/18; **Supplementary Figure S1**). The social-demographic characteristics of 135 individuals were summarized in **Table 1**. Of the 110 identified URFs_0107, the majority of the subjects was male (99.1%, 109/110), younger than 35 years old (77.3%, 85/110), being "single" (77.3%, 85/110), Han ethnicity (84.5%, 93/110), had achieved beyond compulsory education (94.6%, 104/110) and predominantly transmitted by MSM (81.8%, 90/110). All cases of URFs_01B infection were male through homosexual behavior and had similar demographic characteristics with URFs_0107. On the contrary, for URFs_BC and other recombinants, most cases were female and older than 35 years and infected through heterosexual contact.

**TABLE 1** | Demographic characteristics of 135 HIV-1 recombinants infected patients in this study.

| | TOTAL (*n* = 135) | HIV subtype | | | |
|---|---|---|---|---|---|
| | | URF_0107 (*n* = 110) | URF_01B (*n* = 16) | URF_BC (*n* = 5) | Other (*n* = 4) |
| | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** |
| Year | | | | | |
| 2016 | 24 (17.8) | 17 (15.5) | 5 (31.3) | 2 (40.0) | 0 (0) |
| 2017 | 23 (17.0) | 21 (19.1) | 2 (12.5) | 0 (0) | 0 (0) |
| 2018 | 34 (25.2) | 26 (23.6) | 3 (18.8) | 2 (40.0) | 3 (75.0) |
| 2019 | 36 (26.7) | 30 (27.3) | 5 (31.3) | 0 (0) | 1 (25.0) |
| 2020 | 18 (13.3) | 16 (14.5) | 1 (6.3) | 1 (20.0) | 0 (0) |
| Gender | | | | | |
| Male | 128 (94.8) | 109 (99.1) | 16 (100) | 2 (40.0) | 1 (25.0) |
| Female | 7 (5.2) | 1 (0.9) | 0 (0) | 3 (60.0) | 3 (75.0) |
| Age | | | | | |
| <25 | 51 (37.8) | 42 (38.2) | 8 (50.0) | 0 (0) | 1 (25.0) |
| 25–34 | 48 (35.6) | 43 (39.1) | 4 (25.0) | 1 (20.0) | 0 (0) |
| 35–44 | 9 (6.7) | 6 (5.5) | 1 (6.3) | 0 (0) | 2 (50.0) |
| ≥45 | 27 (20.0) | 19 (17.3) | 3 (18.8) | 4 (80.0) | 1 (25.0) |
| Ethnic group | | | | | |
| Han | 113 (83.7) | 93 (84.5) | 14 (87.5) | 4 (80.0) | 2 (50.0) |
| Minority | 22 (16.3) | 17 (15.5) | 2 (12.5) | 1 (20.0) | 2 (50.0) |
| Marriage status | | | | | |
| Single | 99 (73.3) | 85 (77.3) | 12 (75.0) | 1 (20.0) | 1 (25.0) |
| Married | 14 (10.4) | 10 (9.1) | 2 (12.5) | 1 (20.0) | 1 (25.0) |
| Divorced | 21 (15.6) | 14 (12.7) | 2 (12.5) | 3 (60.0) | 2 (50.0) |
| UN | 1 (0.7) | 1 (0.9) | 0 (0) | 0 (0) | 0 (0) |
| Education | | | | | |
| Illiterate | 3 (2.2) | 2 (1.8) | 0 (0) | 0 (0) | 1 (25.0) |
| Primary education | 4 (3.0) | 4 (3.6) | 0 (0) | 0 (0) | 0 (0) |
| Secondary education | 30 (22.2) | 20 (18.2) | 6 (37.5) | 3 (60.0) | 1 (25.0) |
| Higher education | 98 (72.6) | 84 (76.4) | 10 (62.5) | 2 (40.0) | 2 (50.0) |
| Transmission routes | | | | | |
| MSM | 108 (80.0) | 90 (81.8) | 16 (100) | 1 (20.0) | 1 (25.0) |
| Hetero | 26 (19.3) | 20 (18.2) | 0 (0) | 4 (80.0) | 2 (50.0) |
| MTC | 1 (0.7) | 0 (0) | 0 (0) | 0 (0) | 1 (25.0) |
| LAg-Avidity EIA | | | | | |
| Recent | 43 (31.9) | 36 (32.7) | 5 (31.3) | 1 (20.0) | 1 (25.0) |
| LT | 68 (50.4) | 52 (47.3) | 10 (62.5) | 3 (60.0) | 3 (75.0) |
| NA | 24 (17.8) | 22 (20.0) | 1 (6.3) | 1 (20.0) | 0 (0) |

*MSM, men who have sex with men; Hetero, heterosexuals; MTC, mother-to-child; LAg-Avidity EIA, limiting-antigen avidity enzyme immunoassay; Recent, recent infection; LT, long-term infection; and NA, data not available.*

[3] https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html

## The Identification of Potential Transmission Clusters of HIV-1 URFs

To explore the phylogenetic relationship between URFs, the ML trees were reconstructed with *pol* sequences (HXB2: 2268-3278bp; **Figure 1**). Among 135 unique recombinant viruses, 16 potential transmission clusters with bootstrap values 1 and consistent recombination mode were inferred, including 70 subjects with at least two sequences in each cluster. Transmission clusters were composed of URFs_0107 (*n* = 60), URFs_01B (*n* = 8), and URF_BD (*n* = 2), respectively. The remaining 65 recombinants were interspersed widely throughout the phylogenetic tree.

Four larger URFs_0107 transmission clusters (IV, VI, VIII, and IX) were identified, consisting of 8–14 individuals (**Table 2**). Among them, cluster IV (*n* = 10) was observed between 2016 and 2019, and antibody affinity results suggested that 80% were chronically infected. More than half of infections were spread among older men (age range from 58 to 74) by heterosexual. In contrast, for the remaining three clusters, cluster VI (*n* = 14) and IX (*n* = 10) were diagnosed among the younger males who had been infected through a homosexual (*n* = 21) or heterosexual (*n* = 3) route and cluster VIII (*n* = 8) were transmitted only through homosexual behavior. In addition, more than 50% of individuals of three clusters were recent HIV-1 infections.

Also, the other seven URFs_0107 clusters containing two to four individuals were observed in male patients (**Table 2**). Three clusters (I, III, and XI) contained both homosexual and heterosexual transmission, among which cluster I (*n* = 4) patients were all recently infected, while clusters III and XI were chronic HIV-1 infections. Four clusters (II, V, VII, and X) were transmitted by homosexual contact only, and recent HIV-1 infections were detected in cluster II, V, and VII.

## Potential Recombination Hotspots in the *pol* Region

Subsequently, we conducted a simple recombination breakpoint scanning on the 110 URFs consisting of CRF01_AE and

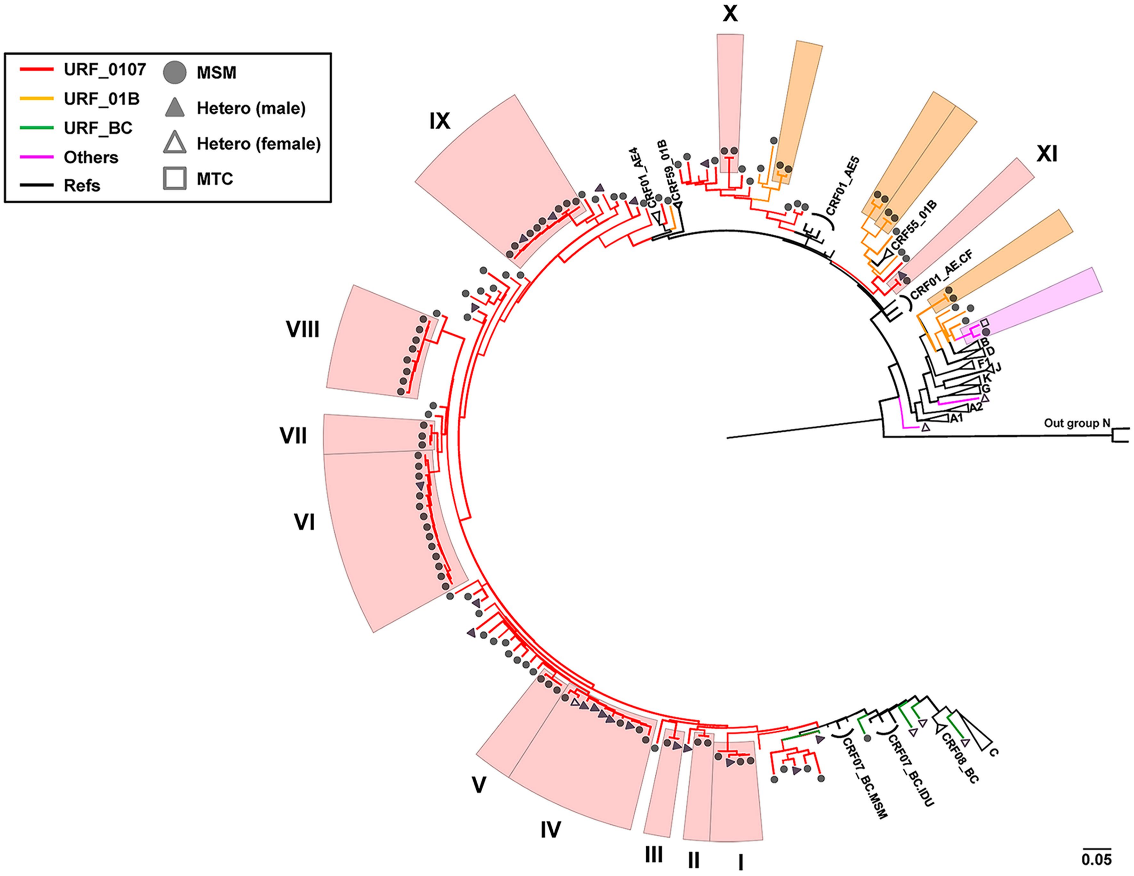


**FIGURE 1** | Phylogenetic tree analysis of all 135 HIV-1 URFs in Shenyang during 2016–2020. The topology of tree was constructed using Maximum-likelihood method under GTR+I+G model with 1,000 bootstrap replicates by IQ-TREE. ALL the reference sequences were downloaded from HIV-1 database and using the subtype N as outgroup. Red branches denoted genotypes URFs_0107, orange branches denoted genotypes URFs_01B, green branches denoted genotypes URFs_BC, and pink branches denoted genotypes other recombinants like URFs_BD, URFs_01A1, and URFs_02A1. Clusters with bootstrap value 1 were defined as the potential transmission clusters and indicated by shadows corresponding to the color of the genotypes, respectively. The sample of transmission routes was represented by different symbol (circle, MSM; triangle, Hetero; and square, MTC). The scale length indicated 5% nucleotide sequence divergence.

**TABLE 2 |** The main epidemiological characteristics of 11 clusters of URFs_0107 strains.

| | TOTAL (*n* =60) | Cluster | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **I (*n* =4)** | **II (*n* =2)** | **III (*n* =2)** | **IV (*n* =10)** | **V (*n* =3)** | **VI (*n* =14)** | **VII (*n* =3)** | **VIII(*n* =8)** | **IX (*n* =10)** | **X (*n* =2)** | **XI (*n* =2)** |
| | | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** | ***N* (%)** |
| Gender | | | | | | | | | | | | |
| Male | 59 (98.3) | 4 (100) | 2 (100) | 2 (100) | 9 (90.0) | 3 (100) | 14 (100) | 3 (100) | 8 (100) | 10 (100) | 2 (100) | 2 (100) |
| Female | 1 (1.7) | 0 (0) | 0 (0) | 0 (0) | 1 (10.0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Age | | | | | | | | | | | | |
| <25 | 27 (45.0) | 1 (25.0) | 0 (0) | 0 (0) | 0 (0) | 2 (66.7) | 8 (57.1) | 2 (66.7) | 5 (62.5) | 6 (60.0) | 2 (100) | 1 (50.0) |
| 25–34 | 18 (30.0) | 3 (75.0) | 1 (50.0) | 2 (100) | 0 (0) | 0 (0) | 3 (21.4) | 1 (33.3) | 3 (37.5) | 4 (40.0) | 0 (0) | 1 (50.0) |
| 35–44 | 3 (5.0) | 0 (0) | 1 (50.0) | 0 (0) | 0 (0) | 0 (0) | 2 (14.3) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| ≥45 | 12 (20.0) | 0 (0) | 0 (0) | 0 (0) | 10 (100) | 1 (33.3) | 1 (7.1) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| Transmission routes | | | | | | | | | | | | |
| MSM | 48 (80.0) | 3 (75.0) | 2 (100) | 1 (50.0) | 4 (40.0) | 3 (100) | 13 (92.9) | 3 (100) | 8 (100) | 8 (80.0) | 2 (100) | 1 (50.0) |
| Hetero | 12 (20.0) | 1 (25.0) | 0 (0) | 1 (50.0) | 6 (60.0) | 0 (0) | 1 (7.1) | 0 (0) | 0 (0) | 2 (20.0) | 0 (0) | 1 (50.0) |
| Diagnosis year | | | | | | | | | | | | |
| 2016 | 9 (15.0) | 0 (0) | 0 (0) | 0 (0) | 3 (30.0) | 0 (0) | 0 (0) | 0 (0) | 4 (50.0) | 1 (10.0) | 0 (0) | 1 (50.0) |
| 2017 | 9 (15.0) | 1 (25.0) | 0 (0) | 0 (0) | 1 (10.0) | 0 (0) | 5 (35.7) | 0 (0) | 2 (25.0) | 0 (0) | 0 (0) | 0 (0) |
| 2018 | 18 (30.0) | 1 (25.0) | 2 (100) | 2 (100) | 3 (30.0) | 1 (33.3) | 3 (21.4) | 2 (66.7) | 1 (12.5) | 3 (30.0) | 0 (0) | 0 (0) |
| 2019 | 18 (30.0) | 2 (50.0) | 0 (0) | 0 (0) | 3 (30.0) | 2 (66.7) | 3 (21.4) | 1 (33.3) | 1 (12.5) | 5 (50.0) | 0 (0) | 1 (50.0) |
| 2020 | 6 (10.0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 3 (21.4) | 0 (0) | 0 (0) | 1 (10.0) | 2 (100) | 0 (0) |
| LAg-Avidity EIA | | | | | | | | | | | | |
| Recent | 27 (45.0) | 4 (100) | 1 (50.0) | 0 (0) | 0 (0) | 3 (100) | 7 (50.0) | 2 (66.7) | 5 (62.5) | 5 (50.0) | 0 (0) | 0 (0) |
| LT | 23 (38.3) | 0 (0) | 1 (50.0) | 2 (100) | 8 (80.0) | 0 (0) | 4 (28.6) | 1 (33.3) | 2 (25.0) | 3 (30.0) | 0 (0) | 2 (100) |
| NA | 10 (16.7) | 0 (0) | 0 (0) | 0 (0) | 2 (20.0) | 0 (0) | 3 (21.4) | 0 (0) | 1 (12.5) | 2 (20.0) | 2 (100) | 0 (0) |

*MSM, men who have sex with men; Hetero, heterosexuals; MTC, mother-to-child; LAg-Avidity EIA, limiting-antigen avidity enzyme immunoassay; Recent, recent infection; LT, long-term infection; and NA, data not available.*

CRF07_BC. After excluding five sequences with ambiguous breakpoints, there were 191 breakpoint positions recorded among the 105 sequences, with some sequences having more than one breakpoint. These positions were plotted as a frequency plot of breakpoints in **Figure 2**. Overall, the breakpoints of URFs_0107 strains were widely distributed, covering almost the 1.1 kb *pol* genomes. One clear recombination peak can be observed with a position of 2,508–2,627 relative to HXB2 genome, near the junction of protease and reverse transcriptase genomes. There were 83.3% (50/60) of clustered and 24.4% (11/45) of non-clustered URFs observed breakpoints at this hotspot region, respectively. Moreover, we also found that some clustered URFs_0107 have similar breakpoints in other regions, such as cluster I and VII at 2848–2867, and the cluster I, IX, and XI at 2908–2967 (position relative to HXB2).

## Variety of Recombination Patterns and Distinct Parental CRF01_AE Origins in URFs_0107

To further dissect the HIV-1 URFs_0107 epidemics, we performed a detailed analysis of recombination patterns and parental origin in 1.1 kb *pol* region (**Figure 3**; **Supplementary Figure S2**), and the 5′ half-genome sequences of some specimens were used to supplement the evidence (**Supplementary Figure S3**). The parental origin of the recombinant CRF01_AE fragments was multiple. Of the 60 clustered URFs_0107 strains, the CRF01_AE fragments of clusters IV-VII originated from CRF01_AE lineage 4 (50%, 30/60), the clusters I-II and IX-XI derived from CRF01_AE lineage 5 (33.3%, 20/60), as well as the parental CRF01_AE of cluster VIII originated from CRF01_AE lineage Thailand (13.3%, 8/60). For the 45 non-clustered URFs_0107 strains, the proportion of parentals derived from CRF01_AE lineage 4, lineage 5, and lineage Thailand was 42.2% (19/45), 42.2% (19/45), and 2.2% (1/45), respectively. In addition, eight cases whose origin is not clear due to the shorter CRF01_AE fragments or ambiguity bases. In contrast, the CRF07_BC fragments of all URF_0107 strains were derived from CRF07_BC lineage predominant in MSM, which were significantly phylogenetically distinct from CRF07_BC strains prevalent among injecting drug users (Zhang et al., 2017).

The recombination structures of URFs_0107 were various. Among the URFs_0107 originals from CRF01_AE lineage 4, the common recombination structure was 01_AE/07_BC/01_AE (57.1%, 28/49), followed by 01_AE/07_BC (34.7%, 17/49), 01_AE/07_BC/01_AE/07_BC (4.1%, 2/49), and 07_BC/01_AE (4.1%, 2/49). And, five patterns were observed in the URFs_0107 originals from CRF01_AE lineage 5, including of 01_AE/07_BC/01_AE (43.6%, 17/39), 01_AE/07_BC (17.9%, 7/39), 07_BC/01_AE (20.5%, 8/39), 07_BC/01_AE/07_BC/01_AE/07_BC (15.4%, 6/39), and 01_AE/07_BC/01_AE/07_BC (2.6%, 1/39). In addition, only one recombination pattern of 01_AE/07_BC/01_AE/07_BC/01_AE was observed for URFs_0107 from CRF01_AE lineage Thailand. Of note, four non-clustered URF_0107 strains were found to have the CRF01_AE lineage 4 fragments closely related to strains in superinfection individual LNA819 under the subregion analysis (**Figure 4**). Our previous study showed that LNA819 was a CRF01_AE4/CRF07_BC

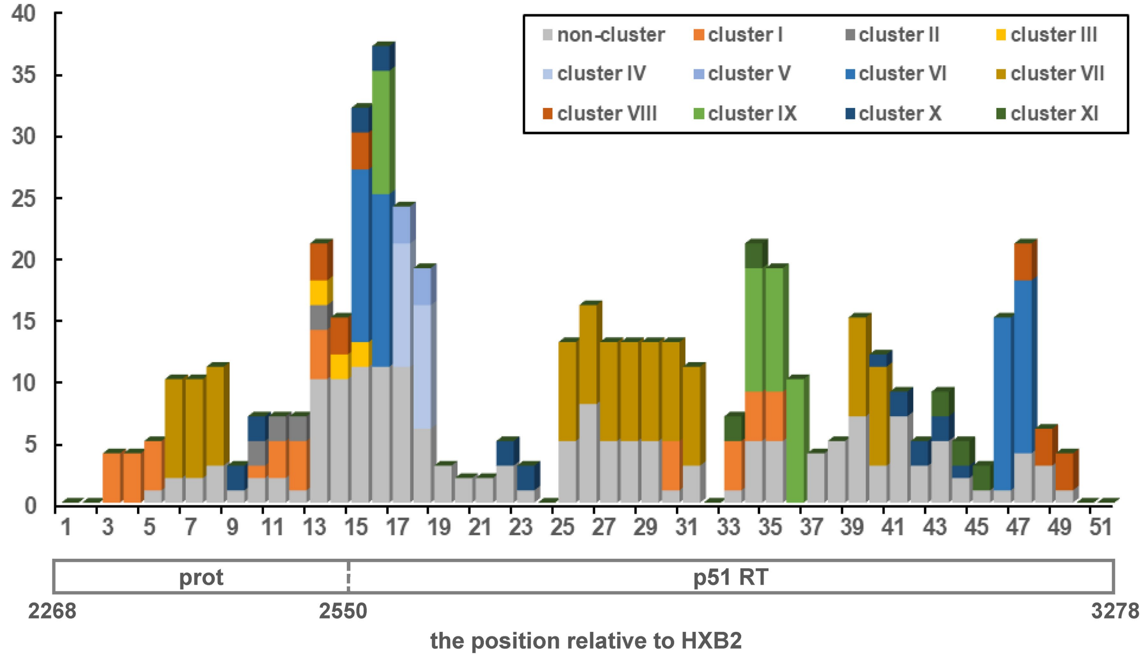


**FIGURE 2** | The distribution and frequency of breakpoints across pol region in HIV-1 URFs_0107 strains (HXB2: 2268-3278). The *x*-axis shown the position relative to the protease (prot) and reverse transcriptase (RT), while the *y*-axis indicated the breakpoint frequency. The 1.1-kb *pol* region was divided into 51 windows using 20 non-overlapping nucleotide steps. Eleven transmission clusters were color-coded, and the non-clustered sequences were uniformly represented in gray.

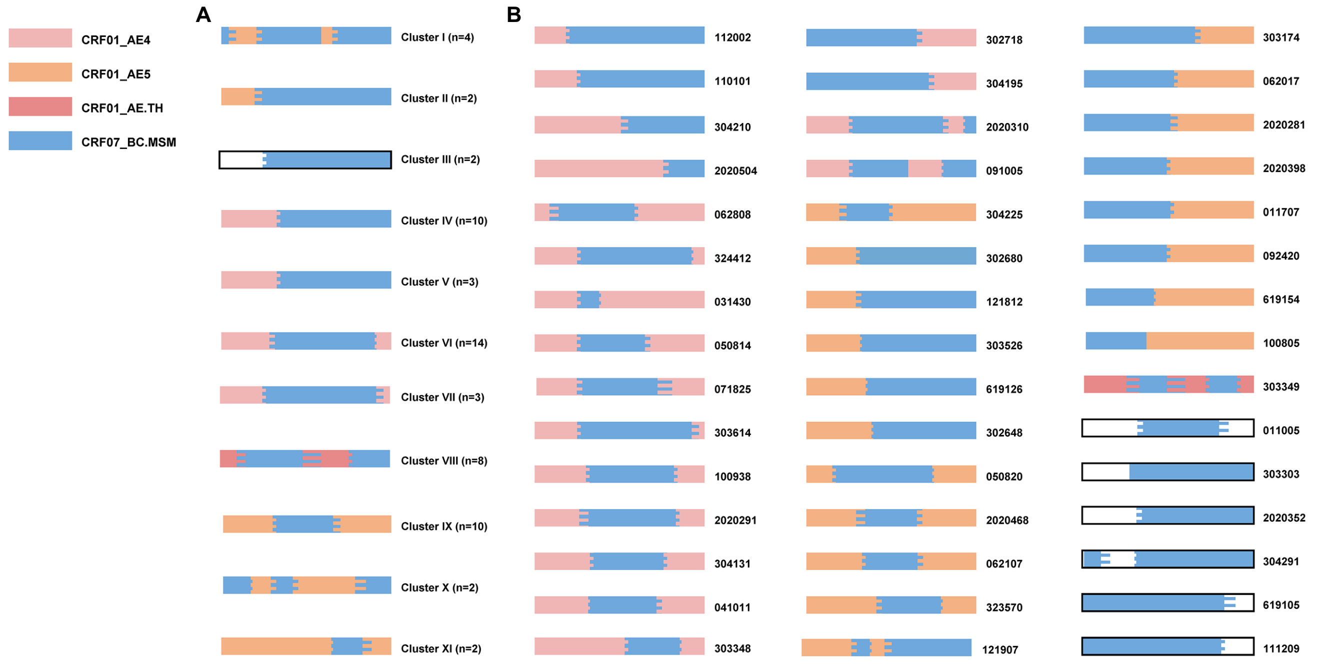


**FIGURE 3 |** The recombination patterns and parental origin of partial *pol* region in URFs_0107 clustered sequences **(A)** and non-clustered sequences **(B)**. The recombination structures were confirmed by RIP, JPHMM, and Simplot software (v 3.5.1). The parent origins of the fragments were represented by different colors, among which blank represented that the definite lineages of CRF01_AE were not identified due to the short sequence fragments or ambiguity bases. The sample codes were indicated on the right side of the mosaic map. The mosaic genetic map was generated by the Recombinant HIV-1 Drawing online tool (www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html).

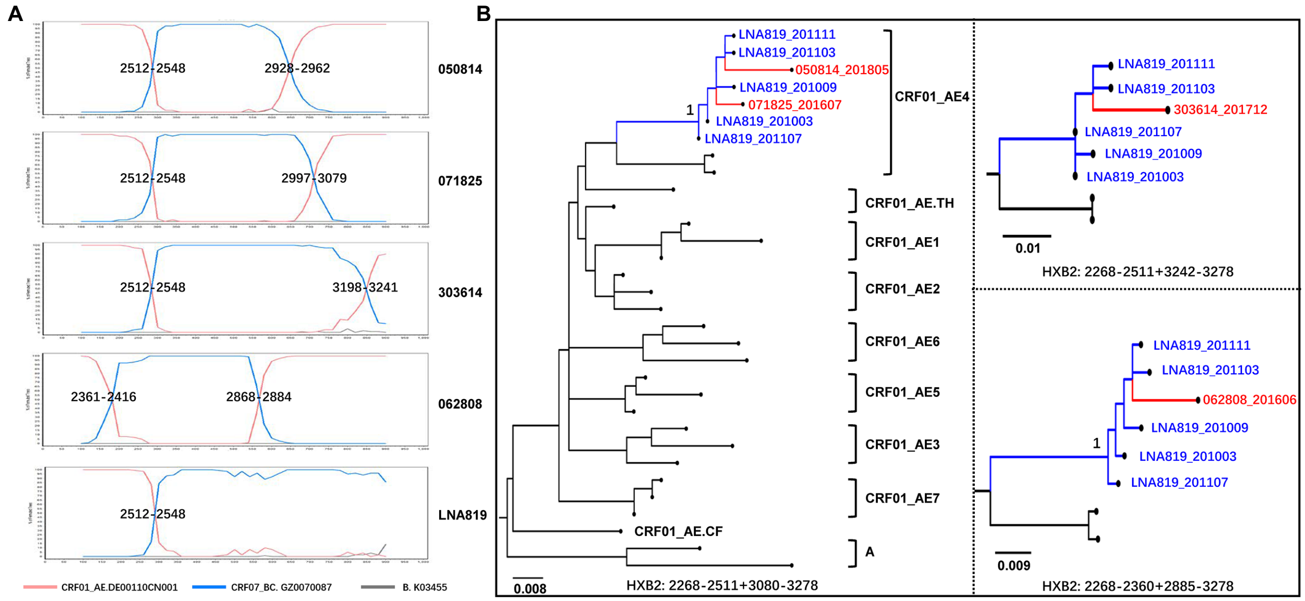


**FIGURE 4 |** The recombination structure **(A)** and evolutionary relationship **(B)** of four URFs_0107 strains and LNA819. **(A)** Recombination analysis of four non-clustered URFs_0107 in the present study and LNA819 related strain (MT857722) by SimPlot (v3.5.1). **(B)** The subregion trees of CRF01_AE homologous region of recombinants were constructed by IQ-TREE under the GTR+I+G model with 1,000 replications. The red branches represented the URFs_0107 in this study, and the blue branches represented the continuous-time CRF01_AE strains in LNA819. The reference sequences were downloaded from the HIV-1 database, including seven lineages of CRF01_AE in China, lineage Thailand, and Africa, as well as subtype A. Only the key bootstrap values are shown.

superinfection with extremely active high-risk behavior, and a variety of distinct recombinants produced in LNA819 have been detected in the other five infected patients (Gao et al., 2021). The recombination structure of four URFs_0107 was 01_AE/07_BC/01_AE, and three of them had identical breakpoint at the position of 2,512–2,548, which also be seen in LNA819 (**Figure 4A**). The topologies of the subregion tree showed that the CRF01_AE fragments from four URFs_0107 were rooted in the pure CRF01_AE strains from LNA819 (five series of sampling points between 2010 and 2011), formed a monophyletic cluster with bootstraps value 1 (**Figure 4B**). Taken together, the similar breakpoints and high homologous parental strains suggested a close evolutionary relationship among four URFs_0107 strains and LNA819.

## DISCUSSION

We performed a detailed characterization of the molecular epidemiology of URFs in Shenyang, northeast China from 2016 to 2020. The recombination patterns of URFs were complicated and dominated by CRF01_AE/CRF07_BC. The parentals of URFs_0107 were derived from the strains circulating among MSM, and the high level of recent infections indicated that URFs_0107 is continually emerging and spreading. Our data suggested the importance of continuous and accurate molecular epidemiology surveillance of recombinants, as well as provided an application method to molecular tracing of HIV-1 URFs for other regions in China where predominantly transmit HIV-1 among MSM population.

First, this study confirmed CRF01_AE/CRF07_BC was the most prevalent HIV-1 URFs in Shenyang over the last 5 years. Multiple subtypes have been reported from the beginning of the HIV epidemic in Shenyang. In the early 2000s, HIV-1 transmission was mainly through blood donation/transfusion and heterosexual contact in Shenyang, and the subtype B/B′ and CRF01_AE were the common prevalence strains (Han et al., 2001, 2010). Thus, the CRF01_AE/B was the dominant recombinants during the same period (Han et al., 2010, 2013b). However, since the year 2006, the prevalence of HIV-1 among MSM population has rapidly expanded, by far, the homosexual has transformed into the predominant risk-group of HIV-1 transmission with dominated circulating subtype CRF01_AE and CRF07_BC (An et al., 2012; Han et al., 2013c). Based on our recent analysis of the *pol* gene, the prevalence of CRF01_AE and subtype B gradually declined in the city during 2008–2016, while a significant increase of subtype CRF07_BC and other CRFs/URFs was reported in Shenyang (Liu et al., 2020). The trend with a higher proportion of URFs indicated the increasing viral complexity of HIV-1 epidemic in Shenyang and the similar situation would also be present in other regions of the national.

Second, the URFs_0107 identified in this study had distinct origin of CRF01_AE fragments, including the CRF01_AE lineage 4 and lineage 5, which have been reported to be widespread among Chinese MSM (Wang et al., 2017), as well as the lineage from Thailand, which is distinct from the seven lineages of CRF01_AE commonly circulating in China (Feng et al., 2013). In fact, the prevalence of CRF01_AE lineage 5 (75.3%) is much higher than lineage 4 (21.2%) in Shenyang, compared with other cities (Han et al., 2013c; An et al., 2021). However, our results showed that the number of URFs_0107 whose parents were CRF01_AE lineage 4 ($n = 49$) was higher than that of lineage 5 ($n = 39$). This is an indication that the URFs_0107 with CRF01_AE4 as parent might have been introduced into the Shenyang from other provinces, or patients infected with the CRF01_AE4/CRF07_BC recombinants have more active risk behaviors, resulting in widespread local transmission. Additionally, four large potential transmission clusters were found (IV, VI, VIII, and IX), indicating that these recombinants are already obtaining the ability to spread in this region (**Table 2**). It has been reported that recombinant forms can display enhanced replicative fitness compared with parental strains (Njai et al., 2006; Lau et al., 2010) and can have increased pathogenicity and virulence (Fischetti et al., 2004). Meanwhile, cross-transmission events between MSM and heterosexual were identified in six transmission clusters (I, III, IV, VI, VIII, IX, and XI) which may increase the risk of further expansion of the URFs_0107 strains to general population in Shenyang (**Table 2**). And, there is a significant imbalance between male ($n = 11$) and female ($n = 1$) patients infected through heterosexual transmission, indicating the possibility of male patients concealing their true sexual orientation.

Third, the determination and intervention of the transmission sources are of great significance for identifying how strains continue to spread and how to avoid new infections effectively. Phylogenetic and distance-based analysis are common strategies to determine the possible transmission networks (Trask et al., 2002; Smith et al., 2009; Little et al., 2014). However, when many recombinants are put together for analysis, they tend to affect the topology of the phylogenetic tree due to the distribution of recombination breakpoints. Recombinants with similar breakpoints are more likely to cluster together in the evolutionary tree, but their parental origin may be different (He et al., 2020). Meanwhile, multiple infections can produce various recombinants, and the strains with different backbone or breakpoints from the same source on the evolutionary tree may not cluster. This phenomenon also appeared in the patient of this study. The parental of CRF01_AE fragment within the four non-clustered strains with different recombination breakpoints was traced back to LNA819, a superinfected person previously identified. The case was diagnosed with HIV-1 CRF01_AE infection in March 2010, re-infected by CRF07_BC in December 2010, and started antiretroviral therapy in 2014 (Luan et al., 2017; Gao et al., 2021). Our analysis showed that four URFs_0107 identified in 2016–2018 were closely related to the evolution of the pure CRF01_AE strains in LNA819, suggesting the possibility of indirect transmission. The patient LNA819 could generate and transmit various recombinants during the treatment-naïve period, or transmit pure CRF_01AE strains to recipients, who could be superinfected with another CRF07_BC strain to develop a series of new URFs_0107 for further spread. It reminded us that the transmission network associated with LNA819 was not effectively interfered, and the related strains could spread lasted up for 8 years. Thus, it is strongly necessary to trace the source of the homologous recombinant fragments for inferring the potential transmission chain, control the early stages of infected patients in timely, monitor the uninfected persons who have contact with the infected, and guide the implementation of Pre-Exposure Prophylaxis (PrEP).

Finally, we found that 58.1% (61/105) of URFs_0107 have similar breakpoints in *pol* region around the position of the junction of protease and reverse transcriptase (**Figure 2**), indicating that this region might be the hotspots of recombination. There may be several reasons for similar breakpoints in recombinants. For example, the similar recombination breakpoints of CRF07_BC and CRF08_BC were attributed to the back-cross between common URF_BC and subtype C in Yunnan, China (McClutchan et al., 2002); CRF74_01B shared four and two breakpoints with CRF33_01B and CRF53_01B, respectively, because CRF74_01B may be a direct descendant of these two CRFs (Cheong et al., 2015). In addition to the above-mentioned recombinants that contain evolutionary relationships, another reason for the formation of common breakpoints is due to the sequence similarity, hairpin structure of genomic RNA, fragile and pause sites, and high-pairing probability (Galetto et al., 2004; Galetto and Negroni, 2005; Delviks-Frankenberry et al., 2011; Jia et al., 2016). The other studies also found hotspots at the genomic junction of protease and reverse transcriptase, which was consistent with our findings (Galli et al., 2008; Smyth et al., 2014).

The major limitations of this study are that (1) only by 1.1-kb *pol* region sequence may underestimate the prevalence of recombination. However, using the pol sequences from the routine drug-resistant test will contribute to the real-time monitor the change of HIV recombinants, which is the most focus in this study; (2) we tried to obtain as many as 5′ half-genome sequences as possible, but considering time and availability of resources, only 14 patients 5'half-genome sequences were amplified to support or verify the recombination analysis of the pol region. In the future, we need to further obtain more 5′-half-genome or the nearly full-length genome sequences of patients to trace the source of the recombinants, clarify the evolutionary relationship; and (3) due to the protection of personal privacy, we do not have enough epidemiological data to determine whether there is a real direct/indirect transmission relationship among clustered URFs.

In conclusion, our study provides molecular evidence that the unique recombination patterns lead to a complex HIV-1 epidemic in Shenyang. In recent years, the transmission of URF_0107 strains among MSM in Shenyang has increased, some of which have caused more widespread transmission. Our study highlights the importance of continued and accurate molecular surveillance to increase our understanding of the evolving HIV-1 URFs epidemic. And tracing the source of recombinants is necessary, which helps provide specific interventions to the most relevant high-risk population.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the First Affiliated Hospital of China Medical University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

HS and XH conceived and designed the study. WS and QL collected the specimen information. SH, GG, LW, YG, and WT performed experimental work. MA and BZ performed data collection. SH performed data analysis and wrote the first draft. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was funded by the National Natural Science Foundation of China (81871637 and 82072272), CAMS Innovation Fund for Medical Sciences (2019-I2M-027), and Science and Technology Project of Shenyang (19-104-4-026).

## ACKNOWLEDGMENTS

We are grateful to the patients who consented to participate and donate their blood samples for this study. We thank all the researchers who designed this study and contributed to the preparation of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.855049/full#supplementary-material

## REFERENCES

An, M., Han, X., Xu, J., Chu, Z., Jia, M., Wu, H., et al. (2012). Reconstituting the epidemic history of HIV strain CRF01_AE among men who have sex with men (MSM) in Liaoning, northeastern China: implications for the expanding epidemic among MSM in China. *J. Virol.* 86, 12402–12406. doi: 10.1128/JVI.00262-12

An, M., Zhao, B., Wang, L., Chu, Z., Xu, J., Ding, H., et al. (2021). The viral founder effect and economic-driven human mobility shaped the distinct epidemic pattern of HIV-1 CRF01_AE in Northeast China. *Front. Med.* 8:769535. doi: 10.3389/fmed.2021.769535

Chang, W., Zhang, M., Ren, Q., Zou, Y., Dong, L., Jia, H., et al. (2020). HIV-1 genetic diversity and recombinant forms among men who have sex with men at a sentinel surveillance site in Xi'an City, China. *Infect. Genet. Evol.* 81:104257. doi: 10.1016/j.meegid.2020.104257

Chen, M., Ma, Y., Chen, H., Dai, J., Luo, H., Yang, C., et al. (2019). Spatial clusters of HIV-1 genotypes in a recently infected population in Yunnan, China. *BMC Infect. Dis.* 19:669. doi: 10.1186/s12879-019-4276-9

Cheong, H. T., Chow, W. Z., Takebe, Y., Chook, J. B., Chan, K. G., Al-Darraji, H. A., et al. (2015). Genetic characterization of a novel HIV-1 circulating recombinant form (CRF74_01B) identified among intravenous drug users in Malaysia: recombination history and phylogenetic linkage with previously defined recombinant lineages. *PLoS One* 10:e0133883. doi: 10.1371/journal.pone.0133883

Delviks-Frankenberry, K., Galli, A., Nikolaitchik, O., Mens, H., Pathak, V. K., and Hu, W. S. (2011). Mechanisms and factors that influence high frequency retroviral recombination. *Viruses* 3, 1650–1680. doi: 10.3390/v3091650

Feng, Y., He, X., Hsi, J. H., Li, F., Li, X., Wang, Q., et al. (2013). The rapidly expanding CRF01_AE epidemic in China is driven by multiple lineages of HIV-1 viruses introduced in the 1990s. *AIDS* 27, 1793–1802. doi: 10.1097/QAD.0b013e328360db2d

Feng, Y., Takebe, Y., Wei, H., He, X., Hsi, J. H., Li, Z., et al. (2016). Geographic origin and evolutionary history of China's two predominant HIV-1 circulating recombinant forms, CRF07_BC and CRF08_BC. *Sci. Rep.* 6:19279. doi: 10.1038/srep19279

Fischetti, L., Opare-Sem, O., Candotti, D., Lee, H., and Allain, J. P. (2004). Higher viral load may explain the dominance of CRF02_AG in the molecular epidemiology of HIV in Ghana. *AIDS* 18, 1208–1210. doi: 10.1097/00002030-200405210-00017

Galetto, R., Moumen, A., Giacomoni, V., Veron, M., Charneau, P., and Negroni, M. (2004). The structure of HIV-1 genomic RNA in the gp120 gene determines a recombination hot spot in vivo. *J. Biol. Chem.* 279, 36625–36632. doi: 10.1074/jbc.M405476200

Galetto, R., and Negroni, M. (2005). Mechanistic features of recombination in HIV. *AIDS Rev.* 7, 92–102.

Galli, A., Lai, A., Corvasce, S., Saladini, F., Riva, C., Deho, L., et al. (2008). Recombination analysis and structure prediction show correlation between breakpoint clusters and RNA hairpins in the pol gene of human immunodeficiency virus type 1 unique recombinant forms. *J. Gen. Virol.* 89, 3119–3125. doi: 10.1099/vir.0.2008/003418-0

Gao, Y., He, S., Tian, W., Li, D., An, M., Zhao, B., et al. (2021). First complete-genome documentation of HIV-1 intersubtype superinfection with transmissions of diverse recombinants over time to five recipients. *PLoS Pathog.* 17:e1009258. doi: 10.1371/journal.ppat.1009258

Han, X., An, M., Zhang, W., Cai, W., Chen, X., Takebe, Y., et al. (2013a). Genome sequences of a novel HIV-1 circulating recombinant form, CRF55_01B, identified in China. *Genome Announc.* 1:e00050-12. doi: 10.1128/genomeA.00050-12

Han, X., An, M., Zhang, W., Zhao, B., Chu, Z., Takebe, Y., et al. (2013b). Genome sequences of a novel HIV-1 circulating recombinant form (CRF59_01B) identified among men who have sex with men in northeastern China. *Genome Announc.* 1:e00315-13. doi: 10.1128/genomeA.00315-13

Han, X., An, M., Zhang, M., Zhao, B., Wu, H., Liang, S., et al. (2013c). Identification of 3 distinct HIV-1 founding strains responsible for expanding epidemic among men who have sex with men in 9 Chinese cities. *J. Acquir. Immune Defic. Syndr.* 64, 16–24. doi: 10.1097/QAI.0b013e3182932210

Han, X., Dai, D., Zhao, B., Liu, J., Ding, H., Zhang, M., et al. (2010). Genetic and epidemiologic characterization of HIV-1 infection in Liaoning Province, China. *J. Acquir. Immune Defic. Syndr.* 53(Suppl. 1), S27–S33. doi: 10.1097/QAI.0b013e3181c7d5bf

Han, X., Jiang, Y., and Shang, H. (2001). Genetic subtyping of HIV-1 in Liaoning province of China. *Zhonghua Liu Xing Bing Xue Za Zhi* 22, 432–434.

Hao, M., Wang, J., He, S., Hao, Y., Ye, J., Xin, R., et al. (2019). Identification of a novel HIV-1 second-generation recombinant form (CRF01_AE/07_BC) in men who have sex with men in Beijing, China. *AIDS Res. Hum. Retrovir.* 35, 500–504. doi: 10.1089/aid.2018.0228

He, S., Gao, Y., An, M., Zhao, B., Wang, L., Ding, H., et al. (2020). Characterization of a novel HIV-1 CRF01_AE/CRF07_BC recombinant strain among men who have sex with men in Liaoning, China. *AIDS Res. Hum. Retrovir.* 37, 70–74. doi: 10.1089/AID.2020.0223

Hemelaar, J. (2013). Implications of HIV diversity for the HIV-1 pandemic. *J. Infect.* 66, 391–400. doi: 10.1016/j.jinf.2012.10.026

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020a). Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781. doi: 10.1016/s2352-3018(20)30252-6

Hemelaar, J., Loganathan, S., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., et al. (2020b). Country-level diversity of the HIV-1 pandemic between 1990 and 2015. *J. Virol.* 95:e01580-20. doi: 10.1128/JVI.01580-20

Jia, L., Li, L., Gui, T., Liu, S., Li, H., Han, J., et al. (2016). Analysis of HIV-1 intersubtype recombination breakpoints suggests region with high pairing probability may be a more fundamental factor than sequence similarity affecting HIV-1 recombination. *Virol. J.* 13:156. doi: 10.1186/s12985-016-0616-1

Jiang, J., Liang, B., Li, K., Yang, Y., Yang, Y., Ning, C., et al. (2020). Genomic characterization of a novel HIV type 1 strain originating from CRF07_BC and CRF01_AE by heterosexual transmission in the Lingshan prefecture of Guangxi Province, China. *AIDS Res. Hum. Retrovir.* 36, 153–160. doi: 10.1089/AID.2019.0182

Jin-ping, M., Wei, S., and Lu, W. (2018). Analysis of epidemiological characteristics of HIV/AIDS in Shenyang from 2008 to 2017. *Mod. Prev. Med.* 45, 2894–2897+2949.

Lau, K. A., Wang, B., Miranda-Saksena, M., Boadle, R., Kamarulzaman, A., Ng, K. P., et al. (2010). Evidence for possible biological advantages of the newly emerging HIV-1 circulating recombinant form from Malaysia - CRF33_01B in comparison to its progenitors - CRF01_AE and subtype B. *Curr. HIV Res.* 8, 259–271. doi: 10.2174/157016210791111151

Li, X., Wu, J., Zhang, Y., Shen, Y., Li, H., Xing, H., et al. (2019). Characterization of a novel HIV-1 second-generation circulating recombinant form (CRF102_0107) among men who have sex with men in Anhui, China. *J. Infect.* 79, 612–625. doi: 10.1016/j.jinf.2019.09.022

Little, S. J., Kosakovsky Pond, S. L., Anderson, C. M., Young, J. A., Wertheim, J. O., Mehta, S. R., et al. (2014). Using HIV networks to inform real time prevention interventions. *PLoS One* 9:e98443. doi: 10.1371/journal.pone.0098443

Liu, M., Han, X., Zhao, B., An, M., He, W., Wang, Z., et al. (2020). Dynamics of HIV-1 molecular networks reveal effective control of large transmission clusters in an area affected by an epidemic of multiple HIV subtypes. *Front. Microbiol.* 11:604993. doi: 10.3389/fmicb.2020.604993

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/jvi.73.1.152-160.1999

Luan, H., Han, X., Yu, X., An, M., Zhang, H., Zhao, B., et al. (2017). Dual infection contributes to rapid disease progression in men who have sex with men in China. *J. Acquir. Immune Defic. Syndr.* 75, 480–487. doi: 10.1097/qai.0000000000001420

McClutchan, F. E., Carr, J. K., Murphy, D., Piyasirisilp, S., Gao, F., Hahn, B., et al. (2002). Precise mapping of recombination breakpoints suggests a common parent of two BC recombinant HIV type 1 strains circulating in China. *AIDS Res. Hum. Retrovir.* 18, 1135–1140. doi: 10.1089/088922202320567879

Moutouh, L., Corbeil, J., and Richman, D. D. (1996). Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proc. Natl. Acad. Sci. U. S. A.* 93, 6106–6111. doi: 10.1073/pnas.93.12.6106

Nájera, R., Delgado, E., Pérez-Alvarez, L., and Thomson, M. M. (2002). Genetic recombination and its role in the development of the HIV-1 pandemic. *AIDS* 16(Suppl. 4), S3–S16. doi: 10.1097/00002030-200216004-00002

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nikolaitchik, O., Keele, B., Gorelick, R., Alvord, W. G., Mazurov, D., Pathak, V. K., et al. (2015). High recombination potential of subtype A HIV-1. *Virology* 484, 334–340. doi: 10.1016/j.virol.2015.06.025

Njai, H. F., Gali, Y., Vanham, G., Clybergh, C., Jennes, W., Vidal, N., et al. (2006). The predominance of human immunodeficiency virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in west Central Africa may be related to its replicative fitness. *Retrovirology* 3:40. doi: 10.1186/1742-4690-3-40

Nora, T., Charpentier, C., Tenaillon, O., Hoede, C., Clavel, F., and Hance, A. J. (2007). Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *J. Virol.* 81, 7620–7628. doi: 10.1128/jvi.00083-07

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2: approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Qi, J., Zhang, D., Fu, X., Li, C., Meng, S., Dai, M., et al. (2015). High risks of HIV transmission for men who have sex with men--a comparison of risk factors of HIV infection among MSM associated with recruitment channels in 15 cities of China. *PLoS One* 10:e0121267. doi: 10.1371/journal.pone.0121267

Sanchez, A. M., DeMarco, C. T., Hora, B., Keinonen, S., Chen, Y., Brinkley, C., et al. (2014). Development of a contemporary globally diverse HIV viral panel by the EQAPOL program. *J. Immunol. Methods* 409, 117–130. doi: 10.1016/j.jim.2014.01.004

Smith, D. M., May, S. J., Tweeten, S., Drumright, L., Pacold, M. E., Kosakovsky Pond, S. L., et al. (2009). A public health model for the molecular surveillance of HIV transmission in San Diego, California. *AIDS* 23, 225–232. doi: 10.1097/QAD.0b013e32831d2a81

Smyth, R. P., Schlub, T. E., Grimm, A. J., Waugh, C., Ellenberg, P., Chopra, A., et al. (2014). Identifying recombination hot spots in the HIV-1 genome. *J. Virol.* 88, 2891–2902. doi: 10.1128/JVI.03014-13

Streeck, H., Li, B., Poon, A. F., Schneidewind, A., Gladden, A. D., Power, K. A., et al. (2008). Immune-driven recombination and loss of control after HIV superinfection. *J. Exp. Med.* 205, 1789–1796. doi: 10.1084/jem.20080281

Tippmann, H. F. (2004). Analysis for free: comparing programs for sequence analysis. *Brief. Bioinform.* 5, 82–87. doi: 10.1093/bib/5.1.82

Trask, S. A., Derdeyn, C. A., Fideli, U., Chen, Y., Meleth, S., Kasolo, F., et al. (2002). Molecular epidemiology of human immunodeficiency virus type 1 transmission in a heterosexual cohort of discordant couples in Zambia. *J. Virol.* 76, 397–405. doi: 10.1128/jvi.76.1.397-405.2002

Vuilleumier, S., and Bonhoeffer, S. (2015). Contribution of recombination to the evolutionary history of HIV. *Curr. Opin. HIV AIDS* 10, 84–89. doi: 10.1097/COH.0000000000000137

Wang, X., He, X., Zhong, P., Liu, Y., Gui, T., Jia, D., et al. (2017). Phylodynamics of major CRF01_AE epidemic clusters circulating in mainland of China. *Sci. Rep.* 7:6330. doi: 10.1038/s41598-017-06573-6

Wu, Z., Xu, J., Liu, E., Mao, Y., Xiao, Y., Sun, X., et al. (2013). HIV and syphilis prevalence among men who have sex with men: a cross-sectional survey of 61 cities in China. *Clin. Infect. Dis.* 57, 298–309. doi: 10.1093/cid/cit210

Xu, J. J., Zhang, M., Brown, K., Reilly, K., Wang, H., Hu, Q., et al. (2010). Syphilis and HIV seroconversion among a 12-month prospective cohort of men who have sex with men in Shenyang, China. *Sex. Transm. Dis.* 37, 432–439. doi: 10.1097/OLQ.0b013e3181d13eed

Yin, Y., Liu, Y., Zhu, J., Hong, X., Yuan, R., Fu, G., et al. (2019). The prevalence, temporal trends, and geographical distribution of HIV-1 subtypes among men who have sex with men in China: a systematic review and meta-analysis. *Epidemiol. Infect.* 147:e83. doi: 10.1017/s0950268818003400

Yuan, R., Cheng, H., Chen, L. S., Zhang, X., and Wang, B. (2016). Prevalence of different HIV-1 subtypes in sexual transmission in China: a systematic review and meta-analysis. *Epidemiol. Infect.* 144, 2144–2153. doi: 10.1017/S0950268816000212

Zai, J., Liu, H., Lu, Z., Chaillon, A., Smith, D., Li, Y., et al. (2020). Tracing the transmission dynamics of HIV-1 CRF55_01B. *Sci. Rep.* 10:5098. doi: 10.1038/s41598-020-61870-x

Zhang, M., Foley, B., Schultz, A. K., Macke, J. P., Bulla, I., Stanke, M., et al. (2010). The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25

Zhang, M., Jia, D., Li, H., Gui, T., Jia, L., Wang, X., et al. (2017). Phylodynamic analysis revealed that epidemic of CRF07_BC strain in men who have sex with men drove its second spreading wave in China. *AIDS Res. Hum. Retrovir.* 33, 1065–1069. doi: 10.1089/AID.2017.0091

Zhao, B., Han, X., Dai, D., Liu, J., Ding, H., Xu, J., et al. (2011). New trends of primary drug resistance among HIV type 1-infected men who have sex with men in Liaoning Province, China. *AIDS Res. Hum. Retrovir.* 27, 1047–1053. doi: 10.1089/AID.2010.0119

Zhao, B., Song, W., Kang, M., Dong, X., Li, X., Wang, L., et al. (2021). Molecular network analysis reveals transmission of HIV-1 drug-resistant strains among newly diagnosed HIV-1 infections in a moderately HIV Endemic City in China. *Front. Microbiol.* 12:797771. doi: 10.3389/fmicb.2021.797771

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 May 2022
doi: 10.3389/fmicb.2022.863123

# The Origin, Epidemiology, and Phylodynamics of Human Immunodeficiency Virus Type 1 CRF47_BF

*Gracelyn Hill[1], Marcos Pérez-Losada[1,2,3], Elena Delgado[4], Sonia Benito[4], Vanessa Montero[4], Horacio Gil[4], Mónica Sánchez[4], Javier E. Cañada-García[4], Elena García-Bodas[4], Keith A. Crandall[1,2]*, Michael M. Thomson[4]* and the Spanish Group for the Study of New HIV Diagnoses*

*[1] Computational Biology Institute, George Washington University, Washington, DC, United States, [2] Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC, United States, [3] CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Porto, Portugal, [4] HIV Biology and Variability Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain*

OPEN ACCESS

***Edited by:***
*Mariza Gonçalves Morgado,*
*Oswaldo Cruz Foundation (FIOCRUZ),*
*Brazil*

***Reviewed by:***
*Gonzalo Jose Bello Bentancor,*
*Oswaldo Cruz Foundation (FIOCRUZ),*
*Brazil*
*Leandro Jones,*
*Consejo Nacional de Investigaciones*
*Científicas y Técnicas (CONICET),*
*Argentina*

****Correspondence:***
*Keith A. Crandall*
*kcrandall@gwu.edu*
*Michael M. Thomson*
*mthomson@isciii.es*

***Specialty section:***
*This article was submitted to*
*Virology,*
*a section of the journal*
*Frontiers in Microbiology*

***Received:*** *26 January 2022*
***Accepted:*** *14 April 2022*
***Published:*** *16 May 2022*

***Citation:***
*Hill G, Pérez-Losada M, Delgado E, Benito S, Montero V, Gil H, Sánchez M, Cañada-García JE, García-Bodas E, Crandall KA, Thomson MM and the Spanish Group for the Study of New HIV Diagnoses (2022) The Origin, Epidemiology, and Phylodynamics of Human Immunodeficiency Virus Type 1 CRF47_BF.*
*Front. Microbiol. 13:863123.*
*doi: 10.3389/fmicb.2022.863123*

CRF47_BF is a circulating recombinant form (CRF) of the human immunodeficiency virus type 1 (HIV-1), the etiological agent of AIDS. CRF47_BF represents one of 19 CRFx_BFs and has a geographic focus in Spain, where it was first identified in 2010. Since its discovery, CRF47_BF has expanded considerably in Spain, predominantly through heterosexual contact (∼56% of the infections). Little is known, however, about the origin and diversity of this CRF or its epidemiological correlates, as very few samples have been available so far. This study conducts a phylogenetic analysis with representatives of all CRFx_BF sequence types along with HIV-1 M Group subtypes to validate that the CRF47_BF sequences share a unique evolutionary history. The CRFx_BF sequences cluster into a single, not well supported, clade that includes their dominant parent subtypes (B and F). This clade also includes subtype D and excludes sub-subtype F2. However, the CRF47_BF sequences all share a most recent common ancestor. Further analysis of this clade couples CRF47_BF protease-reverse transcriptase sequences and epidemiological data from an additional 87 samples collected throughout Spain, as well as additional CRF47_BF database sequences from Brazil and Spain to investigate the origin and phylodynamics of CRF47_BF. The Spanish region with the highest proportion of CRF47_BF samples in the data set was the Basque Country (43.7%) with Navarre next highest at 19.5%. We include in our analysis epidemiological data on host sex, mode of transmission, time of collection, and geographic region. The phylodynamic analysis indicates that CRF47_BF originated in Brazil around 1999–2000 and spread to Spain from Brazil in 2002–2003. The virus spread rapidly throughout Spain with an increase in population size from 2011 to 2015 and leveling off more recently. Three strongly supported clusters associated with Spanish regions (Basque Country, Navarre, and Aragon), together comprising 60.8% of the Spanish samples, were identified, one of which was also associated with transmission among men who have sex with men.

The expansion in Spain of CRF47_BF, together with that of other CRFs and subtype variants of South American origin, previously reported, reflects the increasing relationship between the South American and European HIV-1 epidemics.

**Keywords: HIV, phylodynamics, circulating recombinant form (CRF), Spain, epidemiology, CRF47_BF**

## INTRODUCTION

High genetic diversity of human immunodeficiency virus type 1 (HIV-1) is a defining feature of the AIDS virus. This diversity gain and loss is a hallmark of the evolution of HIV in the context of drug resistance and changing environments (Pennings et al., 2014). A contributing factor in the evolution of HIV is the process of recombination (Rambaut et al., 2004; Vuilleumier and Bonhoeffer, 2015). Genetic recombination is known to impact HIV allelic diversity and subsequent population dynamics at a rate equivalent to the high mutation rate of HIV (Shriner et al., 2004). Genetic diversity within HIV subtypes can be up to 17% sequence divergence across the genome with 17–35% divergence between subtypes (Castro-Nallar et al., 2012a). Yet recombination can even occur between subtypes as HIV variants spread around the globe, leading to circulating recombinant forms or CRFs, as well as unique recombinant forms (URFs) (Castro-Nallar et al., 2012b). There are currently 118 known HIV-1 CRFs according to the Los Alamos HIV Sequence Database (Los Alamos National Laboratory, 2021) involving recombination events between nearly all known subtypes and even between other CRFs [e.g., CRF15_01B is a recombinant form between CRF01 and subtype B (Tovanabutra et al., 2003)]. The CRFs often have their own unique population dynamics and molecular epidemiology compared to their parental strains and often lead to novel infection dynamics and spread. One such CRF is CRF47_BF, discovered in Spain and described in 2010 (Fernández-García et al., 2010) as an intersubtype recombinant form between HIV-1 subtypes B and F. Of the CRFs, among the most abundant are those between B and F subtypes, with 19 CRF_BFs (note that in the Los Alamos HIV Database these are sometimes designated "BF" and sometimes "BF1," even for the same CRF). Of the CRF_BFs, all but two are known from South America (mainly Brazil, but Argentina, Uruguay, Paraguay, Chile, Peru, and Bolivia as well) with a few found both in South America and Europe (CRF66, 75, and 89). Only two CRF_BF have been reported to be found circulating exclusively in Europe, CRF42_BF in Luxembourg (Struck et al., 2015) and CRF47_BF in Spain (Fernández-García et al., 2010). Since its description, CRF47_BF has expanded considerably in Spain, predominantly *via* heterosexual contact, and is now known from Brazil as well, as attested by a CRF47_BF virus collected in this country whose sequence is deposited in the Los Alamos database (Los Alamos National Laboratory, 2021).

The goal of this study is to estimate the temporal and geographic origin of CRF47_BF and the dynamics of diffusion and growth throughout its evolutionary history. Toward this goal, we combine new CRF47_BF sequence data from our lab from strains isolated in Spain with data from other BF strains in the Los Alamos database to examine the origin and evolutionary dynamics of CRF47_BF and their epidemiological correlates.

## MATERIALS AND METHODS

### Sample and Data Collection

Plasma and whole blood samples were collected from HIV-1-infected patients at public hospitals across eight regions in Spain for a molecular epidemiological study of all new HIV-1 diagnoses seen at the participating centers and for antiretroviral drug resistance testing. Epidemiological data from the CRF47_BF patients were collected to link to the HIV sequence data. The epidemiological data included patient gender, the transmission route, the patient's year of HIV diagnosis and date of sample collection, the region from which the sample was collected, the country of origin of the individual, and whether the patient was on antiretroviral (ARV) therapy.

The study was approved by the Committee of Research Ethics of Instituto de Salud Carlos III, Majadahonda, Madrid, Spain (report numbers CEI PI 38_2016-v3 and CEI PI 31_2019-v5). The study did not require written informed consent by the study participants, as it used samples and data collected as part of routine clinical practice and patients' data were anonymized without retaining data allowing individual identification.

### Sequence Analyses

(RT-)PCR was used to amplify the protease-reverse transcriptase (PR-RT) gene region from plasma-extracted RNA or whole blood-extracted DNA using previously described primers (Delgado et al., 2015; **Supplementary Figure 1**). PCR products were sequenced using the Sanger method with an automated capillary sequencer. These data were combined with PR-RT sequences classified as CRF47_BF at the Los Alamos HIV Sequence Database and reference sequences for all subtypes and all CRFx_BFs for this same gene region from the Los Alamos HIV Database. Finally, we conducted a BLAST (Altschul et al., 1990) search against GenBank with the 5′-most 950 nt of PR-RT of all CRF47_BF viruses and included all sequences within 95% similarity. BLAST searches and further analyses (see below) yielded only two additional CRF47_BF sequences not identified as such at the Los Alamos database (with GenBank accessions JF929086, from Spain, and JQ238096, from Brazil).

We conducted two analyses with these data. (1) We included all data to validate the quality of the data and place the CRF47_BF within a broader phylogenetic context. Our initial phylogenetic analysis included subtypes from the HIV-1 M group (subtypes A1, A2, B, C, D, F1, F2, G, H, J, K, and L), as well as the CRF_BF recombinants (TotalCRF_BF.fasta, see **Supplementary Material**). Our final alignment (1,200 bp) included 14 sequences representing all the major subtypes within HIV-1 group M, 5 subtype B sequences, 11 subtype F (F1, F2) sequences, and 34 representatives of all known and distinct CRF_BFs. This alignment also included our more focused (2) CRF47_BF dataset (CRF47_BF.fasta see **Supplementary Material**). By including additional subtypes (including lab strains), we can both verify the monophyly of our target group of CRF47_BF sequences and validate that there are not contaminants or strange recombinants within this group as would be indicated by novel phylogenetic placement. For this second dataset, we included all 99 sequences from CRF47_BF, including 87 obtained by us [7 from a previous study (Fernández-García et al., 2010) and 80 newly derived] from the patients summarized in **Table 1**, and 12 from databases (10 from Spain and 2 from Brazil). We then conducted a focused analysis on the targeted CRF47_BF strains (1,377 aligned bp).

In both analyses, we aligned sequence data using MAFFT (Katoh and Standley, 2013) with the FFT-NS-2 progressive alignment approach since these sequences are relatively similar. Prior to subsequent phylogenetic and phylodynamic analyses, we checked that all sequences showed mosaic structures coincident with CRF47_BF, through two procedures: (1) bootscan analyses and (2) separate phylogenetic trees of B and F1 segments previously defined for CRF47_BF (Fernández-García et al., 2010), including B and F1 subtype references, to ensure that subtype assignment of each segment was identical for all sequences. Phylogenetic analyses were conducted using maximum-likelihood (Felsenstein, 1981; Posada and Crandall, 2021) as implemented by RAxML (Kozlov et al., 2019) *via* the CIPRES web service (Miller et al., 2012). The phylogenetic analyses utilized the best-fit model of evolution (Posada and Crandall, 1998) as determined by ModelTest-NG (Darriba et al., 2020). Phylogenetic analyses were also done using a Bayesian approach as implemented by MrBayes 3.2 (Ronquist et al., 2012) with integrated model selection, 10 million MCMC generations, and codon partitioning. Confidence in the resulting phylogenetic estimates was assessed using the bootstrap approach (Felsenstein, 1985) for the maximum-likelihood analyses with 1,000 pseudoreplicates and with posterior probabilities (pP) in the Bayesian framework. Phylogenetic trees were visualized with iTOL (Letunic and Bork, 2019), as well as mapping of epidemiological characters along the phylogeny.

**TABLE 1** | Summary data from patients with CRF47_BF variant from Spain.

| | Total *N* = 87 | Percent | New sequences |
|---|---|---|---|
| **Gender** | | | |
| Male | 68 | 78.2 | |
| Female | 18 | 20.7 | |
| Transgender | 1 | 1.2 | |
| **Region** | | | |
| Basque Country | 38 | 43.7 | 35 (38 total) |
| Navarre | 17 | 19.5 | 17 |
| Galicia | 14 | 16.1 | 10 (14 total) |
| Aragon | 9 | 10.3 | 9 |
| Comunitat Valenciana | 6 | 6.9 | 6 (13 total) |
| Castilla y Leon | 1 | 1.2 | 1 |
| Castilla-La Mancha | 1 | 1.2 | 1 |
| Madrid | 1 | 1.2 | 1 |
| **Transmission route** | | | |
| Heterosexual | 49 | 56.3 | |
| MSM | 19 | 21.8 | |
| Sexual transmission (unspecified sexuality) | 12 | 13.8 | |
| Other/no data | 7 | 8.1 | |
| **ARV therapy** | | | |
| No | 75 | 86.2 | |
| Yes | 7 | 8.1 | |
| No data | 5 | 5.7 | |
| **Country of origin** | | | |
| Spain | 68 | 78.2 | |
| Brazil | 5 | 5.8 | |
| Colombia | 5 | 5.8 | |
| Morocco | 3 | 3.5 | |
| Nicaragua | 2 | 2.3 | |
| Other | 4 | 4.6 | |

We applied BEAST2 (Bouckaert et al., 2014) to the CRF47_BF dataset to estimate a chronogram and the phylodynamic history of CRF47_BF. First, we validated the existence of temporal signal in the dataset with TempEst v1.5.3 (Rambaut et al., 2016), which determines the correlation of genetic divergence among sequences (measured as root-to-tip distance) with time. For the BEAST2 analysis we ran 10 million generations, two codon partitions (1st + 2nd, and 3rd positions), used an uncorrected log-normal relaxed molecular clock (initial ucld.mean = 1.0 and initial ucld.stdev = 0.333), estimated base frequencies and the HKY + G evolution model. The input file was created using BEAUti. Past population dynamics was estimated *via* Skygrid analysis (Hill and Baele, 2019) using a coalescent Bayesian Skygrid tree prior. We used Tracer (Rambaut et al., 2018) to verify convergence and to visualize the Skygrid plot. We compare the inferred effective population size of the CRF47_BF population in Spain to the proportion of CRF47_BF diagnoses over time across the same study regions and time period.

Finally, known drug resistant mutations were identified in the focused CRF47_BF data using the Stanford HIV Drug Resistance Database's HIVdb v9.0 program (Tang et al., 2012).

## Statistical Analyses

Correlations between cluster membership and epidemiological data were analyzed with Fisher's exact test.

# RESULTS

## Epidemiology and Sequences

We collected samples and epidemiological data from 87 patients throughout eight different regions of Spain (Basque Country, Navarre, Galicia, Aragon, Comunitat Valenciana, Madrid, Castilla-La Mancha, and Castilla y León) (**Figure 1**). Collections were made from 2007 to 2021. Males accounted for 78% of the individuals with CRF47_BF in our study and 56% of individuals reported transmission *via* heterosexual contact (61%

considering only individuals with available data on transmission route) (**Table 1**). The Spanish region with the highest proportion of the CRF47_BF variant in our data set was the Basque Country with 44% of the cases, while Navarre was the next highest (18% of the cases) (see **Table 1** for number of new CRF47_BF sequences and total CRF47_BF sequences per region, and **Figure 1** for the total number of analyzed HIV-1 sequences and prevalence of CRF47_BF among new HIV-1 diagnoses in each region in the sampling periods). Most samples were collected shortly after HIV diagnosis. Patients received ARV therapy after sample collection.

## Phylogenetics

The first phylogenetic analysis was a maximum likelihood phylogenetic estimate of the relationships amongst the CRFx_BFs, including HIV-1 M subtypes as outgroup taxa and subtypes B, F, and CRFx_BFs as ingroup taxa. Our RAxML tree depicted a monophyletic cluster of the subtype B, F, and CRF_BFs relative to the other HIV-1 subtypes (**Supplementary Figure 2**), but including also Subtype D. The backbone structure of the CRF phylogenetic relationships was weakly supported (<70% bootstrap support – indicated by dashed lines), which is not particularly surprising given the potential difficulty in representing evolutionary histories of recombinant HIV-1 forms as bifurcating trees (Posada and Crandall, 2001, 2002). Many of the CRFx_BF forms cluster in strongly supported monophyletic groups themselves (e.g., CRF40_BF, CRF72_BF, CRF75_BF, CRF90_BF, CRF89_BF, etc.), including our target group of CRF47_BF sequences. Many of the other CRFs form weakly supported monophyletic groups (e.g., CRF70_BF, CRF46_BF, CRF38_BF, etc.) and a few form non-monophyletic groupings (e.g., CRF66_BF and CRF71_BF). The subtype B sequences cluster together within the CRFx_BF clade with both a cluster of subtype D and the CRF28_BF sequence nested within this subtype B cluster. Nevertheless, the target group for this study, the CRF47_BF sequences, clearly form a monophyletic group, suggesting independent evolution, and are a sister group to the CRF44_BF clade.

The Bayesian estimated phylogeny for the CRF47_BF sequences shows a monophyletic grouping of the sequences from Spain (**Figure 2**) with the two sequences from Brazil (KJ849798 and JQ238096) branching basally. Within the Spanish cluster, there are three strongly supported clusters, comprising 29 (cluster I), 17 (cluster II), and 13 (cluster III) viruses, respectively, which are associated with the Basque Country ($p$ = 0.0002), Navarre ($p$ = 0.0001), and Aragon ($p$ = 0.0002), respectively. This is indicative of a single introduction of CRF47_BF into Spain with subsequent spread throughout the country and point introductions with subsequent expansion in different regions.

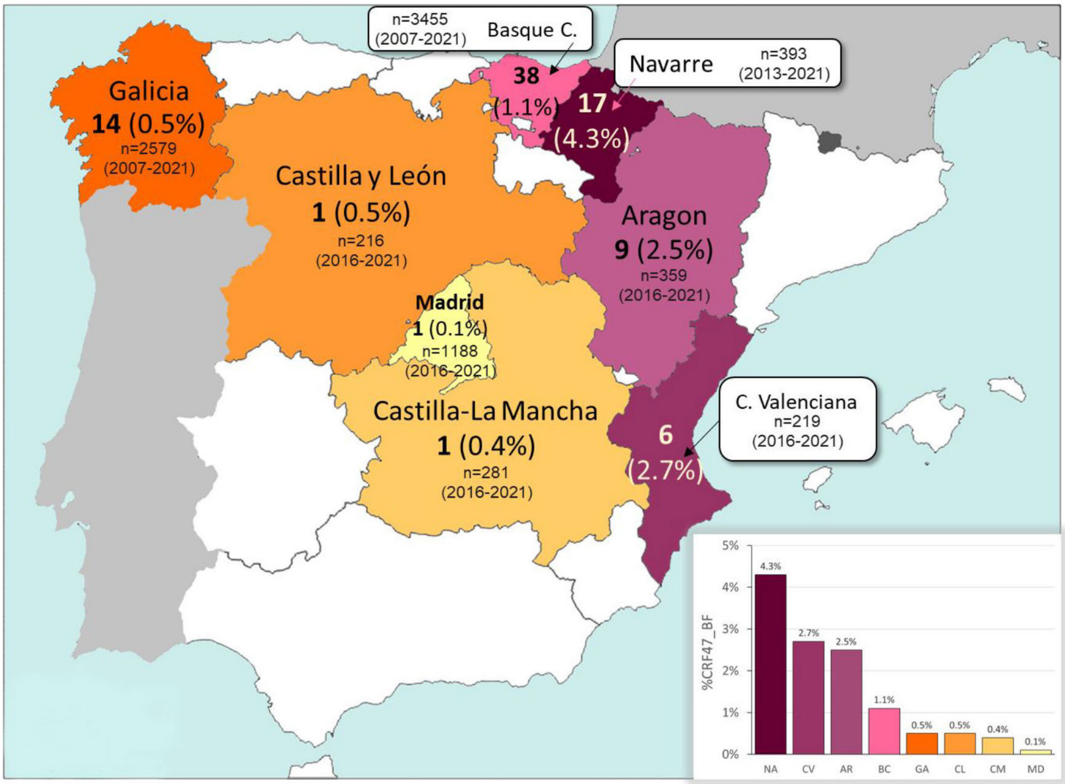


**FIGURE 1 |** Map of Spain with the number and estimated prevalence (percentage between parentheses) of HIV-1 CRF47_BF viruses detected in each of the eight regions analyzed. The number of new HIV-1 diagnoses (*n*) and the period of study in each region are also indicated. The percentage of CRF47_BF in each region is also represented in the inserted bar chart. NA, Navarre; CV, Comunitat Valenciana; AR, Aragon; BC, Basque Country; GA, Galicia; CL, Castilla y León; CM, Castilla-La Mancha; MD, Madrid.

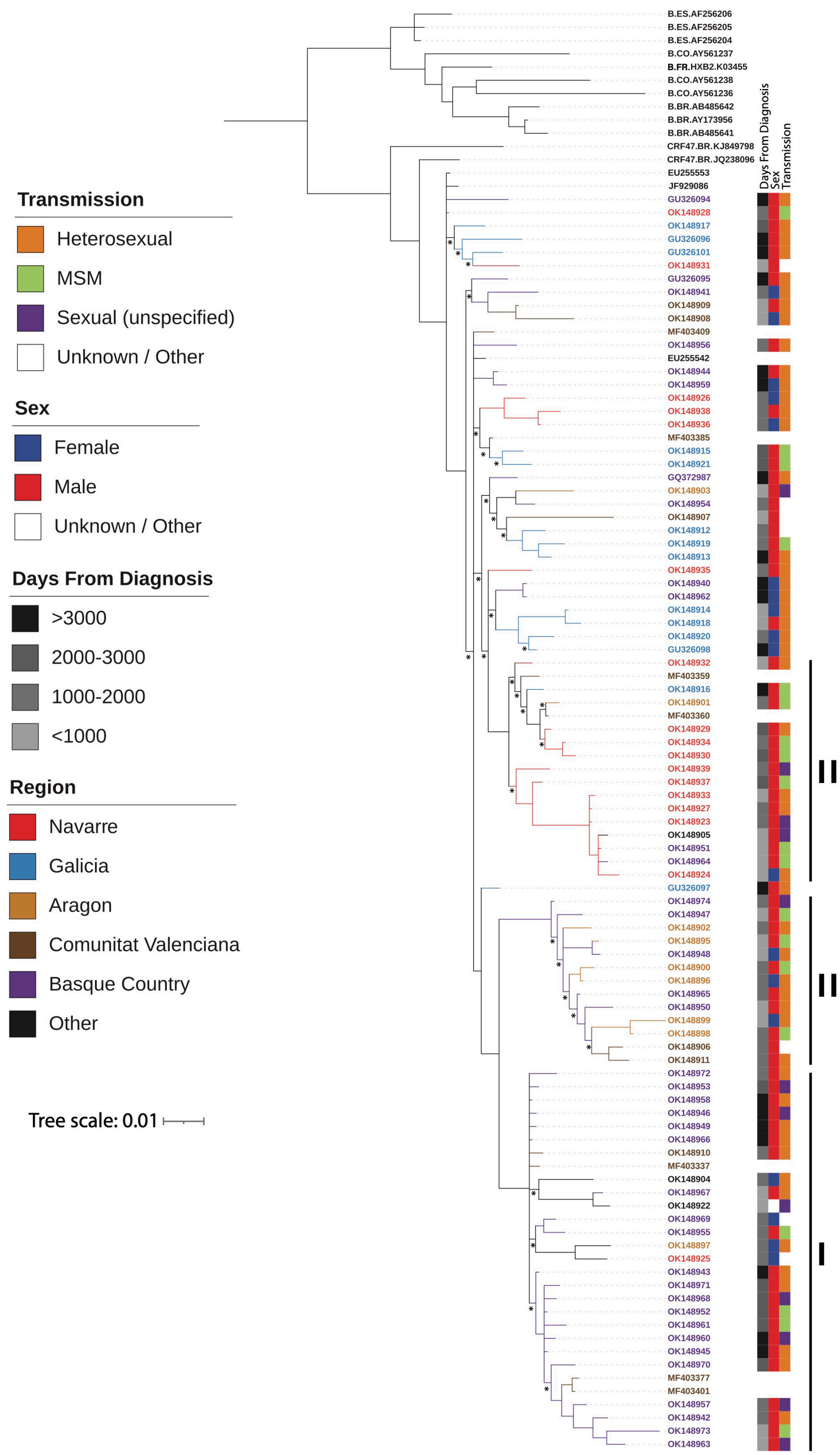


**FIGURE 2 |** Majority-rule consensus Bayesian (MrBayes) phylogenetic estimate of CRF47_BF sequences from Spain (colored by region) and other CRF47_BF sequences from GenBank, as well as a few additional subtype B sequences from Spain, Brazil, and Colombia plus HXB2, all to serve as an outgroup to the CRF47 sequences. Only clade posterior probabilities <0.95 are indicated by an *; all other clades showed posterior probabilities ≥0.95. Epidemiological data are mapped to the right of the phylogeny, including days from diagnosis, sex, transmission, and geographic region. Branch lengths are drawn proportional to the amount of sequence divergence. Clusters corresponding to the Spanish regions of Basque Country (cluster I), Navarre (cluster II), and Aragon (cluster III) are indicated.

The mixing of patient gender throughout the resulting phylogeny supports the epidemiological data suggesting predominantly heterosexual transmission among patients. We also found that cluster II, associated with Navarre, was associated with men who have sex with men (MSM) ($p$ = 0.0388). In this cluster, 14 of 15 individuals with known gender are men.

The bootscan analysis for recombination and separate trees of B and F segments suggest that all the target sequences within the CRF47_BF analyses presented here share the same recombination pattern (as outlined at the Los Alamos HIV Database) (**Figure 3** and **Supplementary Figure 3**). Thus, while recombination can significantly impact phylogenetic interpretations (certainly, for the overall tree presented in **Supplementary Figure 2**), it does not seem to be differentially impacting analyses of our targeted CRF47_BF sequences.

Based on the sample dates, we grouped these in four temporal categories of recency (days between diagnosis date and current date) (>3,000, 2,000–3,000, 1,000–2,000, and <1,000 days from current) for ease of visualization of time over the phylogeny to test for temporal clustering. Thus, the greater the value the closer to the most recent common ancestor, i.e., origin of CRF47_BF. Note that these correspond well to the branch lengths observed leading to samples with <1,000 days from diagnosis having longer branches from the root to the terminal samples and >3,000 having shorter and more basal branches in the phylogram. No temporal clustering was observed as these different time categories were distributed throughout the CRF47_BF phylogeny (**Figure 2**).

## Analysis of Drug Resistance Mutations

To identify drug resistance mutations in the CRF47_BF viruses, we analyzed the sequences with the Stanford HIV Drug Resistance Database's HIVdb program (Tang et al., 2012). We found ARV drug resistance mutations in five patients: M184V or M184I mutations of resistance to nucleoside reverse transcriptase inhibitors (NRTI) in three samples; K103N mutation of resistance to non-nucleoside reverse transcriptase inhibitors (NNRTI) + K65N mutation of resistance to NRTIs in one sample;

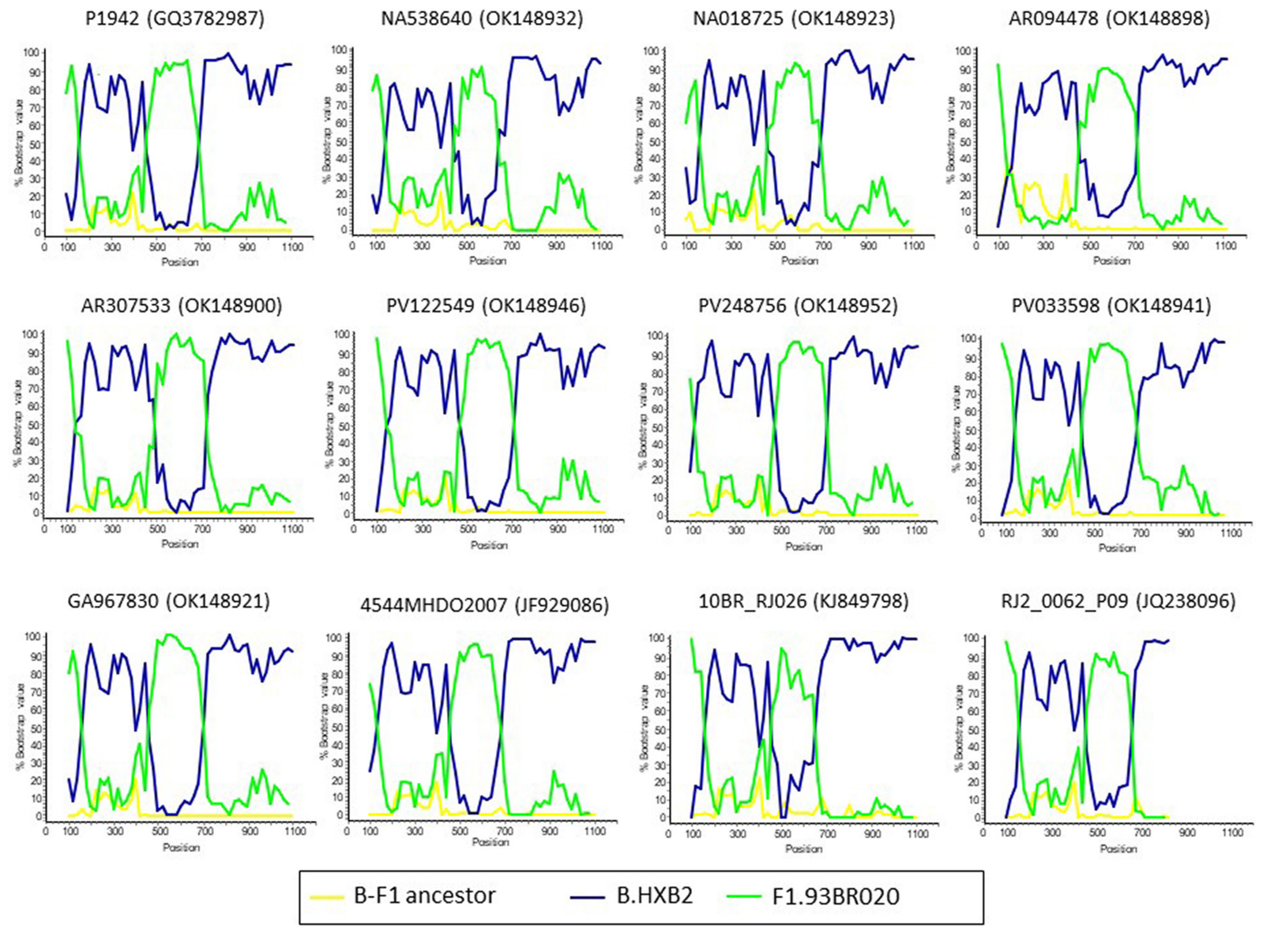


**FIGURE 3 |** Bootscan analyses of PR-RT sequences of CRF47_BF viruses. Simplot v3.5 (Lole et al., 1999) was used for the analyses. Twelve representative profiles are displayed. Names of viruses, with GenBank accessions, are shown above each bootscan plot. P1942 was included as CRF47_BF reference. A reconstructed B-F1 ancestral sequence was used as outgroup. The horizontal axis represents the position from nucleotide 1 of protease and the vertical axis represents bootstrap values supporting clustering with references.

and E138A mutation associated with low level resistance to the NNRTI rilpivirine in one patient. Only one of these patients, with M184I mutation, was ARV drug-experienced.

## Phylodynamics

Our TempEst analysis determined that there was an adequate temporal signal in the dataset ($R^2 = 0.5051$). With time-stamped sequence data, we performed a Bayesian Skygrid coalescent analysis to estimate historical population dynamics (Hill and Baele, 2019) of the CRF47_BF variants throughout Spain. Time labels (tipdates) were determined by the date of sample collection (ranging from 2007 to 2021). Our analysis supports a fairly dynamic population history of the CRF47_BF in Spain over the last 15 years with an initial increase in population size, a subsequent increase from 2011 to 2015, with a leveling off more recently, but seemingly increasing variance (**Figure 4**). This fluctuation in effective population size of CRF47_BF in Spain is not as dynamic as the % CRF47_BF infections among new HIV diagnoses, that fluctuate considerably over this same time period (**Figure 4**), but with similar overall trends. The average effective population size was estimated to be 155 with a mean substitution rate of $1.8128 \times 10^{-3}$ [95% highest posterior density (HPD) interval ($1.3956 \times 10^{-3}$, $2.2548 \times 10^{-3}$)]. Using BEAST, we estimated a chronogram to determine the time of origin for both the CRF47_BF clade as well as the timing of the introduction of CRF47_BF viruses to Spain (**Figure 5**). We estimated the origin of the CRF47_BF clade in Brazil (pP = 1.0), dated to 1999–2000 (95% HPD interval between 1994 and 2003) and timed the introduction of CRF47_BF to Spain (pP = 0.99) to be 2002–2003 (95% HPD interval between 2000 and 2004) (**Figure 5**). Similarly, viral strains seem to have entered once and spread through the Spanish regions of Basque Country (cluster I) (pP = 1.0), Navarre (cluster II) (pP = 1.0), and Aragon (cluster III) (pP = 1.0) between 2009 and 2012 (**Figure 5**). These analyses, hence, suggest that CRF47_BF was probably circulating in Spain for about 8 years before it was identified through DNA sequencing, but clearly at a relatively low frequency. Given the sampling of CRF47 sequences, it appears that the introduction of this recombinant form to Spain was *via* Brazil, supported by very high posterior probabilities (pP = 1.00).

# DISCUSSION

The HIV-1 CRF47_BF was first reported in 2010, detected in nine samples collected in Spain in 2007–2009. Samples have subsequently been collected as this novel variant has spread throughout the country. Our phylogenetic analysis shows that isolates of CRF47_BF form a strongly supported monophyletic group (share a most recent common ancestor) relative to other CRFx_BF sequences, distinct from other CRFx_BF sequences, subtype B, sub-subtype F2 and other Group M subtypes. A focused phylogenetic analysis of the CRF47_BF sequences show a clear single origin in Brazil around 1999–2000 with a subsequent transmission and rapid spread throughout Spain beginning around 2002–2003. Three strongly supported clusters, comprising a majority of viruses and associated with the regions of Basque Country, Navarre, and Aragon, were identified; this suggests that after a single introduction in Spain, CRF47_BF has spread mainly through localized point introductions and subsequent spread in different geographical areas. CRF47_BF is predominant in males (78%) with a predominantly heterosexual transmission (56% of the total, 61%

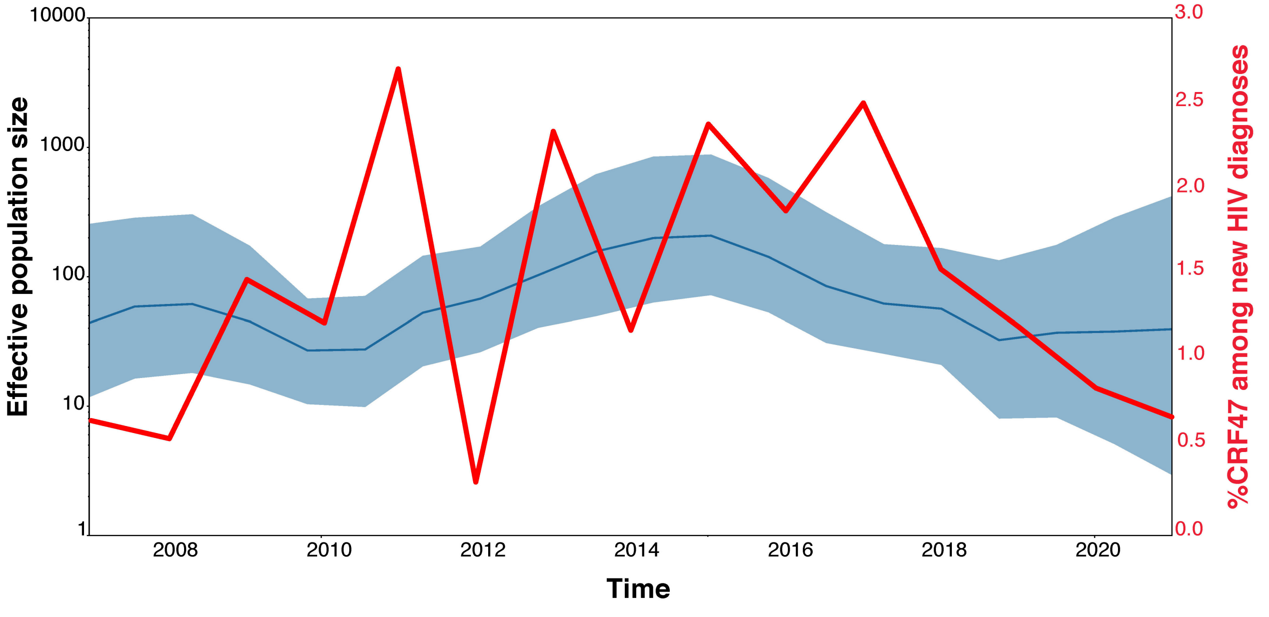


**FIGURE 4 |** Population dynamics of CRF47_BF in Spain. Bayesian Skygrid estimate of fluctuating population size by year compared with the actual proportion of new CRF47_BF samples collected in the study each year among new HIV-1 diagnoses.

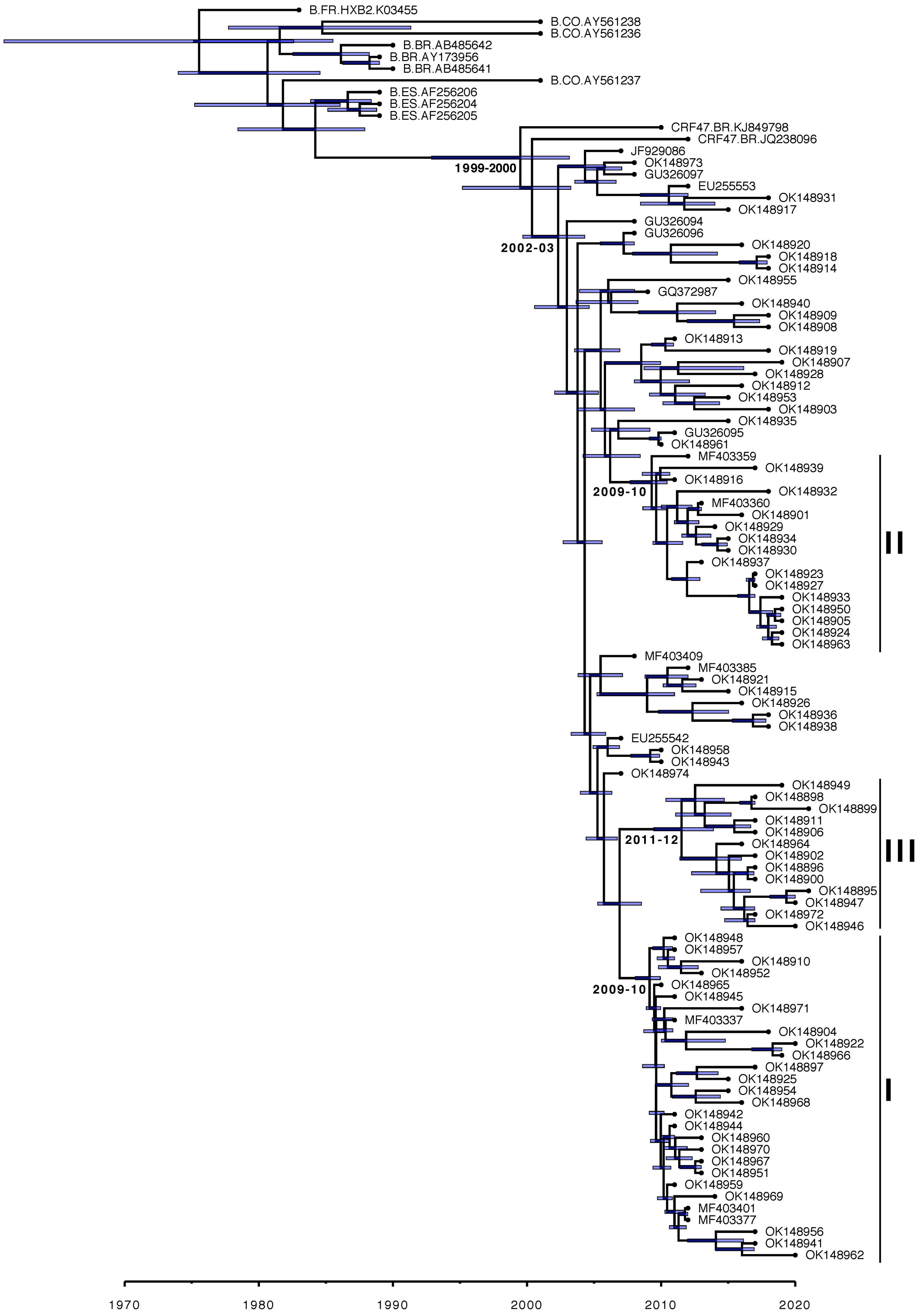


**FIGURE 5 |** Maximum clade credibility Bayesian (BEAST) chronogram estimate of CRF47_BF sequences from Spain and other CRF47_BF sequences from GenBank, as well as a few additional subtype B sequences from Spain, Brazil, and Colombia plus HXB2, all to serve as an outgroup to the CRF47_BF sequences. Clusters associated with the Spanish regions of Basque Country (cluster I), Navarre (cluster II), and Aragon (cluster III) are indicated. Estimated years of emergence of CRF47_BF in Brazil, of its introduction in Spain, and of emergence of the Spanish clusters are indicated besides the corresponding nodes. A total of 95% highest posterior density (HPD) intervals (blue bars) are shown for all time estimates.

of those with data on transmission mode). The phylodynamic analysis and percent of CRF47 among new HIV diagnoses both support a fluctuating population size of CRF47_BF over the last 15 years with periods of expansion and contraction, suggesting that continued monitoring of this novel variant will be important to track its spread.

It is interesting to note that one cluster of 17 individuals, associated with Navarre, where 14 of 15 individuals with available data were male, was significantly associated ($p = 0.0388$) with transmission among MSM. Although three men were reported to be heterosexual, considering the great male preponderance in the cluster, it is probable that they are non-disclosed MSM (Hué et al., 2014; Ragonnet-Cronin et al., 2018). The identification of an MSM-associated cluster within the CRF47_BF clade may be indicative of the diffusion of CRF47_BF from a heterosexual-driven network to a MSM-driven network. A similar phenomenon has been observed for the two other CRFs of South American origin identified by us in Spain: CRF66_BF (Bacqué et al., 2021) and CRF89_BF (Delgado et al., 2021). Such phenomenon may reflect the migration of these CRFs from countries where heterosexual transmission is predominant to Spain, where most currently expanding HIV-1 clusters are associated with MSM (Patiño-Galindo et al., 2017; Gil et al., 2022). It should be pointed out, however, that outside of the Navarre cluster, the male:female ratio was 3.2:1, which contrasts to the 2.4:1 ratio of self-declared heterosexual men to MSM (decreasing to 1.5:1 if all men with non-specified sexual transmission were MSM). This discrepancy could also be explained by the presence of non-disclosed MSM among self-declared heterosexual men outside of the Navarre cluster, and indicates that epidemiological data on transmission route based on self-reported sexual behaviors should be interpreted with caution.

The recent expansion in Spain of CRF47_BF, whose Brazilian origin is first reported here, is one more example of the increasing relationship of the South American and European HIV-1 epidemics, also reflected in the propagation in Europe of other CRFs (12_BF, 17_BF, 60_BC, 66_BF, and 89_BF) (Simonetti et al., 2014; Fabeni et al., 2015, 2020; Bacqué et al., 2021; Delgado et al., 2021) and variants of subtypes F1 and C (Tovanabutra et al., 2003; de Oliveira et al., 2010; Thomson et al., 2012; Lai et al., 2014; Carvalho et al., 2015; Delgado et al., 2015; Vinken et al., 2019) of South American ancestry, which probably derives from increasing migratory flows from South America to Europe.

The repeated introduction and expansion in Spain of multiple CRFs and non-B subtypes (Delgado et al., 2015, 2019; Patiño-Galindo et al., 2017; González-Domenech et al., 2018; Kostaki et al., 2019) justifies the establishment of a HIV-1 molecular epidemiological surveillance system, aimed at promptly detecting the propagation of such variants, as well as rapidly expanding clusters, that could provide information in real-time on changes in the genetic composition and the dynamics of the HIV-1 epidemic to guide the implementation of preventive public health interventions (Paraskevis et al., 2016; German et al., 2017; Oster et al., 2018). Nevertheless, phylogenetic analyses of HIV sequence data should be taken with caution as recombination can impact phylogenetic inference (Schierup and Hein, 2000; Posada and Crandall, 2002) suggesting network approaches might be better suited for representation of such data (Clement et al., 2000). Nevertheless, we focus on a specific set of CRF47_BF sequences with a shared mosaic structure and therefore our results should be robust to the impacts of recombination (see **Figure 3**).

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, OK148895-OK148974.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Committee of Research Ethics of Instituto de Salud Carlos III, Majadahonda, Madrid, Spain (report numbers CEI PI 38_2016-v3 and CEI PI 31_2019-v5). Written informed consent for participation was not required for this study, as it used samples and data collected as part of routine clinical practice and patients' data were anonymized without retaining data allowing individual identification.

## MEMBERS OF THE SPANISH GROUP FOR THE STUDY OF NEW HIV DIAGNOSES

Hospital Universitario de Basurto: José Luis Díaz de Tuesta del Arco, Silvia Hernáez, Sofía Ibarra-Ugarte, Josefa Muñoz, María Carmen Nieto-Toboso, and Miren Zuriñe Zubero-Sulibarria. Hospital Universitario de Cruces, Bilbao: Elena Bereciartua-Bastarrica, Luis Elorduy, Ane Josune Goikoetxea-Agirre, and Leyre López-Soria. Hospital Universitario de Galdakao: María José López de Goikoetxea. Hospital Universitario Donostia, San Sebastián: Maitane Aranzamendi, Gustavo Cilla, José Antonio Iribarren, and Yolanda Salicio. Hospital Universitario Araba, Vitoria: Carmen Gómez, and José Joaquín Portu. Hospital Universitario de Navarra, Pamplona: Aitziber Aguinaga, Carmen Ezpeleta, Carmen Martín-Salas, and María Gracia Ruiz-Alda. Hospital Reina Sofía, Tudela: José Javier García-Irure. Hospital Universitario Sant Joan d'Alacant: Fernando Buñuel and Francisco Jover-Díaz. Complejo Hospitalario Universitario de Vigo: Jorge Julio Cabrera, Antonio Ocampo, and Celia Miralles. Complejo Hospitalario Universitario de Pontevedra: Julio Diz-Aren and Matilde Trigo. Complejo Hospitalario Lucus Augusti, Lugo: María José Gude, Ramón Rabuñal, and Eva María Romay. Complejo Hospitalario Universitario de Ourense: Ricardo Fernández-Rodríguez and Juan García-Costa. Hospital Universitario Miguel Servet, Zaragoza: Piedad Arazo and Ana María Martínez-Sapiña. Centro Sanitario Sandoval, Madrid: Jorge del Romero. Hospital Universitario Río Hortega, Valladolid: Belén Lorenzo-Vidal. Hospital Universitario de Toledo: César Gómez.

## AUTHOR CONTRIBUTIONS

MT, ED, and MP-L conceived of the project. ED collected sequence data from the samples. GH, KC, MP-L, MT, and ED conducted the data analyses. HG performed the data curation. SB, VM, MS, JC-G, and EG-B performed the experimental work. The members of the Spanish Group for the Study of New HIV Diagnoses collected the samples and clinical and epidemiological data for the study. KC, GH, and MP-L wrote the original draft of the manuscript. MT, ED, and HG edited the manuscript. All authors read and approved the manuscript.

## FUNDING

The study was supported by Acción Estratégica en Salud Intramural (AESI) program of Instituto de Salud Carlos III, projects "Estudio sobre Vigilancia Epidemiológica Molecular de la Infección por VIH-1 en España," PI16CIII/00033, and "Epidemiología Molecular del VIH-1 en España y su Utilidad para Investigaciones Biológicas y en Vacunas," PI19CIII/00042; Red de Investigación en SIDA (RIS), Instituto de Salud Carlos III, Plan Nacional I+D+I, project RD16ISCIII/0002/0004; and scientific agreements with the Governments of Galicia (MVI 1004/16) and Basque Country (MVI 1001/16).

## ACKNOWLEDGMENTS

We thank José Antonio Taboada, from Consellería de Sanidade, Government of Galicia, and Daniel Zulaika, from Unidad de Coordinación del Plan de Prevención y Control del SIDA, Osakidetza-Servicio Vasco de Salud, Government of the Basque Country, for their support of this study, and the personnel at the Genomic Unit, Instituto de Salud Carlos III, for technical assistance in sequencing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.863123/full#supplementary-material

## REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Bacqué, J., Delgado, E., Benito, S., Moreno-Lorenzo, M., Montero, V., Gil, H., et al. (2021). Identification of CRF66_BF, a new HIV-1 circulating recombinant form of South American origin. *Front. Microbiol.* 12:774386. doi: 10.3389/fmicb.2021.774386

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10:e1003537. doi: 10.1371/journal.pcbi.1003537

Carvalho, A., Costa, P., Triunfante, V., Branca, F., Rodrigues, F., Santos, C. L., et al. (2015). Analysis of a local HIV-1 epidemic in portugal highlights established transmission of non-B and non-G subtypes. *J. Clin. Microbiol.* 53, 1506–1514.

Castro-Nallar, E., Crandall, K. A., and Pérez-Losada, M. (2012a). Genetic diversity and molecular epidemiology of HIV transmission. *Future Virol.* 7, 239–252.

Castro-Nallar, E., Pérez-Losada, M., Burton, G. F., and Crandall, K. A. (2012b). The evolution of HIV: inferences using phylogenetics. *Mol. Phylogenet. Evol.* 62, 777–792.

Clement, M., Posada, D., and Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* 9, 1657–1659.

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294.

de Oliveira, T., Pillay, D., Gifford, R. J., and UK Collaborative Group on HIV Drug Resistance (2010). The HIV-1 subtype C epidemic in south America is linked to the United Kingdom. *PLoS One* 5:e9311. doi: 10.1371/journal.pone.0009311

Delgado, E., Benito, S., Montero, V., Cuevas, M. T., Fernández-García, A., Sánchez-Martínez, M., et al. (2019). Diverse large HIV-1 non-subtype B clusters are spreading among men who have sex with men in Spain. *Front. Microbiol.* 10:655. doi: 10.3389/fmicb.2019.00655

Delgado, E., Cuevas, M. T., Domínguez, F., Vega, Y., Cabello, M., Fernández-García, A., et al. (2015). Phylogeny and phylogeography of a recent HIV-1 subtype F outbreak among men who have sex with men in Spain deriving from a cluster with a wide geographic circulation in Western Europe. *PLoS One* 10:e0143325. doi: 10.1371/journal.pone.0143325

Delgado, E., Fernández-García, A., Pérez-Losada, M., Moreno-Lorenzo, M., Fernández-Miranda, I., Benito, S., et al. (2021). Identification of CRF89_BF, a new member of an HIV-1 circulating BF intersubtype recombinant form family widely spread in South America. *Sci. Rep.* 11:11442.

Fabeni, L., Alteri, C., Orchi, N., Gori, C., Bertoli, A., Forbici, F., et al. (2015). Recent transmission clustering of HIV-1 C and CRF17_BF strains characterized by NNRTI-related mutations among newly diagnosed men in central Italy. *PLoS One* 10:e0135325. doi: 10.1371/journal.pone.0135325

Fabeni, L., Santoro, M. M., Lorenzini, P., Rusconi, S., Gianotti, N., Costantini, A., et al. (2020). Evaluation of HIV transmission clusters among natives and foreigners living in Italy. *Viruses* 12:791.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Fernández-García, A., Pérez-Alvarez, L., Cuevas, M. T., Delgado, E., Muñoz-Nieto, M., Cilla, G., et al. (2010). Identification of a new HIV type 1 circulating BF intersubtype recombinant form (CRF47_BF) in Spain. *AIDS Res. Hum. Retroviruses* 26, 827–832.

German, D., Grabowski, M. K., and Beyrer, C. (2017). Enhanced use of phylogenetic data to inform public health approaches to HIV among men who have sex with men. *Sex. Health* 14, 89–96. doi: 10.1071/SH16056

Gil, H., Delgado, E., Benito, S., Georgalis, L., Montero, V" Sánchez, M. et al., (2022). Transmission clusters, predominantly associated with men who have sex with men, play a main role in the propagation of HIV-1 in northern Spain (2013–2018). *Front. Microbiol.* 13:782609. doi: 10.3389/fmicb.2022.782609

González-Domenech, C. M., Viciana, I., Delaye, L., Mayorga, M. L., Palacios, R., de la Torre, J., et al. (2018). Emergence as an outbreak of the HIV-1 CRF19_cpx variant in treatment-naïve patients in southern Spain. *PLoS One* 13:e0190544. doi: 10.1371/journal.pone.0190544

Hill, V., and Baele, G. (2019). Bayesian estimation of past population dynamics in BEAST 1.10 using the Skygrid coalescent model. *Mol. Biol. Evol.* 36, 2620–2628. doi: 10.1093/molbev/msz172

Hué, S., Brown, A. E., Ragonnet-Cronin, M., Lycett, S. J., Dunn, D. T., Fearnhill, E., et al. (2014). Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS* 28, 1967–1975. doi: 10.1097/QAD.0000000000000383

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.

Kostaki, E. G., Flampouris, A., Karamitros, T., Chueca, N., Alvarez, M., Casas, P., et al. (2019). Spatiotemporal characteristics of the largest HIV-1 CRF02_AG outbreak in Spain: evidence for onward transmissions. *Front. Microbiol.* 10:370. doi: 10.3389/fmicb.2019.00370

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455.

Lai, A., Bozzi, G., Franzetti, M., Binda, F., Simonetti, F. R., Micheli, V., et al. (2014). Phylogenetic analysis provides evidence of interactions between Italian heterosexual and South American homosexual males as the main source of national HIV-1 subtype C epidemics. *J. Med. Virol.* 86, 729–736.

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259.

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/JVI.73.1.152-160.1999

Los Alamos National Laboratory [LANL] *HIV Databases*. Available online at: https://www.hiv.lanl.gov/content/index (accessed March 20, 2021).

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2012). "The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources," in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond XSEDE '12*, (New York, NY: Association for Computing Machinery), 1–8.

Oster, A. M., France, A. M., and Mermin, J. (2018). Molecular epidemiology and the transformation of HIV prevention. *JAMA* 319, 1657–1658. doi: 10.1001/jama.2018.1513

Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I., and Hatzakis, A. (2016). The application of HIV molecular epidemiology to public health. *Infect. Genet. Evol.* 46, 159–168. doi: 10.1016/j.meegid.2016.06.021

Patiño-Galindo, J. Á, Torres-Puente, M., Bracho, M. A., Alastrué, I., Juan, A., Navarro, D., et al. (2017). The molecular epidemiology of HIV-1 in the Comunidad Valenciana (Spain): analysis of transmission clusters. *Sci. Rep.* 7:11584. doi: 10.1038/s41598-017-10286-1

Pennings, P. S., Kryazhimskiy, S., and Wakeley, J. (2014). Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genet.* 10:e1004000. doi: 10.1371/journal.pgen.1004000

Posada, D., and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.

Posada, D., and Crandall, K. A. (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45.

Posada, D., and Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* 54, 396–402.

Posada, D., and Crandall, K. A. (2021). Felsenstein phylogenetic likelihood. *J. Mol. Evol.* 89, 134–145.

Ragonnet-Cronin, M., Hué, S., Hodcroft, E. B., Tostevin, A., Dunn, D., Fawcet, T., et al. (2018). Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *Lancet HIV* 5, e309–e316. doi: 10.1016/S2352-3018(18)30062-6

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67, 901–904.

Rambaut, A., Lam, T. T., Max, C. L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2:vew007.

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nat. Rev. Genet.* 5, 52–61.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542.

Schierup, M. H., and Hein, J. (2000). Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156, 879–891.

Shriner, D., Rodrigo, A. G., Nickle, D. C., and Mullins, J. I. (2004). Pervasive genomic recombination of HIV-1 in vivo. *Genetics* 167, 1573–1583.

Simonetti, F. R., Lai, A., Monno, L., Binda, F., Brindicci, G., Punzi, G., et al. (2014). Identification of a new HIV-1 BC circulating recombinant form (CRF60_BC) in Italian young men having sex with men. *Infect. Genet. Evol.* 23, 176–181.

Struck, D., Roman, F., De Landtsheer, S., Servais, J.-Y., Lambert, C., Masquelier, C., et al. (2015). Near full-length characterization and population dynamics of the human immunodeficiency virus type I circulating recombinant form 42 (CRF42_BF) in Luxembourg. *AIDS Res. Hum. Retroviruses* 31, 554–558.

Tang, M. W., Liu, T. F., and Shafer, R. W. (2012). The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology* 55, 98–101.

Thomson, M. M., Fernández-García, A., Delgado, E., Vega, Y., Díez-Fuertes, F., Sánchez-Martínez, M., et al. (2012). Rapid expansion of a HIV-1 subtype F cluster of recent origin among men who have sex with men in Galicia, Spain. *J. Acquir. Immune Defic. Syndr.* 59, e49–51.

Tovanabutra, S., Watanaveeradej, V., Viputtikul, K., De Souza, M., Razak, M. H., Suriyanon, V., et al. (2003). A new circulating recombinant form, CRF15_01B, reinforces the linkage between IDU and heterosexual epidemics in Thailand. *AIDS Res. Hum. Retroviruses* 19, 561–567.

Vinken, L., Fransen, K., Cuypers, L., Alexiev, I., Balotta, C., Debaisieux, L., et al. (2019). Earlier initiation of antiretroviral treatment coincides with an initial control of the HIV-1 sub-subtype F1 outbreak among men-having-sex-with-men in Flanders, Belgium. *Front. Microbiol.* 10:613. doi: 10.3389/fmicb.2019.00613

Vuilleumier, S., and Bonhoeffer, S. (2015). Contribution of recombination to the evolutionary history of HIV. *Curr. Opin. HIV AIDS* 10, 84–89.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/fmicb.2022.855232

# HIV Transmembrane Glycoprotein Conserved Domains and Genetic Markers Across HIV-1 and HIV-2 Variants

*Ana Valadés-Alcaraz, Roberto Reinosa and África Holguín**

*HIV-1 Molecular Epidemiology Laboratory, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Microbiology Department, Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP), Red en Investigación Translacional en Infecciones Pediátricas (RITIP), Madrid, Spain*

HIV envelope transmembrane glycoproteins gp41 (HIV-1) and gp36 (HIV-2) present high variability and play a key role in the HIV-host cell membrane's fusion, as a target for human broadly neutralizing antibodies (bnAbs) and drugs. Thus, a better knowledge of amino acid (aa) conservation across structural domains and HIV variants can help to identify conserved targets to direct new therapeutic and diagnostic strategies. All available gp41/gp36 nucleotide sequences were downloaded from Los Alamos National Laboratory (LANL) HIV Sequence Database, selecting 17,078 sequences ascribed to HIV-1 and HIV-2 variants with ≥3 sequences. After aligning and translating into aa with MEGAv6.0, an in-house bioinformatics program (EpiMolBio) was used to identify the most conserved aa and the aa changes that were specific for each variant (V-markers) *vs.* HXB2/BEN (HIV-1/HIV-2) reference sequence. We analyzed the presence of specific aa changes among V-markers affecting infectivity, gp41 structure, function, or resistance to the enfuvirtide viral fusion inhibitor (T-20). We also inferred the consensus sequences per HIV variant, describing in each HIV-1 group (M, N, O, P) the conservation level along the complete gp41 per structural domain and locating in each binding site the anti-gp41 human Abs (bnAbs and non bnAbs) described in LANL. We found 38.3/59.7% highly conserved aa present in ≥90% of the 16,803/275 gp41/gp36 sequences ascribed to 105/3 HIV-1/HIV-2 variants, with 9/12.6% of them showing complete conservation across LANL sequences. The fusion peptide, its proximal region, the N-heptad repeat, and the membrane-proximal external region were the gp41 domains with ≥84% of conserved aa in the HIV-1 consensus sequence, the target of most Abs. No natural major resistance mutations to T-20 were observed. Our results show, for the first time, a complete conservation study of gp41/gp36 per variant in the largest panel of HIV variants analyzed to date, providing useful information for a more rational design of drugs, vaccines, and molecular detection tests targeting the HIV transmembrane glycoprotein.

**Keywords: HIV, transmembrane glycoprotein gp41, transmembrane glycoprotein gp36, conservation, variants, antibody binding domains**

OPEN ACCESS

**Edited by:**
*Joris Hemelaar,*
*University of Oxford, United Kingdom*

**Reviewed by:**
*Kelly Lee,*
*University of Washington,*
*United States*
*Suvankar Ghorai,*
*Raiganj University, India*

***Correspondence:**
*África Holguín*
*africa.holguin@salud.madrid.org*

**Specialty section:**
*This article was submitted to Virology,*
*a section of the journal*
*Frontiers in Microbiology*

**Received:** *14 January 2022*
**Accepted:** *29 March 2022*
**Published:** *27 May 2022*

**Citation:**
*Valadés-Alcaraz A, Reinosa R and Holguín Á (2022) HIV Transmembrane Glycoprotein Conserved Domains and Genetic Markers Across HIV-1 and HIV-2 Variants.*
*Front. Microbiol. 13:855232.*
*doi: 10.3389/fmicb.2022.855232*

## INTRODUCTION

The human immunodeficiency virus (HIV) envelope transmembrane glycoproteins gp41 and gp36 are located on the HIV-1 and HIV-2 virion's membrane forming trimers with gp120 and gp105 glycoproteins, respectively. These proteins mediate the viral fusion with the host cell's membrane, allowing the entry of genetic material and viral proteins into the cells (Blumenthal et al., 2012). Therefore, they are important targets for the development of fusion inhibitors, such as antiretrovirals (ARV) (Qadir and Malik, 2010), antibodies (Abs) (Caillat et al., 2020), and aptamers (Li et al., 2016) used as HIV treatment. HIV-1 gp41, with 345 amino acids (aa), can be segmented into three domains (**Figure 1A**): one exposed ectodomain (aa 1-172), a transmembrane region (TM, aa 173-194), and an intraviral not exposed C-terminal domain (CT, aa 195-345). The ectodomain is exposed and can be divided further into distinct functional regions important for fusion and virus infectivity: an N-terminal hydrophobic region termed as fusion peptide (FP, aa 1-16), necessary to bind the virus to the cell membrane, followed by an N-terminal alpha-helical region or N-heptad repeat region (NHR, aa 33-70). These domains were linked by a fusion peptide proximal region (FPPR, aa 17-32) rich in polar aa and critical for HIV-1 fusion and infectivity because it stabilizes the envelope trimers (Lu et al., 2019). A loop immune-dominant linker with a disulfide bridge (IL, aa 71-113) links the NHR to a C-heptad repeat region (CHR, aa 114-153) (**Figure 1B**). A membrane-proximal external region (MPER, aa 154-172), conformationally flexible, connects the CHR to the TM region (Louis et al., 2016). The gp41 NHR domain is the target of ARV, such as enfuvirtide (T-20), the only clinically approved viral fusion inhibitor for the treatment of HIV infection (Lazzarin, 2005; Oldfield et al., 2005), and MPER for immunogens, as it contains epitopes to broadly neutralizing antibodies (bnAbs), such as 2F5, 4E10, Z13, and 10E8 (Los Alamos HIV Molecular Immunology Database, 2021a).

HIV is one of the most genetically diverse pathogens due to its high mutation and recombination rates, large population size, and rapid replication rate (Hemelaar, 2012). The HIV epidemic is the result of two types of viruses: HIV-1 and HIV-2, which are closely related to SIVcpz (Gao et al., 1999) and SIVsm (Gao et al., 1992), respectively. HIV-1 causes most of the HIV infections worldwide and has been divided according to genetic homology into four groups: M (major or main), N (non-M, non-O) (Simon et al., 1998), O (outlier) (De Leys et al., 1990), and P (Plantier et al., 2009). However, the global HIV epidemic is related to group M (Hemelaar et al., 2019), which has been subdivided into 10 subtypes (A–D, F–H, J–L) and eight sub-subtypes (A1, A2, A3, A4, A5, A6, F1, F2) (Robertson et al., 2000; Salminen, 2000; Leitner et al., 2005; Yamaguchi et al., 2020), at least 118 circulating

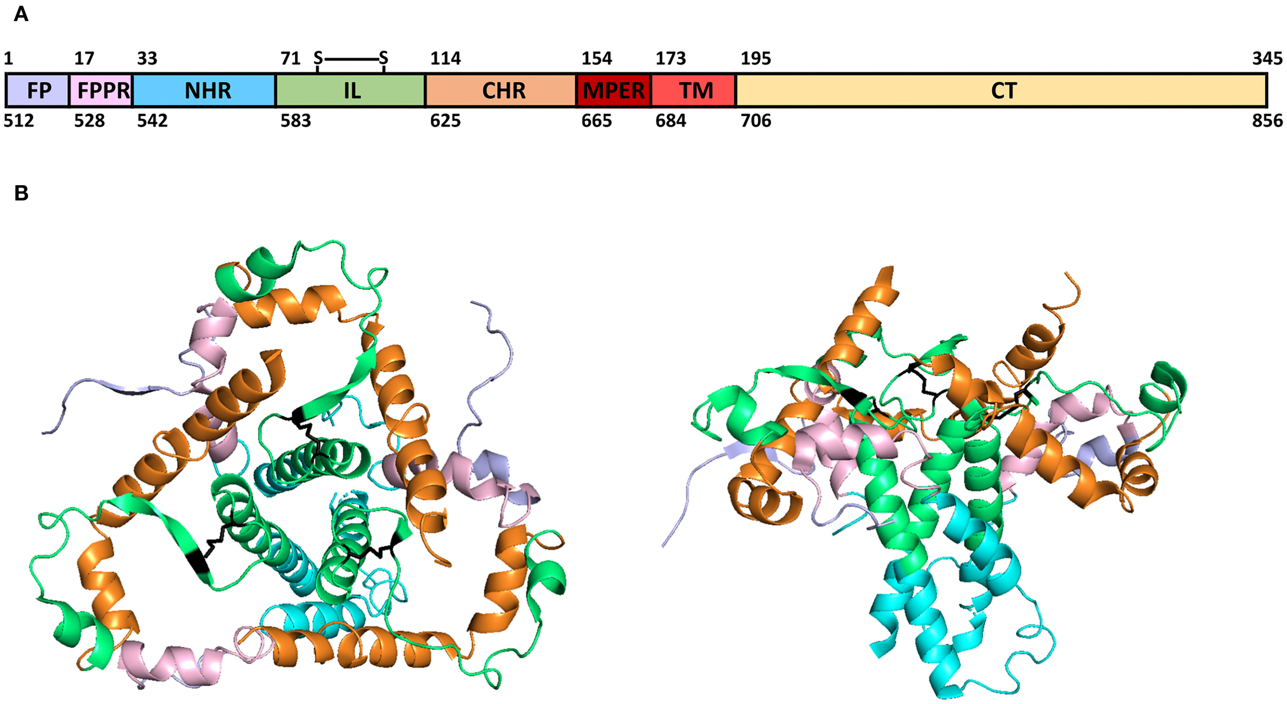


**FIGURE 1 |** HIV-1 gp41 structural domains **(A)** and 3D model structure **(B)**. Color code: light-purple (fusion peptide, FP), light-pink (fusion peptide proximal region, FPPR), blue (N-terminal alpha-helical region, NHR), green (immune-dominant linker, IL), orange (C-terminal heptad repeat region, CHR), dark-red (membrane-proximal external region, MPER), red (transmembrane region, TM), light-yellow (C-terminal domain, CT), and black (disulfide-bridged loop in IL, S-S). **(A)** Over-domains numbering according to HXB2 isolate aa and below-domains numbering according to HXB2 gp160 nucleotides (Genbank Accession number: K03455). **(B)** Gp41 3D model structure (PDB ID: 6OLP) (Berman et al., 2000; Rantalainen and Cottrell, 2019; Torrents de la Peña et al., 2019).

recombinant forms (CRF) (Los Alamos HIV Sequence Database, 2021a) and uncountable unique recombinant forms (URF). HIV-2 has been classified into nine groups (A-I) and two recombinants (CRF01_AB and URF) (Visseaux et al., 2016).

Since the HIV transmembrane glycoprotein is a key target for human bnAbs and anti-HIV drugs, a better knowledge of aa conservation across structural domains and HIV variants can help to identify conserved targets to direct new therapeutic and diagnostic strategies. Furthermore, each HIV variant presents natural polymorphisms and unique aa changes (V-markers) along the viral genome fixed during viral evolution (Arenas et al., 2016), which have not been described to date in each HIV variant, mainly in HIV-1 group M non-B subtypes and recombinants, which are the majority in the pandemic (Hemelaar et al., 2019), and in HIV-2.

We present, for the first time, the most conserved gp41 domains in each HIV-1 variant per structural domain and anti-gp41 antibody binding domains in the largest panel of HIV-1 variants analyzed to date, identifying the V-markers and the consensus transmembrane glycoprotein sequence for each HIV variant (type, group, subtype, sub-subtype, and CRF).

# MATERIALS AND METHODS

## HIV Transmembrane Glycoprotein Sequences

In October and November of 2020, we downloaded all available gp41 (HIV-1, 345 aa) and gp36 (HIV-2, 350 aa) nucleotides sequences from Los Alamos National Laboratory (LANL) HIV Sequence Database (Los Alamos HIV Sequence Database, 2021b) selecting one sequence per patient and grouping them per HIV variant (types, groups, subtypes, sub-subtypes, and CRF). URF sequences were not included in this study. They were aligned, edited, and translated into aa with the MEGAv6.0 program (Molecular Evolutionary Genetics Analysis: https://www.megasoftware.net/) (Tamura et al., 2013). MUSCLE function (Multiple Sequence Comparison by Log Expectation) (Edgar, 2004) was used for alignments with HXB2 (HIV-1 subtype B, GenBank accession number: K03455) or BEN (HIV-2 subtype A, GenBank accession number: M30502) reference sequences, removing nucleotides insertions. Sequences with stop codons in unusual positions and groups, subtypes, sub-subtypes, and CRF with <3 sequences were excluded from the study, except group P, as it was necessary to establish the HIV-1 aa sequence consensus.

## Gp41/Gp36 aa Conservation and Inferred Consensus Sequences

Using a bioinformatics tool developed in our laboratory (EpiMolBio program), we analyzed the gp41/gp36 aa conservation of HIV variants with at least three available sequences (except group P, with two sequences). We also inferred the aa consensus sequences for gp41 HIV-1/gp36 HIV-2 and each HIV-1/HIV-2 variant, providing the most conserved aa in each residue.

The new EpiMolBio bioinformatics tool reported the percentage of sequences with a conserved aa in each position of any protein, establishing a color code for this study depending on the frequency of each conserved aa in gp41/gp36: white (<90%), light-blue/light-pink (≥90–<100%), and dark blue-green/fuchsia (100% or complete conservation across considered sequences), respectively. We also studied the level of conserved aa per gp41 structural domain in each HIV-1 non-M group, group M, and HIV-1 consensus sequences. For the analysis, we summed the aa conservation percentages (percentages of most conserved aa) of gp41 residues involved in each secondary structural domain and then divided the sum by the total number of residues per domain.

Finally, we used WebLogo (https://weblogo.berkeley.edu/logo.cgi) (Crooks et al., 2004) to generate a figure showing the HIV-1 and HIV-1 group M gp41, as well as the HIV-2 gp36 consensus sequences, including in each protein position the most frequent aa present in the corresponding alignment. The HIV-1 group M gp41 consensus sequence was generated after the alignment of group M variants consensus sequences, the HIV-1 gp41 consensus sequence was generated after the alignment of the HIV-1 groups (M, N, O, P) consensus sequences, and the HIV-2 gp36 consensus sequence was generated after the alignment of the HIV-2 groups A, B, and the CRF01_AB consensus sequences. The aa letters were represented as large as their conservation percentage with a color code according to their side-chain: in black, non-polar aliphatic (glycine, G; alanine, A; valine, V; leucine, L; methionine, M; isoleucine, I); in dark blue-green, aromatic (phenylalanine, F; tyrosine, Y; tryptophan, W); in fuchsia, polar uncharged (serine, S; threonine, T; cysteine, C; proline, P; asparagine, N; glutamine; Q); in light-blue, positively charged (lysine, K; arginine, R; histidine, H); and in light-pink, negatively charged (aspartic acid, D; glutamic acid, E). Deletions were represented by a yellow "X".

## HIV-1 Monoclonal Human Antibodies Location

We analyzed the overall conservation in each gp41 secondary-structure domain across the four HIV-1 groups and in the HIV-1 consensus sequence. We also identified the aa conservation level in each HIV-1 group and each anti-gp41 human Abs (bnAbs and non-bnAbs) binding domain described in the LANL HIV Immunology Database (Los Alamos HIV Molecular Immunology Database, 2021a), showing lineal epitopes in blue and non-linear epitopes recognized by bnAbs in orange.

## Gp41/Gp36 Natural Polymorphisms and V-Markers Across HIV Variants

We described the gp41/gp36 natural polymorphisms and V-markers in HIV-1 and HIV-2 variants using the EpiMolBio program. To identify the aa changes present in ≥90% sequences in each HIV-1 or HIV-2 variant (natural polymorphisms), we compared all gp41 sequences with the HXB2 isolate, HIV-1 consensus, and HIV-1 group M consensus sequences, and all gp36 sequences with BEN isolate, and the HIV-2 consensus sequence. Among the natural polymorphisms found, we identified the exclusive V-markers of each HIV-1 or HIV-2 variant, not present in any other HIV variant. The color code for that analysis was light-blue (≥90–<100%) and dark blue-green (100%) for HIV-1 and light-pink (≥90–<100%) and fuchsia (100%) for HIV-2.

We also studied the presence of major T-20 resistance mutations according to the 2019 edition of the International Antiviral Society–USA (2019 IAS-USA) drug resistance mutations list (Wensing et al., 2019) among the V-markers found. Moreover, we looked for L44M change due to association with 1.8-fold resistance to T-20 *in vitro* (Mink et al., 2005).

Furthermore, we examined the presence of natural polymorphisms and specific V-markers on four gp41 positions (N160, W161, F162, and W169), described as key for HIV-1 neutralization by 10E8 (Huang et al., 2012; Kwon et al., 2016), which is a highly potent bnAb-recognizing gp41 MPER (epitope NWFDISNWLWYIK, gp41 positions 160-172) (Los Alamos HIV Molecular Immunology Database, 2021b). Other key reasons to study 10E8 bnAbs were that it was not detected by some serological diagnostic tests targeting gp41 (Smith et al., 2021) and was recently used to design new strategies for the development of a more efficient HIV-1 vaccine (Kuchar et al., 2021).

Finally, some gp41 changes (S23P, S23A, T25A, T27A, and I48P) affecting infectivity, gp41 structure or function (Alsahafi et al., 2015; Lu et al., 2019), were also studied across HIV-1 variants.

# RESULTS

## Analyzed Gp41/Gp36 Sequences and Inferred Consensus Sequences

We downloaded all 18,348 HIV transmembrane glycoprotein sequences from the LANL database. Once sequences with stop codons in unusual positions were discarded and after excluding variants with <3 sequences (except HIV-1 group P), a total of 17,078 gp41/gp36 sequences from 108 variants, including types, groups, subtypes, sub-subtypes, and CRF, were finally used in this study: 16,803 gp41 (HIV-1, 105 variants) and 275 gp36 (HIV-2, three variants) (**Table 1**). Among the HIV-1 gp41 sequences, 99 belonged to non-M groups (N, O, and P) and 16,704 were ascribed to group M (nine subtypes, six sub-subtypes, and 87 CRF). The gp41 sequences from group M sub-subtype A5, subtype F, CRF30_0206, CRF84_A1D, CRF91_01C, CRF94_cpx, CRF97_01B, CRF101_01B, and CRF102_0107 were not available in LANL. Gp36 sequences from groups E, H, and I were also absent. The variants with the highest representation in HIV-1 group M were subtype B (48.5%), subtype C (23.9%), and recombinant CRF01_AE (13.1%). In HIV-2, the most represented group was A (85.5%).

Consensus sequences at aa level were inferred by EpiMolBio for HIV-1 group M (**Figure 2A**), HIV-1 (**Figure 2B**), and HIV-2 (**Figure 2C**) to study the homology at aa level across variants. The gp41 HIV-1 consensus sequence was generated after aligning the four HIV-1 groups (M, N, O, P) consensus sequences and the HIV-1 group M consensus after aligning 102 group M variants with at least three sequences. The gp36 HIV-2

**TABLE 1** | Gp41/gp36 LANL sequences were analyzed in this study.

| Variants | | | | N° SEQS |
|---|---|---|---|---|
| | | | N | 11 |
| HIV-1 | Non-M Groups | | O | 86 |
| | | | P* | 2 |
| HIV-1 | Group M | Subtypes | A | 20 |
| | | | A1 | 752 |
| | | | A2 | 8 |
| | | | A3 | 3 |
| | | | A4 | 2 |
| | | | A5 | 0 |
| | | | A6 | 167 |
| | | | B | 8106 |
| | | | C | 3985 |
| | | | D | 183 |
| | | | F | 0 |
| | | | F1 | 122 |
| | | | F2 | 15 |
| | | | G | 153 |
| | | | H | 11 |
| | | | J | 8 |
| | | | K | 3 |
| | | | L | 3 |
| HIV-1 | Group M | CRF | CRF01_AE | 2186 |
| | | | CRF02_AG | 291 |
| | | | CRF03_AB | 5 |
| | | | CRF04_cpx | 8 |
| | | | CRF05_DF | 4 |
| | | | CRF06_cpx | 14 |
| | | | CRF07_BC | 126 |
| | | | CRF08_BC | 62 |
| | | | CRF09_cpx | 5 |
| | | | CRF10_CD | 3 |
| | | | CRF11_cpx | 26 |
| | | | CRF12_BF | 18 |
| | | | CRF13_cpx | 10 |
| | | | CRF14_BG | 12 |
| | | | CRF15_01B | 9 |
| | | | CRF16_A2D | 4 |
| | | | CRF17_BF | 6 |
| | | | CRF18_cpx | 5 |
| | | | CRF19_cpx | 5 |
| | | | CRF20_BG | 4 |
| | | | CRF21_A2D | 4 |
| | | | CRF22_01A1 | 13 |
| | | | CRF23_BG | 2 |
| | | | CRF24_BG | 4 |
| | | | CRF25_cpx | 5 |
| | | | CRF26_A5U | 5 |
| | | | CRF27_cpx | 4 |
| | | | CRF28_BF | 5 |
| | | | CRF29_BF | 7 |
| | | | CRF30_0206 | 0 |

*(Continued)*

**TABLE 1** | Continued

| Variants | | | | N° SEQS |
|---|---|---|---|---|
| HIV-1 | Group M | CRF | CRF31_BC | 3 |
| | | | CRF32_06A6 | 3 |
| | | | CRF33_01B | 7 |
| | | | CRF34_01B | 3 |
| | | | CRF35_AD | 21 |
| | | | CRF36_cpx | 3 |
| | | | CRF37_cpx | 4 |
| | | | CRF38_BF | 1 |
| | | | CRF39_BF | 3 |
| | | | CRF40_BF | 4 |
| | | | CRF41_CD | 3 |
| | | | CRF42_BF | 13 |
| | | | CRF43_02G | 5 |
| | | | CRF44_BF | 3 |
| | | | CRF45_cpx | 5 |
| | | | CRF46_BF | 8 |
| | | | CRF47_BF | 3 |
| | | | CRF48_01B | 3 |
| | | | CRF49_cpx | 5 |
| | | | CRF50_A1D | 4 |
| | | | CRF51_01B | 7 |
| | | | CRF52_01B | 3 |
| | | | CRF53_01B | 4 |
| | | | CRF54_01B | 3 |
| | | | CRF55_01B | 9 |
| | | | CRF56_cpx | 4 |
| | | | CRF57_BC | 7 |
| | | | CRF58_01B | 6 |
| | | | CRF59_01B | 8 |
| | | | CRF60_BC | 5 |
| | | | CRF61_BC | 3 |
| | | | CRF62_BC | 3 |
| | | | CRF63_02A | 15 |
| | | | CRF64_BC | 8 |
| | | | CRF65_cpx | 6 |
| | | | CRF66_BF | 3 |
| | | | CRF67_01B | 2 |
| | | | CRF68_01B | 3 |
| | | | CRF69_01B | 7 |
| | | | CRF70_BF | 3 |
| | | | CRF71_BF | 15 |
| | | | CRF72_BF | 5 |
| | | | CRF73_BG | 2 |
| | | | CRF74_01B | 3 |
| | | | CRF75_BF | 3 |
| | | | CRF76_01B | 2 |
| | | | CRF77_cpx | 4 |
| | | | CRF78_cpx | 3 |
| | | | CRF79_0107 | 3 |
| | | | CRF80_0107 | 2 |
| | | | CRF81_cpx | 2 |

*(Continued)*

**TABLE 1** | Continued

| Variants | | | | N° SEQS |
|---|---|---|---|---|
| HIV-1 | Group M | CRF | CRF82_cpx | 6 |
| | | | CRF83_cpx | 11 |
| | | | CRF84_A1D | 0 |
| | | | CRF85_BC | 11 |
| | | | CRF86_BC | 3 |
| | | | CRF87_cpx | 3 |
| | | | CRF88_BC | 3 |
| | | | CRF89_BF | 3 |
| | | | CRF90_BF1 | 6 |
| | | | CRF91_01C | 0 |
| | | | CRF92_C2U | 5 |
| | | | CRF93_cpx | 3 |
| | | | CRF94_cpx | 0 |
| | | | CRF95_02B | 5 |
| | | | CRF96_cpx | 3 |
| | | | CRF97_01B | 0 |
| | | | CRF98_06B | 1 |
| | | | CRF99_BF | 2 |
| | | | CRF100_01C | 3 |
| | | | CRF101_01B | 0 |
| | | | CRF102_0107 | 0 |
| | | | CRF103_01B | 4 |
| HIV-2 | Groups | | A | 235 |
| | | | B | 34 |
| | | | C | 1 |
| | | | D | 1 |
| | | | E | 0 |
| | | | F | 2 |
| | | | G | 1 |
| | | | H | 0 |
| | | | I | 0 |
| HIV-2 | | CRF | CRF01_AB | 6 |

*In red, variants with <3 sequences. Only HIV-1 group P, with 2 sequences, was included in the study (with an asterisk). N°, number; SEQS, sequences; CRF, circulating recombinant forms.*

consensus sequence was inferred after aligning the 275 gp36 LANL sequences ascribed to groups A, B, and CRF01_AB. **Supplementary Tables 1**, **2** show the inferred HIV-1 and HIV-2 transmembrane consensus sequences in this study, respectively.

## Amino Acid Conservation of Gp41 and Gp36 Across HIV Variants

We identified the gp41 and gp36 residues with ≥90% and 100% conservation across sequences from each analyzed variant (**Table 2**). The gp41 HIV-1 consensus aa sequence had 38.3% of the 345 gp41 residues conserved in ≥90% of gp41 sequences, and 9% of aa were 100% conserved. Higher conservation of gp36 HIV-2 aa consensus sequence was observed compared to HIV-1 gp41 aa consensus sequence, with 59.7% of 350 gp36 residues conserved in ≥90% of gp36

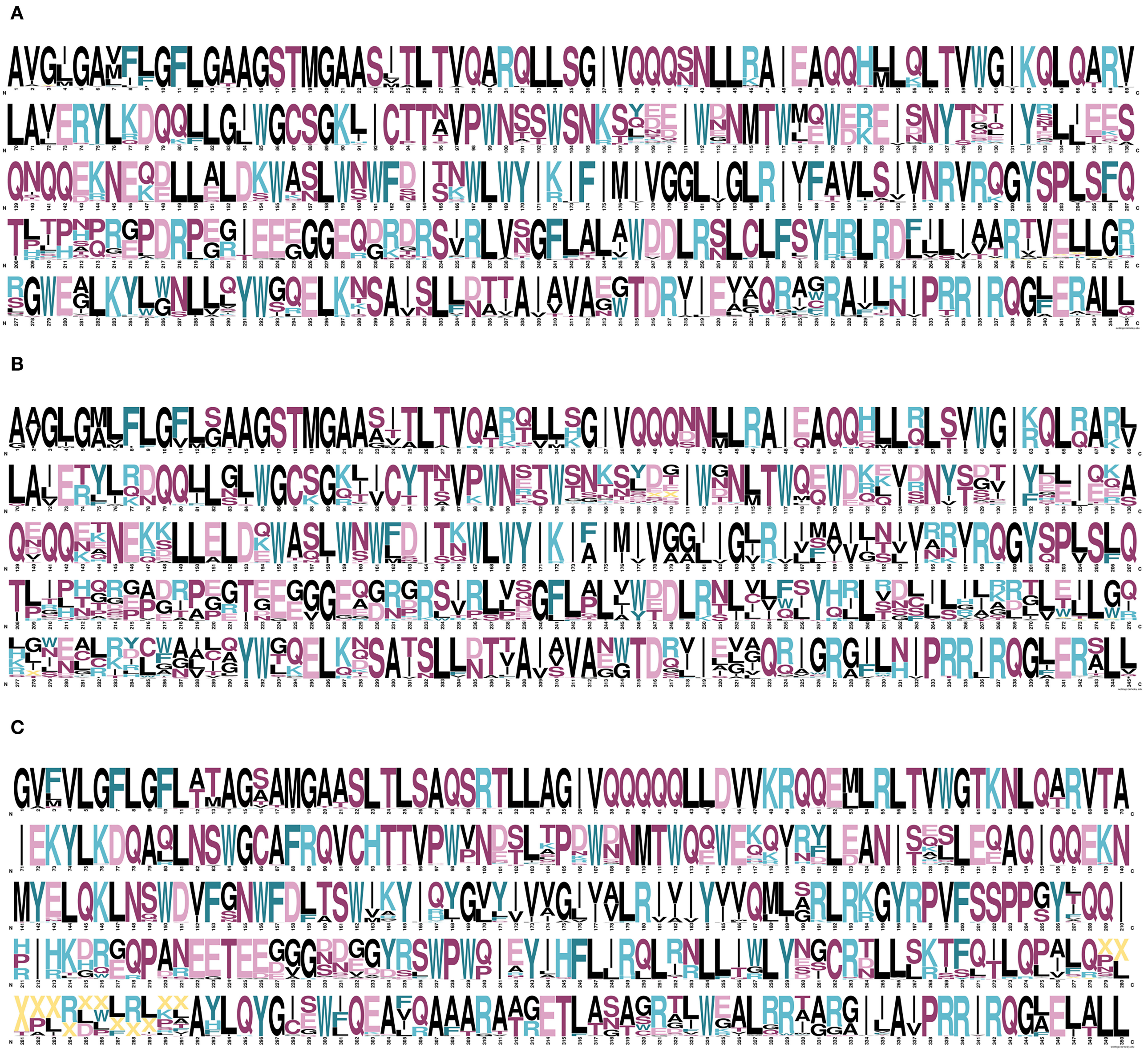


**FIGURE 2 |** Inferred consensus transmembrane sequences for HIV-1 group M **(A)**, HIV-1 **(B)**, and HIV-2 **(C)**. Each residue includes the most frequent aa present in the corresponding alignment. The aa code letters were represented in proportion to their percentage of conservation. Color code according to their side-chain: non-polar aliphatic (glycine, G; alanine, A; valine, V; leucine, L; methionine, M; isoleucine, I) in black; aromatic (phenylalanine, F; tyrosine, Y; tryptophan, W) in dark blue-green; polar uncharged (serine, S; threonine, T; cysteine, C; proline, P; asparagine, N; glutamine, Q) in fuchsia; positively charged (lysine, K; arginine, R; histidine, H) in light-blue, and negatively charged (aspartic acid, D; glutamic acid, E) in light-pink. Deletions were represented by "X" in yellow.

sequences, and 12.6% of aa totally conserved. Thus, 4 out of 10 aa positions were ≥90% conserved in the HIV-1 gp41 consensus sequence, rising to 6 out of 10 in the HIV-2 gp36 consensus sequence.

For HIV-1 groups, the highest percentage of conserved aa in ≥90%/100% of their HIV-1 gp41 LANL sequences were found in groups P (88.7%/88.7%) and N (85.8%/73%), followed by O (62.3%/33.6%) and M (60.3%/7.5%). Therefore, group P was the HIV-1 non-M group with the highest number of highly conserved gp41 residues, while group O had the fewest.

Among the 16,704 group M gp41 sequences, CRF46_BF had the lowest number of aa present in ≥90% of sequences (64.1%, eight sequences) and CRF41_CD the highest (98.8%, three sequences). When considering the completely conserved (100%) residues, subtype C showed the lowest number (3.2%, 3,985 sequences) and CRF41_CD the highest (98.8%, three sequences). Thus, among group M variants, CRF41_CD was the most conserved variant, while subtype C was the least conserved.

Regarding HIV-2, the variant with the highest percentage of ≥90%/100% conserved aa in gp36 protein was CRF01_AB

**TABLE 2 |** Number and percentage of highly (≥90%, 100%) conserved aa per HIV-1/HIV-2 variant.

| | | | | | Conservation | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | ≥90% | | 100% | |
| **Variants** | | | | **SEQS** | **N° AA** | **%** | **N° AA** | **%** |
| HIV-1 | HIV-1 Consensus | | | 4 | 132 | 38.3 | 31 | 9 |
| | | | N | 11 | 296 | 85.8 | 252 | 73 |
| | Non-M Groups | | O | 86 | 215 | 62.3 | 116 | 33.6 |
| | | | P | 2 | 306 | 88.7 | 306 | 88.7 |
| HIV-1 | Group M Consensus | | | 102 | 208 | 60.3 | 26 | 7.5 |
| | Group M | Subtypes | A | 20 | 272 | 78.8 | 188 | 54.5 |
| | | | A1 | 752 | 241 | 69.9 | 60 | 17.4 |
| | | | A2 | 8 | 236 | 68.4 | 236 | 68.4 |
| | | | A3 | 3 | 296 | 85.8 | 296 | 85.8 |
| | | | A6 | 167 | 287 | 83.2 | 112 | 32.5 |
| | | | B | 8106 | 247 | 71.6 | 22 | 6.4 |
| | | | C | 3985 | 240 | 69.6 | 11 | 3.2 |
| | | | D | 183 | 257 | 74.5 | 58 | 16.8 |
| | | | F1 | 122 | 263 | 76.2 | 95 | 27.5 |
| | | | F2 | 15 | 241 | 69.9 | 199 | 57.7 |
| | | | G | 153 | 247 | 71.6 | 90 | 26.1 |
| | | | H | 11 | 274 | 79.4 | 204 | 59.1 |
| | | | J | 8 | 241 | 69.9 | 241 | 69.9 |
| | | | K | 3 | 289 | 83.8 | 289 | 83.8 |
| | | | L | 3 | 276 | 80 | 276 | 80 |
| HIV-1 | Group M | CRF | CRF01_AE | 2186 | 260 | 75.4 | 43 | 12.5 |
| | | | CRF02_AG | 291 | 260 | 75.4 | 66 | 19.1 |
| | | | CRF03_AB | 5 | 311 | 90.1 | 311 | 90.1 |
| | | | CRF04_cpx | 8 | 249 | 72.2 | 249 | 72.2 |
| | | | CRF05_DF | 4 | 270 | 78.3 | 270 | 78.3 |
| | | | CRF06_cpx | 14 | 255 | 73.9 | 204 | 59.1 |
| | | | CRF07_BC | 126 | 286 | 82.9 | 142 | 41.2 |
| | | | CRF08_BC | 62 | 281 | 81.4 | 173 | 50.1 |
| | | | CRF09_cpx | 5 | 290 | 84.1 | 290 | 84.1 |
| | | | CRF10_CD | 3 | 291 | 84.3 | 291 | 84.3 |
| | | | CRF11_cpx | 26 | 246 | 71.3 | 170 | 49.3 |
| | | | CRF12_BF | 18 | 245 | 71 | 185 | 53.6 |
| | | | CRF13_cpx | 10 | 285 | 82.6 | 236 | 68.4 |
| | | | CRF14_BG | 12 | 306 | 88.7 | 268 | 77.7 |
| | | | CRF15_01B | 9 | 243 | 70.4 | 243 | 70.4 |
| | | | CRF16_A2D | 4 | 273 | 79.1 | 273 | 79.1 |
| | | | CRF17_BF | 6 | 246 | 71.3 | 246 | 71.3 |
| | | | CRF18_cpx | 5 | 270 | 78.3 | 270 | 78.3 |
| | | | CRF19_cpx | 5 | 271 | 78.6 | 271 | 78.6 |
| | | | CRF20_BG | 4 | 291 | 84.3 | 291 | 84.3 |
| | | | CRF21_A2D | 4 | 272 | 78.8 | 272 | 78.8 |
| | | | CRF22_01A1 | 13 | 266 | 77.1 | 202 | 58.6 |
| | | | CRF24_BG | 4 | 299 | 86.7 | 299 | 86.7 |
| | | | CRF25_cpx | 5 | 270 | 78.3 | 270 | 78.3 |
| | | | CRF26_A5U | 5 | 257 | 74.5 | 257 | 74.5 |
| | | | CRF27_cpx | 4 | 254 | 73.6 | 254 | 73.6 |
| | | | CRF28_BF | 5 | 268 | 77.7 | 268 | 77.7 |
| | | | CRF29_BF | 7 | 236 | 68.4 | 236 | 68.4 |
| | | | CRF31_BC | 3 | 291 | 84.3 | 291 | 84.3 |
| | | | CRF32_06A6 | 3 | 321 | 93 | 321 | 93 |
| | | | CRF33_01B | 7 | 266 | 77.1 | 266 | 77.1 |
| | | | CRF34_01B | 3 | 327 | 94.8 | 327 | 94.8 |
| | | | CRF35_AD | 21 | 289 | 83.8 | 219 | 63.5 |
| | | | CRF36_cpx | 3 | 304 | 88.1 | 304 | 88.1 |
| | | | CRF37_cpx | 4 | 277 | 80.3 | 277 | 80.3 |
| | | | CRF39_BF | 3 | 283 | 82 | 283 | 82 |
| | | | CRF40_BF | 4 | 276 | 80 | 276 | 80 |
| | | | CRF41_CD | 3 | 341 | 98.8 | 341 | 98.8 |
| | | | CRF42_BF | 13 | 322 | 93.3 | 301 | 87.2 |
| | | | CRF43_02G | 5 | 277 | 80.3 | 277 | 80.3 |
| | | | CRF44_BF | 3 | 286 | 82.9 | 286 | 82.9 |
| | | | CRF45_cpx | 5 | 265 | 76.8 | 265 | 76.8 |
| | | | CRF46_BF | 8 | 221 | 64.1 | 221 | 64.1 |
| | | | CRF47_BF | 3 | 295 | 85.5 | 295 | 85.5 |
| | | | CRF48_01B | 3 | 299 | 86.7 | 299 | 86.7 |
| | | | CRF49_cpx | 5 | 268 | 77.7 | 268 | 77.7 |
| | | | CRF50_A1D | 4 | 298 | 86.4 | 298 | 86.4 |
| | | | CRF51_01B | 7 | 286 | 82.9 | 286 | 82.9 |
| | | | CRF52_01B | 3 | 302 | 87.5 | 302 | 87.5 |
| | | | CRF53_01B | 4 | 288 | 83.5 | 288 | 83.5 |
| | | | CRF54_01B | 3 | 310 | 89.9 | 310 | 89.9 |
| | | | CRF55_01B | 9 | 282 | 81.7 | 282 | 81.7 |
| | | | CRF56_cpx | 4 | 338 | 98 | 338 | 98 |
| | | | CRF57_BC | 7 | 247 | 71.6 | 247 | 71.6 |
| | | | CRF58_01B | 6 | 294 | 85.2 | 294 | 85.2 |
| | | | CRF59_01B | 8 | 295 | 85.5 | 295 | 85.5 |
| | | | CRF60_BC | 5 | 288 | 83.5 | 288 | 83.5 |
| | | | CRF61_BC | 3 | 330 | 95.7 | 330 | 95.7 |
| | | | CRF62_BC | 3 | 322 | 93.3 | 322 | 93.3 |
| | | | CRF63_02A | 15 | 302 | 87.5 | 259 | 75.1 |
| | | | CRF64_BC | 8 | 252 | 73 | 252 | 73 |
| | | | CRF65_cpx | 6 | 294 | 85.2 | 294 | 85.2 |
| | | | CRF66_BF | 3 | 282 | 81.7 | 282 | 81.7 |
| | | | CRF68_01B | 3 | 320 | 92.8 | 320 | 92.8 |
| | | | CRF69_01B | 7 | 280 | 81.2 | 280 | 81.2 |
| | | | CRF70_BF | 3 | 286 | 82.9 | 286 | 82.9 |
| | | | CRF71_BF | 15 | 253 | 73.3 | 212 | 61.4 |
| | | | CRF72_BF | 5 | 257 | 74.5 | 257 | 74.5 |
| | | | CRF74_01B | 3 | 300 | 87 | 300 | 87 |
| | | | CRF75_BF | 3 | 316 | 91.6 | 316 | 91.6 |
| | | | CRF77_cpx | 4 | 304 | 88.1 | 304 | 88.1 |
| | | | CRF78_cpx | 3 | 279 | 80.9 | 279 | 80.9 |
| | | | CRF79_0107 | 3 | 318 | 92.2 | 318 | 92.2 |
| | | | CRF82_cpx | 6 | 299 | 86.7 | 299 | 86.7 |
| | | | CRF83_cpx | 11 | 319 | 92.5 | 298 | 86.4 |

*(Continued)*

**TABLE 2** | Continued

| | | | | | Conservation | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | ≥90% | | 100% | |
| Variants | | | | SEQS | N° AA | % | N° AA | % |
| HIV-1 | Group M | CRF | CRF85_BC | 11 | 308 | 89.3 | 251 | 72.8 |
| | | | CRF86_BC | 3 | 305 | 88.4 | 305 | 88.4 |
| | | | CRF87_cpx | 3 | 312 | 90.4 | 312 | 90.4 |
| | | | CRF88_BC | 3 | 303 | 87.8 | 303 | 87.8 |
| | | | CRF89_BF | 3 | 290 | 84.1 | 290 | 84.1 |
| | | | CRF90_BF1 | 6 | 258 | 74.8 | 258 | 74.8 |
| | | | CRF92_C2U | 5 | 234 | 67.8 | 234 | 67.8 |
| | | | CRF93_cpx | 3 | 277 | 80.3 | 277 | 80.3 |
| | | | CRF95_02B | 5 | 294 | 85.2 | 294 | 85.2 |
| | | | CRF96_cpx | 3 | 305 | 88.4 | 305 | 88.4 |
| | | | CRF100_01C | 3 | 293 | 84.9 | 293 | 84.9 |
| | | | CRF103_01B | 4 | 302 | 87.5 | 302 | 87.5 |
| HIV-2 | HIV-2 Consensus | | | 3 | 209 | 59.7 | 44 | 12.6 |
| | Groups | | A | 235 | 231 | 66 | 55 | 15.7 |
| | | | B | 34 | 251 | 71.7 | 159 | 45.4 |
| | | CRF | CRF01_AB | 6 | 290 | 82.9 | 290 | 82.9 |

*Data showed the number of aa conserved in ≥90 or 100% of sequences per analyzed variant. Gp41 (HIV-1), 345 aa; gp36 (HIV-2), 350 aa. Gp41 HIV-1 consensus sequence was generated after the alignment of the four HIV-1 groups (M, N, O, P) consensus sequences, the HIV-1 group M consensus after the alignment of 102 group M variants with ≥3 sequences, and the HIV-2 consensus after the alignment of three HIV-2 variants. N°, number; SEQS, sequences; AA, amino acids; %, aa percentage; CRF, circulating recombinant forms.*

(82.9%/82.9%), followed by group B (71.7%/45.4%) and group A (66%/15.7%).

## Gp41 Conservation in Each HIV-1 Group per Structural Domain and Anti-Gp41 Antibody Binding Sites

When we studied the level of conserved aa in each gp41 structural domain, we observed that gp41 conservation differed between HIV-1 groups and structural domains (**Table 3**), ranging from 66.4% to 98.9% conservation. The gp41 domain in the HIV-1 consensus sequence with the highest conservation was NHR (86.2%), followed by FPPR (84.9%), MPER (84.2%), and FP (84%). We observed a high conservation percentage in FPPR (96%) and NHR (95.4%) domains in the group M consensus sequence, despite the lower overall gp41 conservation *vs.* the other three HIV-1 groups. Group N gp41 presented the most conserved FP (98.9%) and FPPR (98.8%) domains, group O the best conserved MPER domain (95.7%), and group P the most conserved NHR (98.7%), TM (95.5%), IL (95.3%), and CT (93.7%) gp41 domains (**Table 3**). CHR domain presented the highest conservation (95.0%) in groups N and P. Considering all analyzed HIV-1 variants, the FP, FPPR, NHR, and MPER were the gp41 domains with the highest (≥84%) number of conserved aa, being the target for most anti-gp41 Abs. The least conserved domains in the HIV-1 consensus sequence were TM (69%) and CT (66.4%) located inside the virus.

The conservation level in each domain for human anti-gp41 bnAbs and non-bnAbs across HIV-1 groups was also described (**Figure 3**), located between the FP and MPER domains since TM and CT (the less conserved gp41 domains according to our data) are not exposed. The bnAbs are directed to CHR (2F5) and mainly to the MPER domain (2F5, Z13, 4E10, 10E8, and derivatives). No anti-gp41 Ab-binding sites were shown in LANL in the first 13 aa of the FP domain, in the 17 first positions of NHR, in the last 9 residues of IL, and the first 14 aa of CHR. The same was true in the target peptide to the fusion inhibitor T-20 (GIVQQQNNLL, NHR, residues 36-45), even though it presented high conservation across HIV-1 groups (**Figure 3**).

## Natural Polymorphisms and V-Markers

All-natural polymorphisms (aa present in ≥90% sequences) found in each HIV variant are shown in **Supplementary Table 3** (HIV-1) and **Supplementary Table 4** (HIV-2). The number of polymorphisms in gp41 across HIV-1 non-B variants (different from subtype B) increased when using the HXB2 subtype B sequence as reference. HIV-2 group A presented fewer polymorphisms in gp36 when HIV-2 BEN subtype A group was used as reference. **Figure 4** shows the natural polymorphisms that could be considered as exclusive V-markers for HIV-1 gp41 non-M groups (86 V-markers, **Figure 4A**), for HIV-1 group M gp41 variants (120 V-markers, **Figure 4B**), and HIV-2 gp36 variants (24 V-markers, **Figure 4C**). The specific V-markers per variant can be found in **Supplementary Tables 3, 4**.

No V-markers associated with T-20 major resistance were found in HIV-1 groups N, O, P, group M, and HIV-2 variants. Only M44 was present in the group P consensus gp41 sequence (**Supplementary Tables 1, 3**). Group O gp41 consensus sequence carried D42 V-marker, but no T42, a residue associated with high T-20 resistance.

After the analysis of the four gp41 key positions (N160, W161, F162, and W169) for HIV-1 neutralization by the anti-MPER bnAb 10E8, we observed high (90–100%) conservation of W161 and W169 in the transmembrane consensus sequences across HIV-1 variants. L162 appeared in group O and L/M162 in group P gp41 consensus sequence. Residue 160 showed the highest variability, carrying the group P and many HIV-1 group M variants another aa, mainly serine.

Finally, regarding changes in gp41 residues (A23, P23, A25, A27, and P48) affecting infectivity, gp41 structure, or function, we observed that A23 appeared in groups P and O and in CRF63_02A gp41 consensus sequences, while A25 appeared in group O, subtype L, CRF41_CD, and CRF47_BF group M variants. V48 appeared in the CRF90_BF1 gp41 consensus sequence but P48 did not.

# DISCUSSION

This is the most up-to-date descriptive study related to HIV-1/HIV-2 transmembrane envelope proteins, providing the conservation level and the V-markers in each HIV variant, identifying the conserved gp41 domains in each HIV-1 group

**TABLE 3 |** Percentage of aa conservation in each gp41 structural domain across HIV-1 groups and HIV-1 consensus sequences.

| | % AA Conservation HIV-1 GP41 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Gp41 structural domains** | **FP** | **FPPR** | **NHR** | **IL** | **CHR** | **MPER** | **TM** | **CT** | **GP41** |
| Gp41 residues | 1–16 | 17–32 | 33–70 | 71–113 | 114–153 | 154–172 | 173–194 | 195–345 | 1–345 |
| Group N | **98.9** | 98.8 | 96.5 | 95.0 | 95.0 | 90.9 | 94.6 | 93.5 | 94.6 |
| Group O | **97.7** | 87.6 | 95.3 | 79.2 | 83.8 | 95.7 | 89.0 | 86.0 | 87.3 |
| Group P | 93.8 | 90.6 | **98.7** | 95.3 | 95.0 | 89.5 | 95.5 | 93.7 | 94.3 |
| Cons. Group M | 86.7 | **96.0** | 95.4 | 86.6 | 79.6 | 86.1 | 91.9 | 81.0 | 85.1 |
| Cons. HIV-1 | 84.0 | 84.9 | **86.2** | 73.3 | 70.7 | 84.2 | 69.0 | 66.4 | 72.8 |

*For the analysis, we used the aa conservation percentages per residue (percentages of most conserved aa) of each HIV-1 group and HIV-1 consensus sequence provided in* ***Supplementary Table 1*** *with color code. After identifying the residues of each secondary structural domain, the sum of the aa conservation percentages of each of them was divided by the total number of residues per domain. In color, the highest conservation level in each structural domain. In bold, the most conserved structural domain in each HIV-1 group and HIV-1 consensus sequence. Gp41 residues according to HXB2 HIV-1 isolate. N°, number; AA, amino acids; Cons, aa consensus sequence; FP, fusion peptide; FPPR, fusion peptide proximal region; NHR, N-terminal alpha-helical region; IL, immune-dominant linker; CHR, C-terminal heptad repeat region; MPER, membrane-proximal external region; TM, transmembrane region; CT, C-terminal domain.*

per structural domain and per anti-gp41 antibody binding site in the most extensive panel of HIV-1 gp41/HIV-2 gp36 sequences ($n = 17{,}078$) and variants (105 HIV-1 variants, three HIV-2 variants) analyzed to date. Our in-house bioinformatics tool was also used to establish consensus sequences for gp41/gp36 proteins to study the aa level conservation across HIV variants.

The results have shown that the degree of conservation of the protein can differ across HIV variants and transmembrane structural domains, as our group previously described with fewer sequences and variants (Holguín et al., 2007). The higher level of conserved residues across variants in gp36 (275 sequences) *vs.* gp41 (16,803 sequences) could be explained by the lower number of gp36 sequences in LANL due to the lower HIV-2 prevalence and worldwide infections (1–2 of 38 million HIV infections) explained by its lower transmissibility and virulence *vs.* HIV-1 (Azevedo-Pereira and Santos-Costa, 2016; Clinical Info and gov, 2019; Kapoor and Padival, 2021). A higher level of fully conserved aa observed in HIV-1 non-M groups (99 sequences) *vs.* group M (16,704 sequences) could also be explained by the lower prevalence of groups N, O, and P in the pandemic (Mourez et al., 2013). The same happens in most CRF, with few available gp41 sequences in LANL, due to their low prevalence and the absence of gp41 sequencing in countries where they circulate, with sequencing or research not always available.

All gp41 secondary structure ectodomains presented >70% of conservation (HIV-1 consensus sequence), which supports their key role in the viral cycle and the importance of structure maintenance for virus-cell membranes fusion and viral entry (Pancera et al., 2014). Our data revealed that the best-conserved gp41 domains were N-HR (86.2%), FPPR (84.9%), MPER (84.2%), and FP (84%). It is essential to highlight that precisely NHR, FP, and FPPR gp41 domains have recently been implicated in the interaction with the fusion inhibitor T-20, which prevents the virus from entering the cell (Xu et al., 2019).

Regarding the gp41-CT region (aa 195-345), located inside the virus, it was the least conserved domain in the HIV-1 and group M gp41 consensus sequences, despite its important role in gp41 structure and function (Fernandez and Freed, 2018). CT domain is involved in envelope conformation (Castillo-Menendez et al., 2018), being essential for efficient envelope incorporation into budding HIV-1 particles (Murakami and Freed, 2000), and requiring interaction with gag matrix protein (MA) (Wyma et al., 2000; Eastep et al., 2021). However, the domains in each protein involved in this interaction are still unknown (Fernandez and Freed, 2018). A link between the matrix trimers' formation and the binding between MA and gp41 CT has also been reported (Alfadhli et al., 2016).

The critical role of the FPPR or polar region at the N terminus of gp41 for HIV-1 fusion and infectivity by stabilizing envelope trimers (Lu et al., 2019) could explain the high gp41 aa conservation percentage found in this domain. S23 within FPPR is structurally essential for maintaining HIV-1 envelope trimer, viral fusogenicity, and infectivity (Lu et al., 2019). Single or combined mutations S23P, T25A, and T27A in the FPPR region abolished or significantly decreased HIV-1 infectivity without affecting viral production, and S23A change significantly reduced HIV-1 infectivity and fusogenicity but not envelope expression and cleavage (Lu et al., 2019). In our study, only S23A and T25A substitutions appeared in a low number of variants with low prevalence in the pandemic. The absence of I48P gp41 change in our sequence set could be explained by its high impact on the quaternary conformation and function of the envelope glycoprotein trimer (Alsahafi et al., 2015).

Analyzing the genetic variability of HIV glycoprotein transmembrane within its immunodominant epitopes is important for understanding its possible impact on HIV Abs detection (Dorn et al., 2000; Dong et al., 2005; Smith et al., 2021). HIV-1 bnAbs can neutralize most HIV-1 strains from diverse genetic and geographic backgrounds (Binley et al., 2004; Wang and Zhang, 2020). The anti-gp41 bnAbs can recognize the MPER (Huang et al., 2012) and FP domains (Yuan et al., 2019), as well as the gp120/gp41 interphase (Huang et al., 2014; Scharf et al., 2014; Wang and Zhang, 2020). The conformational plasticity of FP could facilitate the recognition

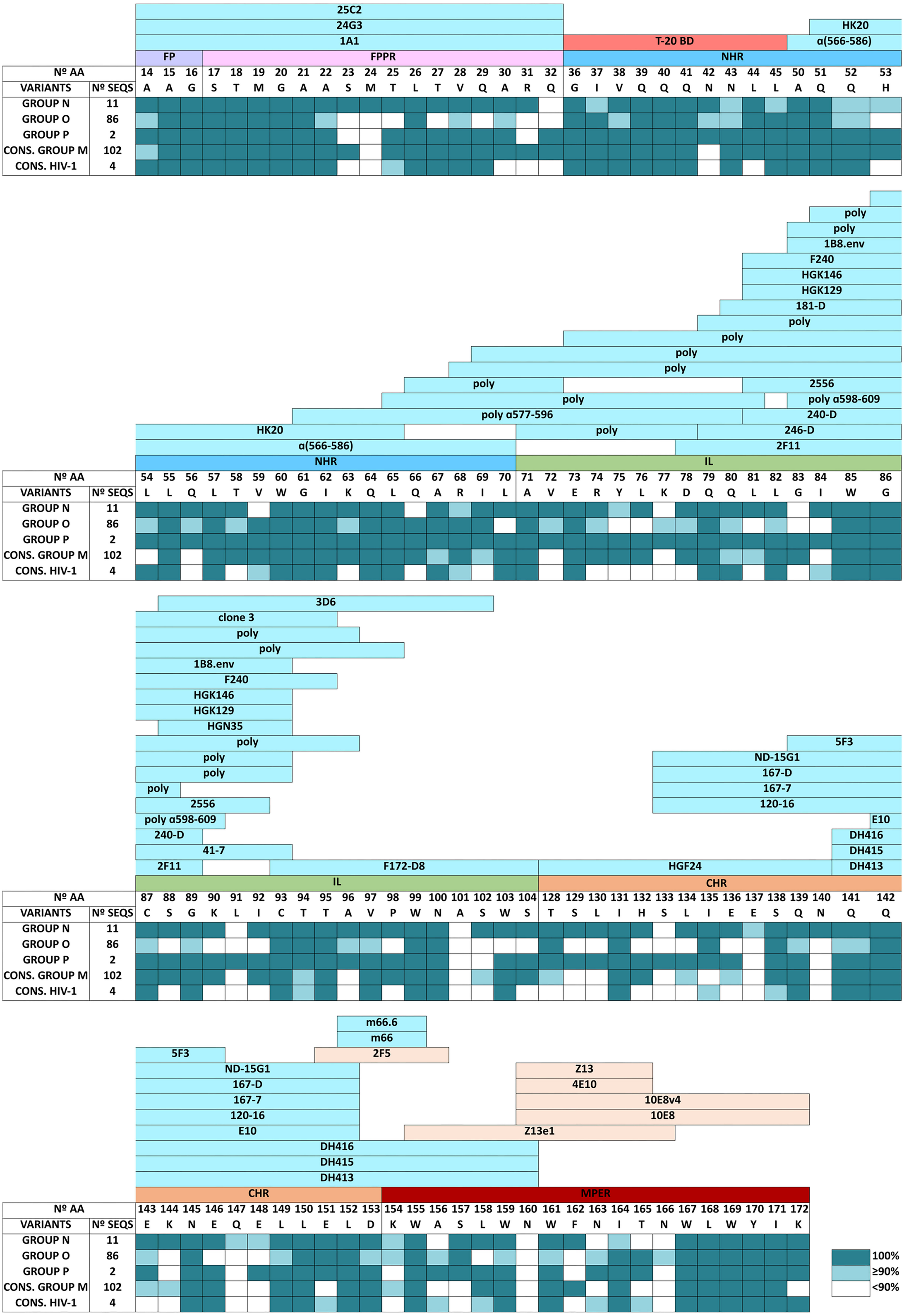


**FIGURE 3 |** Amino acid conservation level in gp41 residues involved in anti-gp41 human antibody binding domains described in LANL HIV Immunology Database across HIV-1 groups and HIV-1 consensus sequences. Data showed the percentage of the most conserved aa with the following conservation color code: white (aa

*(Continued)*

**FIGURE 3 |** conserved <90%), light-blue (aa conserved ≥90%–<100%), and dark blue-green (aa conserved 100%). Anti-gp41 human antibodies include bnAbs and non-bnAbs described in Los Alamos HIV Immune Database (Los Alamos HIV Molecular Immunology Database, 2021a). Linear epitopes are shown in blue and non-linear epitopes, recognized by broadly neutralizing antibodies (bnAbs), are shown in orange. We also indicated the T-20 fusion inhibitor binding domain in red (T-20BD). N°, number; AA, amino acids; SEQS, sequences; Cons, aa consensus sequence; FP, fusion peptide; FPPR, fusion peptide proximal region; NHR, N-terminal alpha-helical region; IL, immune-dominant linker; CHR, C-terminal heptad repeat region; MPER, membrane-proximal external region; TM, transmembrane region; CT, C-terminal domain; G, glycine; A, alanine; V, valine; L, leucine; M, methionine; I, isoleucine; F, phenylalanine; Y, tyrosine; W, tryptophan; S, serine; T, threonine; C, cysteine; P, proline; N, asparagine; Q, glutamine; K, lysine; R, arginine; H, histidine; D, aspartic acid; E, glutamic acid; poly, polyclonal.

**A**

| Nº AA | 11 | 30 | 35 | 42 | 44 | 49 | 53 | 68 | 78 | 81 | 90 | 92 | 104 | 106 | 123 | 125 | 134 | 143 | 148 | 154 | 157 | 179 | 180 | 184 | 187 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. HIV-1 | F | A | S | N | L | E | H | R | D | L | K | I | S | K | E | D | L | E | K | Q | S | G | G | L | I |
| V-MARKERS | V | T | K/H | D | M | Q | E | K | N | I | Q | V | T | T | K/L | R | E | N/K | S | E | Q | A | A | I | L |

| Nº AA | 188 | 191 | 192 | 195 | 196 | 197 | 202 | 211 | 213 | 215 | 217 | 218 | 219 | 220 | 226 | 229 | 238 | 245 | 246 | 247 | 251 | 252 | 254 | 255 | 258 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. HIV-1 | M | L | N | R | R | V | S | P | Q | G | D | R | P | E | G | Q | V | V | W | D | N | L | L | F | H |
| V-MARKERS | S | I/G | T | A | N | I | Q | L | G | P | G | I | A | P | V | A | L | L | Y | T | T | I | V | W/Q | Q |

| Nº AA | 259 | 261 | 267 | 269 | 270 | 272 | 275 | 276 | 277 | 278 | 279 | 280 | 281 | 282 | 284 | 286 | 287 | 288 | 289 | 290 | 293 | 298 | 310 | 317 | 324 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. HIV-1 | R | R | L | R | T | E | G | Q | L | G | N | E | A | L | Y | W | A | I | L | Q | G | N | A | R | R |
| V-MARKERS | L/I | S/V | I | D | L | T | W | I | H | L | I | N | C | R/C | D | F | G | A | C | A/G | L/T | D/Q | S | Q | Q |

**B**

| Nº AA | 1 | 2 | 3 | 4 | 7 | 9 | 23 | 46 | 48 | 49 | 62 | 74 | 76 | 77 | 80 | 81 | 101 | 107 | 108 | 109 | 110 | 113 | 123 | 125 | 129 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. GROUP M | A | V | G | I | V | L | S | R | I | E | I | R | L | K | Q | L | S | S | Y | E | E | D | E | S | N |
| V-MARKERS | X | X | T | F | A | I | A | Q/M | V | D | V | S | V | E | T | R | D/I | N | V/R | Q | S/A | Q/G | Q | G | E |

| Nº AA | 133 | 136 | 140 | 144 | 147 | 156 | 177 | 188 | 196 | 197 | 210 | 211 | 213 | 215 | 216 | 219 | 220 | 229 | 233 | 234 | 235 | 236 | 242 | 245 | 247 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. GROUP M | R | E | N | K | Q | A | I | F | R | V | T | P | P | G | P | P | E | Q | R | S | I | R | L | A | D |
| V-MARKERS | G/D | G | I | I | L | S/T | V | C | K | C | H/S | R | S | D | L/H | H | R | P | K | F | R/E | P/C | S | F | I |

| Nº AA | 249 | 254 | 264 | 268 | 269 | 270 | 271 | 272 | 278 | 280 | 281 | 282 | 284 | 285 | 286 | 287 | 292 | 294 | 295 | 297 | 298 | 299 | 300 | 301 | 302 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. GROUP M | L | L | I | A | R | T | V | E | G | E | A | L | Y | L | W | N | W | Q | E | K | N | S | A | I | S |
| V-MARKERS | I | I | T/G | X | K/X | A/X | G/X | V | V | Q | S | I | H | G | K | E | L | K | L | Q | Y | W | G | L | E/T |

| Nº AA | 304 | 306 | 307 | 308 | 310 | 311 | 312 | 313 | 315 | 316 | 319 | 322 | 323 | 325 | 326 | 330 | 331 | 335 | 340 | 342 | 343 | 344 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. GROUP M | L | T | T | A | A | V | A | E | T | D | I | V | Q | A | G | L | H | R | L | R | A | L |
| V-MARKERS | K/I | S/F | A | S | L/T | T | D | Q | A | I | A/L | T | D | L | I/V | Q/V | R | A | P | K/I | L | R |

**C**

| Nº AA | 12 | 53 | 99 | 119 | 127 | 170 | 178 | 182 | 190 | 205 | 240 | 253 | 280 | 282 | 283 | 284 | 285 | 287 | 288 | 289 | 299 | 331 | 345 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CONS. HIV-2 | A | M | V | V | S | V | A | V | S | G | Q | N | X | X | X | R | X | L | R | L | S | A | L |
| V-MARKERS | T | L | P | I | T | L | V | I | A | S | P | R | L | P | L | X | D/L | X | X | X | E | G | A |

**FIGURE 4 |** Location of exclusive transmembrane glycoprotein V-markers in non-M groups **(A)**, HIV-1 group M variants **(B)**, and HIV-2 variants **(C)**. V-markers in gp41 HIV-1 non-M groups are described with respect to the HIV-1 consensus sequence (yellow color). V-markers in gp41 HIV-1 group M variants are described with respect to the HIV-1 group M consensus sequence (light-blue color). V-markers in gp36 HIV-2 variants are described with respect to the HIV-2 consensus sequence (light-pink color). These V-markers and their corresponding variant can be also found in **Supplementary Table 3** (for HIV-1) and **Supplementary Table 4** (for HIV-2). V-markers, aa changes that were specific for each variant present in ≥90% of gp41 or gp36 sequences of an exclusive HIV-1 or HIV-2 variant; N°, number; AA, amino acids; CONS, aa consensus sequence; G, glycine; A, alanine; V, valine; L, leucine; M, methionine; I, isoleucine; F, phenylalanine; Y, tyrosine; W, tryptophan; S, serine; T, threonine; C, cysteine; P, proline; N, asparagine; Q, glutamine; K, lysine; R, arginine; H, histidine; D, aspartic acid; E, glutamic acid; X, aa deletion.

of the virus by bnAbs (Yuan et al., 2019). It is also known that HIV-1 variability can impact on bnAbs reactivity in HIV diagnostic tests targeting gp41, leading to non-reactive results with different serological diagnostic assays (Smith et al., 2021). Furthermore, the identification of HIV antibody binding domains is critical for vaccine development studies (Kuchar et al., 2021). Many bnAbs directed to gp41 have been described (Los Alamos HIV Molecular Immunology Database, 2021a), and the identification of the aa conservation level across HIV variants on the recognized epitopes by each bnAb and in those key gp41 residues for viral neutralization is of particular interest. The high variability found in key gp41 residue 160 across some HIV-1 variants could explain the previously reported failure of 10E8 recognition by some different serological diagnostic tests (Smith et al., 2021). We could not analyze the gp41 aa conservation in the gp41 target sequence per each failing diagnostic test because manufacturers do not provide detailed information regarding which part of the gp41 sequence was targeted in their HIV diagnostic assays detecting the transmembrane protein, which can also differ across assays. The provided information in the **Supplementary Tables** of the manuscript can help manufacturers and other researchers design new gp41-based molecular and serological diagnostic tests to identify those HIV-1 variants whose diagnosis could be compromised by viral genetic variability.

Natural infection by HIV-2 also leads to the elicitation of high titers of bnAbs against primary HIV-2 strains (De Silva et al., 2012; Kong et al., 2012; Özkaya Sahin et al., 2012), although not all bnAbs to HIV-2 neutralize HIV-1 variants (Björling et al., 1993). In fact, MPER-specific Abs induced by vaccination with recombinant gp36 proteins in rats did not neutralize HIV-2 (Behrendt et al., 2012). It is known that HIV types present different mechanisms for the processing of envelope glycoproteins from a smaller env precursor in HIV-2 (gp140) than in HIV-1 (gp160) (Rey et al., 1989). To date, unfortunately, the exact aa residues of each secondary structure in HIV-2 gp36 have not been specified. However, some conserved gp36 epitopes have been reported (Jadhav et al., 2011), as well as immunogenic sites, antibody binding sites in the TM and IL region of HIV-2 transmembrane gp36 (Chiodi et al., 1993), and other epitopes recognized by bnAbs in HIV-2 envelope gp140 (Kong et al., 2012).

Previous studies showed the importance of consensus sequence establishment to guide vaccine development (Ellenberger et al., 2002; Sliepen et al., 2019). For the first time, our study also provides the aa consensus sequence of the transmembrane glycoprotein in each HIV variant (type, group, subtype, sub-subtype, and CRF). Moreover, we showed the conservation level across their sequences, which could be helpful to look for highly conserved peptides to direct new ARV, Abs, aptamers, probes, or primers to control or diagnose HIV infection regardless of the HIV variant. Furthermore, we showed the first identification of specific natural polymorphisms of gp41 and gp36 that can be considered as V-markers for all HIV-1 and HIV-2 variants, which should be considered in the new strategies for developing HIV-1 vaccines based on epitopes recognized by bnAbs (Kuchar et al., 2021). The exclusive HIV V-markers identified in gp41/gp36 sequences could help in faster and preliminary HIV variant identification if required, before doing the phylogenetic study, the gold standard method for correct HIV variant characterization. New studies are required to evaluate the structural and biological impact of the different levels of aa conservation in gp41 across HIV-1 variants and the specific V-markers found in the viral transmembrane protein.

Long-term exposure to the first entry inhibitor T-20 induces drug-resistant mutations (Pérez-Alvarez et al., 2006). Interestingly, none of the variants had V-markers associated with major resistance to T-20, as we previously reported testing 79 different HIV variants from naïve patients (Holguín et al., 2007). Thus, no natural major resistance mutations to T-20 were observed. The L44M change found in group P consensus sequence, previously associated with 1.8-fold resistance to T-20 *in vitro* (Mink et al., 2005), was previously found in T-20 naïve subjects from China (Chang et al., 2021), maybe reflecting a resistance transmission during primoinfection. New HIV-1 fusion inhibitors are under development (Luque and Camarasa, 2021) and previous studies showed that optimized T-20 derivates could have been effective inhibitors of infection for multiple HIV-1 variants (Chen et al., 2019).

The main limitation of the study was the absence of LANL sequences from some HIV-1 (sub-subtype A5, subtype F, CRF30_0206, CRF84_A1D, CRF91_01C, CRF94_cpx, CRF97_01B, CRF101_01B, CRF102_0107) and HIV-2 variants (groups E, H, I), as well as the scarce number of HIV gp41/gp36 sequences in other 14 HIV variants with <3 sequences in LANL, which meant that they could not be included in the analysis (except group P).

The information provided in this manuscript aims to help other researchers studying the biological, therapeutical, diagnostic, or structural role of gp41 to identify the natural polymorphisms and specific V-markers per variant in each gp41/gp36 residue or epitope according to their interest. This study will also be useful for a more rational design of anti-gp41 drugs and vaccines and future HIV molecular diagnostic tests directed to transmembrane HIV protein.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Files**, further inquiries can be directed to the corresponding authors.

## AUTHOR CONTRIBUTIONS

AV-A downloaded and analyzed the HIV LANL sequences under study, validated some EpiMolBio functions necessary for sequences analyses, performed the computations, discussed results, and wrote the first draft of the manuscript. RR developed the in-house EpiMolBio bioinformatics program, validated some EpiMolBio functions necessary for sequences analyses, discussed results, and reviewed the final version of the manuscript. ÁH designed and supervised the study, reviewed and edited the manuscript, funding application, and project administration. All authors approved the final version submitted.

## FUNDING

This research and AV-A's contract were funded by Instituto de Salud Carlos III (PI18/00904 Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016) and co-financed by the European Regional Development Fund "A way to achieve Europe" (ERDF). RR was funded by FONDOS FUR 2020/0285. The funder had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

## ACKNOWLEDGMENTS

We thank Paul Devlin for proofreading the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.855232/full#supplementary-material

## REFERENCES

Alfadhli, A., Mack, A., Ritchie, C., Cylinder, I., Harper, L., Tedbury, P. R., et al. (2016). Trimer enhancement mutation effects on HIV-1 matrix protein binding activities. *J. Virol*. 90, 5657–5664. doi: 10.1128/JVI.00509-16

Alsahafi, N., Debbeche, O., Sodroski, J., and Finzi, A. (2015). Effects of the I559P gp41 change on the conformation and function of the human immunodeficiency virus (HIV-1) membrane envelope glycoprotein trimer. *PLoS ONE* 10:e0122111. doi: 10.1371/journal.pone.0122111

Arenas, M., Lorenzo-Redondo, R., and Lopez-Galindez, C. (2016). Influence of mutation and recombination on HIV-1 *in vitro* fitness recovery. *Mol. Phylogenet. Evol*. 94, 264–270. doi: 10.1016/j.ympev.2015.09.001

Azevedo-Pereira, J. M., and Santos-Costa, Q. (2016). HIV interaction with human host: HIV-2 as a model of a less virulent infection. *AIDS Rev*. 18, 44–53. Available online at: https://www.aidsreviews.com/resumen.php?id=1327&indice=2016181&u=unp

Behrendt, R., Fiebig, U., Kurth, R., and Denner, J. (2012). Induction of antibodies binding to the membrane proximal external region of gp36 of HIV-2. *Intervirology* 55, 252–256. doi: 10.1159/000324483

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res*. 28, 235–242. doi: 10.1093/nar/28.1.235

Binley, J. M., Wrin, T., Korber, B., Zwick, M. B., Wang, M., Chappey, C., et al. (2004). Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *J. Virol*. 78, 13232–13252. doi: 10.1128/JVI.78.23.13232-13252.2004

Björling, E., Scarlatti, G., von Gegerfelt, A., Albert, J., Biberfeld, G., Chiodi, F., et al. (1993). Autologous neutralizing antibodies prevail in HIV-2 but not in HIV-1 infection. *Virology* 193, 528–530. doi: 10.1006/viro.1993.1160

Blumenthal, R., Durell, S., and Viard, M. (2012). HIV entry and envelope glycoprotein-mediated fusion. *J. Biol. Chem*. 287, 40841–40849. doi: 10.1074/jbc.R112.406272

Caillat, C., Guilligay, D., Sulbaran, G., and Weissenhorn, W. (2020). Neutralizing antibodies targeting HIV-1 gp41. *Viruses* 12:E1210. doi: 10.3390/v12111210

Castillo-Menendez, L. R., Witt, K., Espy, N., Princiotto, A., Madani, N., Pacheco, B., et al. (2018). Comparison of uncleaved and mature human immunodeficiency virus membrane envelope glycoprotein trimers. *J. Virol*. 92, e00277–e00218. doi: 10.1128/JVI.00277-18

Chang, L., Zhao, J., Guo, F., Ji, H., Zhang, L., Jiang, X., et al. (2021). HIV-1 gp41 genetic diversity and enfuvirtide resistance-associated mutations among enfuvirtide-naïve patients in southern China. *Virus Res*. 292:198215. doi: 10.1016/j.virusres.2020.198215

Chen, G., Cook, J. D., Ye, W., Lee, J. E., and Sidhu, S. S. (2019). Optimization of peptidic HIV-1 fusion inhibitor T20 by phage display. *Protein. Sci*. 28, 1501–1512. doi: 10.1002/pro.3669

Chiodi, F., Björling, E., Samuelsson, A., and Norrby, E. (1993). Antigenic and immunogenic sites of HIV-2 glycoproteins. *Chem. Immunol*. 56, 61–77. doi: 10.1159/000319156

Clinical Info, H. I. V., and gov. (2019). *HIV-2 Infection*. Available online at: https://clinicalinfo.hiv.gov/es/node/9288 (accessed September 15, 2021).

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*. 14, 1188–1190. doi: 10.1101/gr.849004

De Leys, R., Vanderborght, B., Vanden Haesevelde, M., Heyndrickx, L., van Geel, A., Wauters, C., et al. (1990). Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of West-Central African origin. *J. Virol*. 64, 1207–1216. doi: 10.1128/jvi.64.3.1207-1216.1990

De Silva, T. I., Aasa-Chapman, M., Cotten, M., Hué, S., Robinson, J., Bibollet-Ruche, F., et al. (2012). Potent autologous and heterologous neutralizing antibody responses occur in HIV-2 infection across a broad range of infection outcomes. *J. Virol*. 86, 930–946. doi: 10.1128/JVI.06126-11

Dong, X.-N., Ying, J., Wu, Y., and Chen, Y.-H. (2005). Genetic variability of principal neutralizing determinants on HIV-1 gp41 and its correlation with subtypes. *Immunol. Lett*. 101, 104–107. doi: 10.1016/j.imlet.2005.04.013

Dorn, J., Masciotra, S., Yang, C., Downing, R., Biryahwaho, B., Mastro, T. D., et al. (2000). Analysis of genetic variability within the immunodominant epitopes of envelope gp41 from human immunodeficiency virus type 1 (HIV-1) group M and its impact on HIV-1 antibody detection. *J. Clin. Microbiol*. 38, 773–780. doi: 10.1128/JCM.38.2.773-780.2000

Eastep, G. N., Ghanam, R. H., Green, T. J., and Saad, J. S. (2021). Structural characterization of HIV-1 matrix mutants implicated in envelope incorporation. *J. Biol. Chem*. 296:100321. doi: 10.1016/j.jbc.2021.100321

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32, 1792–1797. doi: 10.1093/nar/gkh340

Ellenberger, D. L., Li, B., Lupo, L. D., Owen, S. M., Nkengasong, J., Kadio-Morokro, M. S., et al. (2002). Generation of a consensus sequence from prevalent and incident HIV-1 infections in West Africa to guide AIDS vaccine development. *Virology* 302, 155–163. doi: 10.1006/viro.2002.1577

Fernandez, M. V., and Freed, E. O. (2018). Meeting Review: 2018 International workshop on structure and function of the lentiviral gp41 cytoplasmic tail. *Viruses* 10:613. doi: 10.3390/v10110613

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., et al. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 397, 436–441. doi: 10.1038/17130

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., et al. (1992). Human infection by genetically diverse SIV SM-related HIV-2 in West Africa. *Nature* 358, 495–499. doi: 10.1038/358495a0

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med*. 18, 182–192. doi: 10.1016/j.molmed.2011.12.001

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis*. 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Holguín, A., De Arellano, E. R., and Soriano, V. (2007). Amino acid conservation in the gp41 transmembrane protein and natural polymorphisms associated with enfuvirtide resistance across HIV-1 variants. *AIDS Res. Hum. Retroviruses* 23, 1067–1074. doi: 10.1089/aid.2006.0256

Huang, J., Kang, B. H., Pancera, M., Lee, J. H., Tong, T., Feng, Y., et al. (2014). Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-120 interface. *Nature* 515, 138–142. doi: 10.1038/nature13601

Huang, J., Ofek, G., Laub, L., Louder, M. K., Doria-Rose, N. A., Longo, N. S., et al. (2012). Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* 491, 406–412. doi: 10.1038/nature11544

Jadhav, S., Tripathy, S., Kulkarni, S., Chaturbhuj, D., Ghare, R., Bhattacharya, J., et al. (2011). Genetic conservation in gp36 transmembrane sequences of Indian HIV type 2 isolates. *AIDS Res. Hum. Retroviruses* 27, 1337–1343. doi: 10.1089/aid.2011.0063

Kapoor, A. K., and Padival, S. (2021). *HIV-2. In StatPearls*. StatPearls Publishing. Available online at: http://www.ncbi.nlm.nih.gov/books/NBK572083/~ [Internet].

Kong, R., Li, H., Bibollet-Ruche, F., Decker, J. M., Zheng, N. N., Gottlieb, G. S., et al. (2012). Broad and potent neutralizing antibody responses elicited in natural HIV-2 infection. *J. Virol.* 86, 947–960. doi: 10.1128/JVI.06155-11

Kuchar, M., Kosztyu, P., Daniel Lišková, V., Cerný, J., Petroková, H., and Vróblová, E., et al. (2021). Myomedin scaffold variants targeted to 10E8 HIV-1 broadly neutralizing antibody mimic gp41 epitope and elicit HIV-1 virus-neutralizing sera in mice. *Virulence* 12, 1271–1287. doi: 10.1080/21505594.2021.1920251

Kwon, Y. D., Georgiev, I. S., Ofek, G., Zhang, B., Asokan, M., Bailer, R. T., et al. (2016). Optimization of the solubility of HIV-1-neutralizing antibody 10E8 through somatic variation and structure-based design. *J. Virol.* 90, 5899–5914. doi: 10.1128/JVI.03246-15

Lazzarin, A. (2005). Enfuvirtide: the first HIV fusion inhibitor. *Exp. Opin. Pharmacother.* 6, 453–464. doi: 10.1517/14656566.6.3.453

Leitner, T., Korber, B., Daniels, M., Calef, C., and Foley, B. (2005). HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005. *HIV Seq. Compend.* 2005, 41–48. Available online at: https://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2005/partI/leitner.pdf

Li, K., Xiu, C. L., Gao, L. M., Shi, M., and Zhai, Y. (2016). Subtractive SELEX using agar beads for screening DNA aptamers with specific affinity to HIV gp41 antigen. *Nan Fang Yi Ke Da Xue Xue Bao* 36, 1592–1598. doi: 10.3969/j.issn.1673-4254.2016.12.01

Los Alamos HIV Molecular Immunology Database (2021a). *Gp160 Ab Epitope Map*. Available online at: https://www.hiv.lanl.gov/content/immunology/maps/ab/gp160.html (accessed September 15, 2021).

Los Alamos HIV Molecular Immunology Database (2021b). *Search Antibody Database 10E8 MAb*. Available online at: https://www.hiv.lanl.gov/content/immunology/ab_search.html?results=Searchandprotein=gp160andinrange=671-683andmab_name=10E8 (accessed February 27, 2022).

Los Alamos HIV Sequence Database (2021a). *HIV Circulating Recombinant Forms (CRFs)*. Available online at: https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html#CRF01 (accessed September 15, 2021).

Los Alamos HIV Sequence Database (2021b). *Sequence Search Interface*. Available online at: https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html (accessed November 19, 2020).

Louis, J. M., Baber, J. L., Ghirlando, R., Aniana, A., Bax, A., and Roche, J. (2016). Insights into the conformation of the membrane proximal regions critical to the trimerization of the HIV-1 gp41 ectodomain bound to dodecyl phosphocholine micelles. *PLoS ONE* 11:e0160597. doi: 10.1371/journal.pone.0160597

Lu, W., Chen, S., Yu, J., Behrens, R., Wiggins, J., Sherer, N., et al. (2019). The polar region of the HIV-1 envelope protein determines viral fusion and infectivity by stabilizing the gp120-gp41 association. *J. Virol.* 93, e02128–e02118. doi: 10.1128/JVI.02128-18

Luque, F. J., and Camarasa, M. (2021). HIV-1 envelope spike MPER: from a vaccine target to a new druggable pocket for novel and effective fusion inhibitors. *Chem. Med. Chem.* 16, 105–107. doi: 10.1002/cmdc.202000411

Mink, M., Mosier, S. M., Janumpalli, S., Davison, D., Jin, L., Melby, T., et al. (2005). Impact of human immunodeficiency virus type 1 gp41 amino acid substitutions selected during enfuvirtide treatment on gp41 binding and antiviral potency of enfuvirtide in vitro. *J. Virol.* 79, 12447–12454. doi: 10.1128/JVI.79.19.12447-12454.2005

Mourez, T., Simon, F., and Plantier, J. C. (2013). Non-M variants of human immunodeficiency virus type 1. *Clin. Microbiol. Rev.* 26, 448–461. doi: 10.1128/CMR.00012-13

Murakami, T., and Freed, E. O. (2000). The long cytoplasmic tail of gp41 is required in a cell type-dependent manner for HIV-1 envelope glycoprotein incorporation into virions. *Proc. Natl. Acad. Sci. U.S.A.* 97, 343–348. doi: 10.1073/pnas.97.1.343

Oldfield, V., Keating, G. M., and Plosker, G. (2005). Enfuvirtide: a review of its use in the management of HIV infection. *Drugs* 65, 1139–1160. doi: 10.2165/00003495-200565080-00007

Özkaya Sahin, G., Holmgren, B., da Silva, Z., Nielsen, J., Nowroozalizadeh, S., and Esbjörnsson, J., et al. (2012). Potent intratype neutralizing activity distinguishes human immunodeficiency virus type 2 (HIV-2) from HIV-1. *J. Virol.* 86, 961–971. doi: 10.1128/JVI.06315-11

Pancera, M., Zhou, T., Druz, A., Georgiev, I. S., Soto, C., Gorman, J., et al. (2014). Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* 514, 455–461. doi: 10.1038/nature13808

Pérez-Alvarez, L., Carmona, R., Ocampo, A., Asorey, A., Miralles, C., Pérez de Castro, S., et al. (2006). Long-term monitoring of genotypic and phenotypic resistance to T20 in treated patients infected with HIV-1. *J. Med. Virol.* 78, 141–147. doi: 10.1002/jmv.20520

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nat. Med.* 15, 871–872. doi: 10.1038/nm.2016

Qadir, M. I., and Malik, S. A. (2010). HIV fusion inhibitors. *Rev. Med. Virol.* 20, 23–33. doi: 10.1002/rmv.631

Rantalainen, K., and Cottrell, C. A. (2019). Full length HIV-1 Env AMC011 in complex with PGT151 Fab. *PLoS Pathog.* (2019) 15:e1007920. doi: 10.2210/pdb6olp/pdb

Rey, M. A., Krust, B., Laurent, A. G., Guétard, D., Montagnier, L., and Hovanessian, A. G. (1989). Characterization of an HIV-2-related virus with a smaller sized extracellular envelope glycoprotein. *Virology* 173, 258–267. doi: 10.1016/0042-6822(89)90242-0

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–55. doi: 10.1126/science.288.5463.55d

Salminen, M. (2000). HIV inter-subtype recombination – Consequences for the epidemic. *AIDS Rev.* 3, 178–189. Available online at: https://www.researchgate.net/profile/Mika-Salminen/publication/254903484_HIV_Inter-subtype_Recombination_-_Consequences_for_the_Epidemic/links/53df80660cf2aede4b4905b9/HIV-Inter-subtype-Recombination-Consequences-for-the-Epidemic.pdf

Scharf, L., Scheid, J. F., Lee, J. H., West, A. P., Chen, C., Gao, H., et al. (2014). Antibody 8ANC195 reveals a site of broad vulnerability on the HIV-1 envelope spike. *Cell Rep.* 7, 785–795. doi: 10.1016/j.celrep.2014.04.001

Simon, F., Mauclère, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M. C., Saragosti, S., et al. (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med.* 4, 1032–1037. doi: 10.1038/2017

Sliepen, K., Han, B. W., Bontjer, I., Mooij, P., Garces, F., Behrens, A.-J., et al. (2019). Structure and immunogenicity of a stabilized HIV-1 envelope trimer based on a group-M consensus sequence. *Nat Commun.* 10:2355. doi: 10.1038/s41467-019-10262-5

Smith, T., Masciotra, S., Luo, W., Sullivan, V., Switzer, W. M., Johnson, J. A., et al. (2021). Broadly neutralizing HIV-1 antibody reactivity in HIV tests: implications for diagnostics. *AIDS* 35, 1561–1565. doi: 10.1097/QAD.0000000000002898

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Torrents de la Peña, A., Rantalainen, K., Cottrell, C. A., Allen, J. D., van Gils, M. J., Torres, J. L., et al. (2019). Similarities and differences between native HIV-1 envelope glycoprotein trimers and stabilized soluble trimer mimetics. *PLoS Pathog.* 15:e1007920. doi: 10.1371/journal.ppat.1007920

Visseaux, B., Damond, F., Matheron, S., Descamps, D., and Charpentier, C. (2016). HIV-2 molecular epidemiology. *Infect. Genet. Evol.* 46, 233–240. doi: 10.1016/j.meegid.2016.08.010

Wang, Q., and Zhang, L. (2020). Broadly neutralizing antibodies and vaccine design against HIV-1 infection. *Front. Med.* 14, 30–42. doi: 10.1007/s11684-019-0721-9

Wensing, A. M., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Günthard, H. F., Paredes, R., et al. (2019). 2019 Update of the drug resistance mutations in

HIV-1. *Top. Antivir. Med*. 27, 111–121. Available online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6892618/pdf/tam-27-111.pdf

Wyma, D. J., Kotov, A., and Aiken, C. (2000). Evidence for a stable interaction of gp41 with Pr55Gag in immature human immunodeficiency virus type 1 particles. *J. Virol*. 74, 9381–9387. doi: 10.1128/JVI.74.20.9381-9387.2000

Xu, W., Pu, J., Su, S., Hua, C., Su, X., Wang, Q., et al. (2019). Revisiting the mechanism of enfuvirtide and designing an analog with improved fusion inhibitory activity by targeting triple sites in gp41. *AIDS* 33, 1545–1555. doi: 10.1097/QAD.0000000000002208

Yamaguchi, J., Vallari, A., McArthur, C., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2020). Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *J. Acquir. Immune Defic. Syndr*. 83, 319–322. doi: 10.1097/QAI.0000000000002246

Yuan, M., Cottrell, C. A., Ozorowski, G., van Gils, M. J., Kumar, S., Wu, N. C., et al. (2019). Conformational plasticity in the HIV-1 fusion peptide facilitates recognition by broadly neutralizing antibodies. *Cell Host Microbe* 25, 873–883. doi: 10.1016/j.chom.2019.04.011

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 27 May 2022
doi: 10.3389/fmicb.2022.863084

# Viruses Previously Identified in Brazil as Belonging to HIV-1 CRF72_BF1 Represent Two Closely Related Circulating Recombinant Forms, One of Which, Designated CRF122_BF1, Is Also Circulating in Spain

OPEN ACCESS

**Edited by:**
Alfredo Garzino-Demo,
University of Maryland, Baltimore,
United States

**Reviewed by:**
Kenji Sadamasu,
Tokyo Metropolitan Institute of Public Health, Japan
Brian T. Foley,
Los Alamos National Laboratory (DOE), United States

***Correspondence:**
Michael M. Thomson
mthomson@isciii.es

**Specialty section:**
This article was submitted to Virology,
a section of the journal
Frontiers in Microbiology

**Received:** 26 January 2022
**Accepted:** 28 March 2022
**Published:** 27 May 2022

**Citation:**
Cañada-García JE, Delgado E, Gil H, Benito S, Sánchez M, Ocampo A, Cabrera JJ, Miralles C, García-Bodas E, Mariño A, Ordóñez P, Gude MJ, Ezpeleta C and Thomson MM (2022) Viruses Previously Identified in Brazil as Belonging to HIV-1 CRF72_BF1 Represent Two Closely Related Circulating Recombinant Forms, One of Which, Designated CRF122_BF1, Is Also Circulating in Spain. Front. Microbiol. 13:863084. doi: 10.3389/fmicb.2022.863084

***Javier E. Cañada-García[1], Elena Delgado[1], Horacio Gil[1], Sonia Benito[1], Mónica Sánchez[1], Antonio Ocampo[2], Jorge Julio Cabrera[3,4], Celia Miralles[2], Elena García-Bodas[1], Ana Mariño[5], Patricia Ordóñez[6], María José Gude[7], Carmen Ezpeleta[8] and Michael M. Thomson[1]****

*[1] HIV Biology and Variability Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Majadahonda, Spain, [2] Department of Internal Medicine, Complejo Hospitalario Universitario de Vigo, Vigo, Spain, [3] Department of Microbiology, Complejo Hospitalario Universitario de Vigo, Vigo, Spain, [4] Microbiology and Infectology Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain, [5] Infectious Diseases Unit, Complejo Hospitalario Universitario de Ferrol, Ferrol, Spain, [6] Department of Microbiology, Complejo Hospitalario Universitario de Ferrol, Ferrol, Spain, [7] Department of Microbiology, Hospital Universitario Lucus Augusti, Lugo, Spain, [8] Department of Clinical Microbiology, Complejo Hospitalario de Navarra, Pamplona, Spain*

Circulating recombinant forms (CRFs) are important components of the HIV-1 pandemic. Those derived from recombination between subtype B and subsubtype F1, with 18 reported, most of them of South American origin, are among the most diverse. In this study, we identified a HIV-1 BF1 recombinant cluster that is expanding in Spain, transmitted mainly *via* heterosexual contact, which, analyzed in near full-length genomes in four viruses, exhibited a coincident BF1 mosaic structure, with 12 breakpoints, that fully coincided with that of two viruses (10BR_MG003 and 10BR_MG005) from Brazil, previously classified as CRF72_BF1. The three remaining Brazilian viruses (10BR_MG002, 10BR_MG004, and 10BR_MG008) previously identified as CRF72_BF1 exhibited mosaic structures highly similar, but not identical, to that of the Spanish viruses and to 10BR_MG003 and 10BR_MG005, with discrepant subtypes in two short genome segments, located in *pol* and gp120$^{env}$. Based on these results, we propose that the five viruses from Brazil previously identified as CRF72_BF1 actually belong to two closely related CRFs, one comprising 10BR_MG002, 10BR_MG004, and 10BR_MG008, which keep their CRF72_BF1 designation, and the other, designated CRF122_BF1, comprising 10BR_MG003, 10BR_MG005, and the viruses of the identified Spanish cluster. Three other BF1 recombinant genomes, two from Brazil and one from Italy, previously identified as unique recombinant forms, were classified as CRF72_BF1. CRF122_BF1, but not CRF72_BF1, was associated with protease

L89M substitution, which was reported to contribute to antiretroviral drug resistance. Phylodynamic analyses estimate the emergence of CRF122_BF1 in Brazil around 1987. Given their close phylogenetic relationship and similar structures, the grouping of CRF72_BF1 and CRF122_BF1 in a CRF family is proposed.

**Keywords: HIV-1, circulating recombinant forms, molecular epidemiology, phylogeny, phylodynamics**

## INTRODUCTION

HIV-1 is characterized by high genetic diversity and rapid evolution, derived from elevated mutation and recombination rates. Through these mechanisms, the HIV-1 group M, the causative agent of the AIDS pandemic, has evolved into numerous circulating genetic forms, known as subtypes, of which 10 have been identified (A–D, F–H, J–L), subsubtypes (A1–A6, F1, and F2), and circulating recombinant forms (CRFs), 118 of which are currently listed in the Los Alamos HIV Sequence Database (Los Alamos National Laboratory, 2021). In addition, geographic variants and clusters, some representing substantial proportions of viruses in certain areas, have been identified through phylogenetic analyses within subtypes, subsubtypes, and CRFs (Thomson and Nájera, 2005; Delgado et al., 2015, 2019). Genetic characterization of HIV-1 variants is of public health relevance, as it allows tracking their geographic spread and estimating their population growth and the efficacy of preventive interventions (Paraskevis et al., 2016; Rife et al., 2017; Vasylyeva et al., 2020). It has also biological and clinical relevance, as different biological properties have been associated with some HIV-1 variants (Kiwanuka et al., 2008; Pérez-Álvarez et al., 2014; Kouri et al., 2015; Venner et al., 2016; Cid-Silva et al., 2018; Song et al., 2019).

The number of CRFs is increasing incessantly, due to both the continuous generation of recombinant forms where diverse HIV-1 variants meet in the same population (Nájera et al., 2002), some of which become circulating through introduction into transmission networks, and the identification of old previously undocumented CRFs. The proportion of CRFs in the HIV-1 pandemic has increased over time, representing around 17% of infections in 2010–2015 (Hemelaar et al., 2020). Among CRFs, those derived from subtype B and subsubtype F1 are among the most numerous, 18 of which have been reported in the literature, most of them originated in South America. The most widely circulating CRF from South America is CRF12_BF, which circulates at high prevalences in Argentina and Uruguay, where unique recombinant forms (URFs) related to CRF12_BF are frequently found (Thomson et al., 2000, 2002; Carr et al., 2001). Four other CRF_BFs (numbers 17, 38, 44, and 89) related to CRF12_BF, as evidenced by shared breakpoints and phylogenetic clustering, were subsequently identified in different South American countries (Ruchansky et al., 2009; Delgado et al., 2010, 2021; Aulicino et al., 2012). Due to their common ancestry and similar structures, these five CRFs and related URFs have been proposed to constitute a "family" of recombinant viruses (Thomson and Nájera, 2005; Zhang et al., 2010; Delgado et al., 2021). By contrast, Brazilian CRF_BFs (De Sá Filho et al., 2006; Guimarães et al., 2008; Sanabani et al., 2010; Pessôa et al., 2014a,b, 2016; Reis et al., 2017, 2019) and CRF66_BF (the latter found mainly in Paraguay and Paraguayans living in Spain) (Bacqué et al., 2021) are unrelated to CRF12_BF. Similarly to the viruses of the CRF12_BF family, close relations have been reported between some Brazilian CRF_BFs: CRF28_BF and CRF29_BF (De Sá Filho et al., 2006) and CRF70_BF and CRF71_BF (Pessôa et al., 2014a).

In this study, we report the spread of a BF1 cluster in Spain whose viruses exhibit a mosaic structure identical to two Brazilian viruses previously identified as CRF72_BF1 (Pessôa et al., 2014b, 2016), which would represent a new CRF, with the three other viruses classified as CRF72_BF1 showing highly similar, but not identical, structures. We propose that viruses previously identified as CRF72_BF1 actually belong to two closely related CRFs that constitute a CRF family.

## MATERIALS AND METHODS

### Samples

Plasma and whole blood samples from HIV-1-infected individuals were collected in Spain for antiretroviral drug resistance tests and for a molecular epidemiological study. The study was approved by the Committee of Research Ethics of Instituto de Salud Carlos III, Majadahonda, Madrid, Spain. It did not require written informed consent by the study participants, as it used samples and data collected as part of routine clinical practice, and patients' data were anonymized without retaining data allowing individual identification.

### PCR Amplification and Sequencing

An ∼1.4-kb pol fragment in protease–reverse transcriptase (Pr–RT) was amplified from plasma-extracted RNA or from whole blood-extracted DNA by (RT-)PCR followed by nested PCR, as described previously (Delgado et al., 2015), and sequenced with the Sanger method using a capillary automated sequencer. Near full-length genome (NFLG) sequences were obtained for selected samples by RT-PCR/nested PCR amplification from plasma RNA in five overlapping segments and sequenced by the Sanger method, as described (Delgado et al., 2002; Sierra et al., 2005; Cañada et al., 2021). Newly derived sequences are deposited in GenBank under accessions OL982311–OL982317 and OL982320–OL982323.

### Phylogenetic Sequence Analyses

Sequences were aligned with MAFFT v7 (Katoh and Standley, 2013). Initial phylogenetic trees with all Pr–RT sequences obtained by us were constructed *via* approximate maximum likelihood with FastTree v2.1.10 (Price et al., 2010), using the general time-reversible evolutionary model with CAT approximation to account for among-site rate heterogeneity, with assessment of node support with Shimodaira–Hasegawa (SH)-like local support values (Guindon et al., 2010). Subsequent maximum likelihood (ML) trees with sequences of interest were constructed with W-IQ-Tree (Trifinopoulos et al., 2016), using the best-fit substitution model selected by the ModelFinder program (Kalyaanamoorthy et al., 2017), with assessment of node support with the ultrafast bootstrap approximation approach (Hoang et al., 2018). Trees were visualized with MEGA v7.0 (Kumar et al., 2016).

Mosaic structures were analyzed by bootscanning (Salminen et al., 1995) with SimPlot v1.3.5 (Lole et al., 1999). In these analyses, trees were constructed using the neighbor-joining method with the Kimura two-parameter model and a window width of 250 nucleotides. The subtype affiliations of recombinant segments identified with SimPlot, whose breakpoints were more precisely located in the midpoint of transitions between BF1 subtype-discriminating nucleotides (here defined as those differing between the 75% consensus sequences of subtype B and the Brazilian F1 strain), were further phylogenetically analyzed *via* ML and Bayesian inference. These analyses were performed with IQ-Tree; PhyML (Guindon et al., 2010), using the best-fit evolutionary model selected by the SMS program (Lefort et al., 2017) and node support assessment with the approximate likelihood ratio test, SH–like procedure; and MrBayes v3.2 (Ronquist et al., 2012), using the GTR + G + I substitution model, with two simultaneous independent runs and eight chains 2–5 million generations long, ensuring that both runs reached convergence, as determined by an average standard deviation of split frequencies $<0.01$, discarding the first 50% of the trees in the posterior distribution as burn-in. For these analyses, we used a reconstructed BF1 ancestral sequence as outgroup, considering the phylogenetic relationship between B and F subtypes (Zhu et al., 1998), obtained with IQ-Tree. The use of a reconstructed ancestral sequence as outgroup is similar to the approach used in other studies (Travers et al., 2004; Thomson and Fernández-García, 2011; Seager et al., 2014) to prevent the long-branch attraction artifact (Bergsten, 2005) that could be caused by an outgroup whose distance to the ingroup is relatively long compared with the within-ingroup distances. This artifact can result in collapse or a substantial decrease in node support of the clades of the ingroup, particularly in short genome segments.

Phylogenetic trees and alignments used for their construction have been deposited in TreeBase, with accession URL http://purl.org/phylo/treebase/phylows/study/TB2:S29595.

## Antiretroviral Drug Resistance Determination

Antiretroviral (ARV) drug resistance was analyzed with the HIVdb program of the Stanford University's HIV Drug Resistance Database (Rhee et al., 2003; Shafer, 2006).

## Temporal and Geographic Estimations of Clade Ancestors

The time and the most probable location of the most recent common ancestor (MRCA) of the newly defined CRF were estimated using Pr–RT sequences with the Bayesian Markov chain Monte Carlo (MCMC) coalescent method implemented in BEAST v1.10.4 (Suchard et al., 2018). Before the BEAST analysis, the existence of temporal signal in the dataset was assessed with TempEst v1.5.3 (Rambaut et al., 2016). The BEAST analysis was performed using the SRD06 codon-based evolutionary model (where the third codon position is in a partition different from the first and second positions) (Shapiro et al., 2006). We also specified an uncorrelated lognormal relaxed clock and a Bayesian SkyGrid coalescent tree prior (Gill et al., 2013). The MCMCs were run for 20 million generations. The runs were performed in duplicate, and the posterior tree files were combined with LogCombiner v1.10.4. Proper mixing of the chains was assessed with Tracer v1.6, ensuring that effective sample size values of all parameters were $>200$. The posterior distribution of trees was summarized in a maximum clade credibility (MCC) tree with TreeAnnotator v1.10.4, after discarding 10% of the MCMC chain as burn-in. MCC trees were visualized with FigTree v1.4.2 (Rambaut).[1] Parameter uncertainty was summarized in 95% highest posterior density (HPD) intervals.

# RESULTS

## Identification of a HIV-1 Cluster of F1 Subsubtype in Protease–Reverse Transcriptase Propagating in Spain

In a molecular epidemiological study in Spain, based on Pr–RT sequences, we detected frequent grouping in clusters, several of which were of F1 subsubtype in Pr–RT (Thomson et al., 2012; Delgado et al., 2015, 2019; Gil et al., 2022). One of them, designated F1_2, which is the focus of the present study, comprised 14 individuals, 13 of them from the region of Galicia, northwest Spain (**Table 1**). Years of HIV-1 diagnoses were between 2007 and 2019, and transmission was predominantly heterosexual ($n$ = 8), but there were three men who have sex with men (two others had non-specified sexual transmission, and no data on transmission route were available for another individual) (**Table 1**). Most individuals were Spanish, but three were Brazilian, one was Swiss, and one was Ukrainian. To determine whether other sequences from databases belonged to this cluster, we performed BLAST searches in the HIV Sequence Database (Los Alamos National Laboratory, 2021), incorporating the most similar sequences in the phylogenetic analyses. This allowed identifying three additional sequences that belonged to the F1_2 cluster, from Brazil, Portugal, and Germany (**Figure 1**). All but two of the viruses collected in Spain and the virus from Germany branched in a subcluster. Viruses from the F1_2 cluster were most closely related to viruses of the Brazilian F1 strain and to Brazilian CRF_BFs with Pr–RT derived from it.

## Analyses of Near Full-Length Genome Sequences

In order to determine whether the F1_2 cluster was of uniform subtype or recombinant, we obtained NFLG sequences from

[1] http://tree.bio.ed.ac.uk/software/figtree/

**TABLE 1 |** Epidemiological data of patients and GenBank accessions of sequences.

| Sample ID | City of sample collection | Region of sample collection | Country of origin | Year of HIV diagnosis | Year of sample collection | Gender* | Transmission route* | PR–RT GenBank accession | NFLG GenBank accession |
|---|---|---|---|---|---|---|---|---|---|
| X2592 | Vigo | Galicia | Spain | 2008 | 2008 | F | HT | GU326146 | – |
| X2632 | Ferrol | Galicia | Spain | 2009 | 2009 | M | MSM | GU326158 | KC113006<br>JX140660 |
| X2657 | Vigo | Galicia | Spain | 2009 | 2009 | F | HT | GU326163 | – |
| GA076319 | Vigo | Galicia | Ukraine | 2010 | 2010 | F | HT | OL982311 | – |
| GA099170 | Vigo | Galicia | Switzerland | 2011 | 2011 | F | HT | – | OL982312 |
| GA330265 | Vigo | Galicia | Brazil | 2007 | 2007 | Trans | Sexual | OL982313 | – |
| GA486085 | Ferrol | Galicia | Spain | 2018 | 2018 | F | n.a. | – | OL982314 |
| GA501952 | Vigo | Galicia | Spain | 2014 | 2014 | F | HT | OL982315 | – |
| GA513250 | Vigo | Galicia | Spain | 2012 | 2012 | M | Sexual | OL982316 | – |
| GA522821 | Lugo | Galicia | Brazil | 2019 | 2019 | M | MSM | OL982317 | – |
| GA817166 | Vigo | Galicia | Spain | 2008 | 2016 | F | HT | OL982320 | – |
| GA874035 | Vigo | Galicia | Spain | 2012 | 2012 | M | HT | – | OL982321 |
| GA903064 | Vigo | Galicia | Spain | 2012 | 2012 | F | HT | OL982322 | – |
| NA584314 | Pamplona | Navarre | Brazil | 2017 | 2017 | M | MSM | OL982323 | – |

**n.a., datum not available; Trans, transgender; HT, heterosexual; MSM, man who has sex with men; sexual, unspecified sexual transmission.*

three individuals from two cities through amplification from plasma RNA. A fourth NFLG sequence had been obtained previously from the virus culture supernatant (Sanchez et al., 2014). Preliminary analyses of the NFLG with Recombination Identification Program[2] indicated that the genomes were BF1 recombinant. To determine whether they belonged to a known CRF, we constructed a phylogenetic tree in which genomes of all CRF_BFs were included. The tree showed that viruses of the F1_2 cluster grouped in a strongly supported clade with viruses classified as CRF72_BF1, with two of them, 10BR_MG003 and 10BR_MG005, being the most closely related to the viruses of the F1_2 cluster, and the other three, 10BR_MG002, 10BR_MG004, and 10BR_MG008, branching in a sister clade (**Figure 2**).

Bootscan analyses of NFLG sequences showed that the viruses of the F1_2 cluster were BF1 recombinant, exhibiting mosaic structures fully coincident with those of 10BR_MG003 and 10BR_MG005, and slightly different from the three other viruses classified as CRF72_BF1 (**Figure 3** and **Supplementary Figure 1**). The differences between these three viruses were observed in a short *pol* segment, around the protease–reverse transcriptase junction, where grouping with subtype references was discrepant, and in the 5′ segment of gp120, where the location of a BF1 breakpoint differed. The mosaic structures determined with bootscanning were confirmed by ML and Bayesian phylogenetic analyses of partial genome segments, which confirmed the coincidence of the mosaic structures of the four F1_2 viruses and the Brazilian 10BR_MG003 and 10BR_MG005 viruses and the subtype discrepancy in two genome segments (HXB2 positions 2429–2618 and 6432–6519) of these viruses with 10BR_MG002, 10BR_MG004, and 10BR_MG008 (**Figure 4**). These analyses, therefore, allowed determining that viruses of the identified Spanish BF1 cluster, together with the Brazilian viruses 10BR_MG003 and 10BR_MG005, previously classified as CRF72_BF1, belong to a CRF, which was designated CRF122_BF1, which is closely related to, but different from, the three other viruses previously classified as CRF72_BF1, 10BR_MG002, 10BR_MG004, and 10BR_MG008, whose original CRF designation is maintained. The mosaic structures of both CRFs, as inferred from bootscan analyses, ML and Bayesian phylogenetic trees of partial sequences, and examination of intersubtype transitions of subtype-discriminating nucleotides, are shown in **Figure 5**.

Three additional BF1 recombinant NFLGs, originally identified as unique recombinant forms, two from Brazil, 99UFRJ-2 (Thomson et al., 2004) and BREPM1029 (Sa-Filho et al., 2007), and one from Italy, IT_BF_PRIN_454 (Bruselles et al., 2009), in their published analyses, exhibited mosaic structures similar to CRF72_BF1 and CRF122_BF1. To determine whether they belonged to one of these CRFs, we constructed a phylogenetic tree with NFLG sequences including the three mentioned genomes, which showed that all of them grouped with CRF72_BF1 viruses (**Supplementary Figure 2**). Bootscan analyses showed mosaic structures of 99UFRJ-2 and IT_BF_PRIN_454 coincident with that of CRF72_BF1; however, the bootscan plot of BREPM1029 failed to show clustering with the subtype B references in the protease–RT junction (HXB2 positions 2429–2618) (**Supplementary Figure 3**). Examination of subtype-discriminating nucleotides suggested that the 2429–2618 segment was of subtype B, as in CRF72_BF1, in 99UFRJ-2 and IT_BF_PRIN_454, which was confirmed by phylogenetic analyses of this fragment (**Supplementary Figure 4a**). However, in BREPM1029, the subtype B fragment in the protease–RT junction appeared to be slightly shorter, located between HXB2 positions 2479 and 2618, which was confirmed by phylogenetic trees (**Supplementary Figure 4b**). Phylogenetic analyses also showed that in all three genomes the 6432–6519 segment in gp120 was of subtype B, as in CRF72_BF1 and unlike CRF122_BF1 (**Supplementary Figure 4c**). These results allowed to confidently classify 99UFRJ-2 and IT_BF_PRIN_454 as CRF72_BF1 viruses. As to BREPM1029, given its strong

[2] https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html

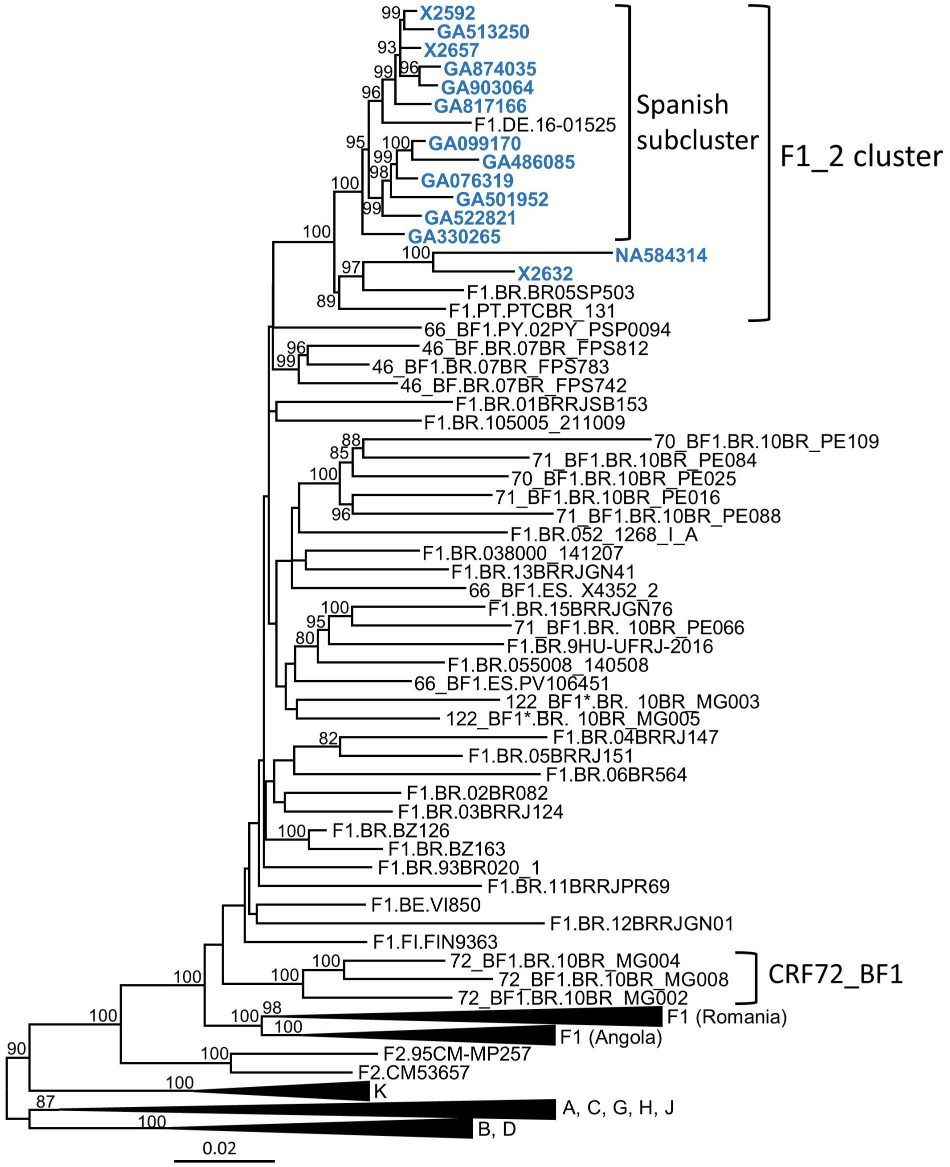


**FIGURE 1 |** Maximum likelihood phylogenetic tree of Pr–RT sequences of the F1_2 cluster. Names of sequences obtained by us, all collected in Spain, are in blue. Only ultrafast bootstrap values ≥ 80% are shown. In database sequences, the country of sample collection is indicated before the virus name with the two-letter ISO country code: BE, Belgium; BR, Brazil, DE, Germany; ES, Spain; FI, Finland; PT, Portugal; PY, Paraguay. The scale indicates substitutions/site. *10BR_MG003 and 10BR_MG005 were originally identified as CRF72_BF1 (Pessôa et al., 2014b, 2016), but analyses described in this study have reclassified them as CRF122_BF1.

phylogenetic clustering with CRF72_BF1 references and its minimal difference in mosaic structure with CRF72_BF1, with a breakpoint displaced only around 50 nt relative to this CRF, it seems reasonable to also classify it as CRF72_BF1, although we cannot definitively discern whether its breakpoint displacement is due to a different recombination event or to mutations occurring near the CRF72_BF1 breakpoint.

## Differences in Amino Acid Residues

We analyzed amino acid residues in viral proteins differing between CRF72_BF1 and CRF122_BF1 viruses and conserved within each CRF. We found 10 such amino acid residues (**Table 2**). One of them is in position 89 of protease, where CRF72_BF1 has leucine, which is the subtype B consensus, while CRF122_BF1 has methionine, which is the F1 subsubtype consensus. Protease L89M substitution has been reported to contribute, together with other protease mutations, to resistance to some protease inhibitor drugs (Calazans et al., 2005; Marcelin et al., 2008; Wensing et al., 2019).

## Antiretroviral Drug Resistance Mutations

Primary ARV drug resistance mutations were detected in two CRF122_BF1 viruses, both from Brazil, one (10BR_MG003, collected in 2010) with L90M protease inhibitor (PI) resistance

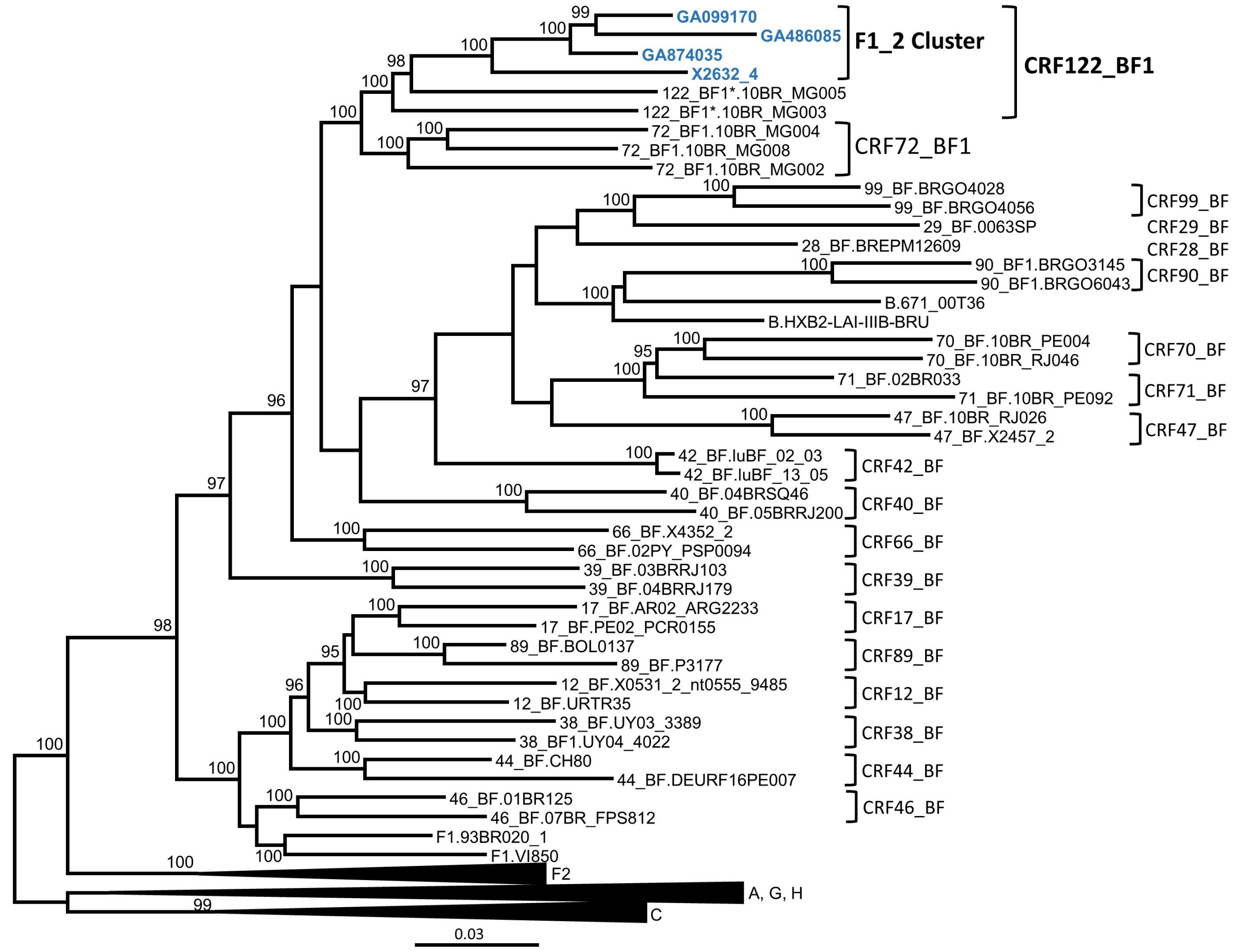


**FIGURE 2 |** Maximum likelihood tree of NFLG sequences of viruses of the F1_2 cluster. References of published CRF_BFs and of HIV-1 subtypes are also included in the analysis. Names of sequences obtained by us are in blue. In reference sequences, the subtype or CRF is indicated before the virus name. Only ultrafast bootstrap values ≥ 90% are shown. The scale indicates substitutions/site. *10BR_MG003 and 10BR_MG005 were originally identified as CRF72_BF1 (Pessôa et al., 2014b, 2016), but analyses described in this study have reclassified them as CRF122_BF1.

mutation and the other (BR05SP503, collected in 2005) with D30N PI resistance mutation and M41L, D67N, M184V, and T215Y mutations of resistance to nucleoside RT inhibitors.

## Temporal and Geographic Estimation of CRF122_BF1 Origin

To estimate the time and place of origin of CRF122_BF1, Pr–RT sequences were analyzed with the Bayesian coalescent method implemented in BEAST 1.10.4. Prior to this analysis, we performed TempEst analyses to determine whether there was an adequate temporal signal in the dataset. We found that the temporal signal, assessed by the correlation between root-to-tip distance and time, increased by masking the positions of codons with drug resistance mutations in any of the sequences ($r^2$ = 0.5265; **Supplementary Figure 5**). Therefore, the BEAST analysis was performed with a sequence alignment where these codons had been removed. In this analysis, the substitution rate was estimated at 1.829 × $10^{-3}$ subs/site/year (95% HPD, 1.118 × $10^{-3}$–2.542 × $10^{-3}$ subs/site/year) and the time of the MRCA of CRF122_BF1 was estimated around 1987 (95% HPD, 1976–1998), with its most probable location being Brazil (location PP = 0.89) (**Figure 6**). The introduction of CRF122_BF1 in Spain (according to the MRCA of the Spanish cluster) was estimated in the Galician city of Vigo (location PP = 0.992) around 2002 (95% HPD, 1998–2005).

## DISCUSSION

The results presented here indicate that the five Brazilian viruses previously classified as CRF72_BF1 actually belong to two closely related CRFs, one of which is circulating in Spain. Consequently, the CRF comprising two Brazilian viruses previously classified as CRF72_BF1 and the four Spanish viruses with coincident mosaic structures is given a new designation, CRF122_BF1, while the

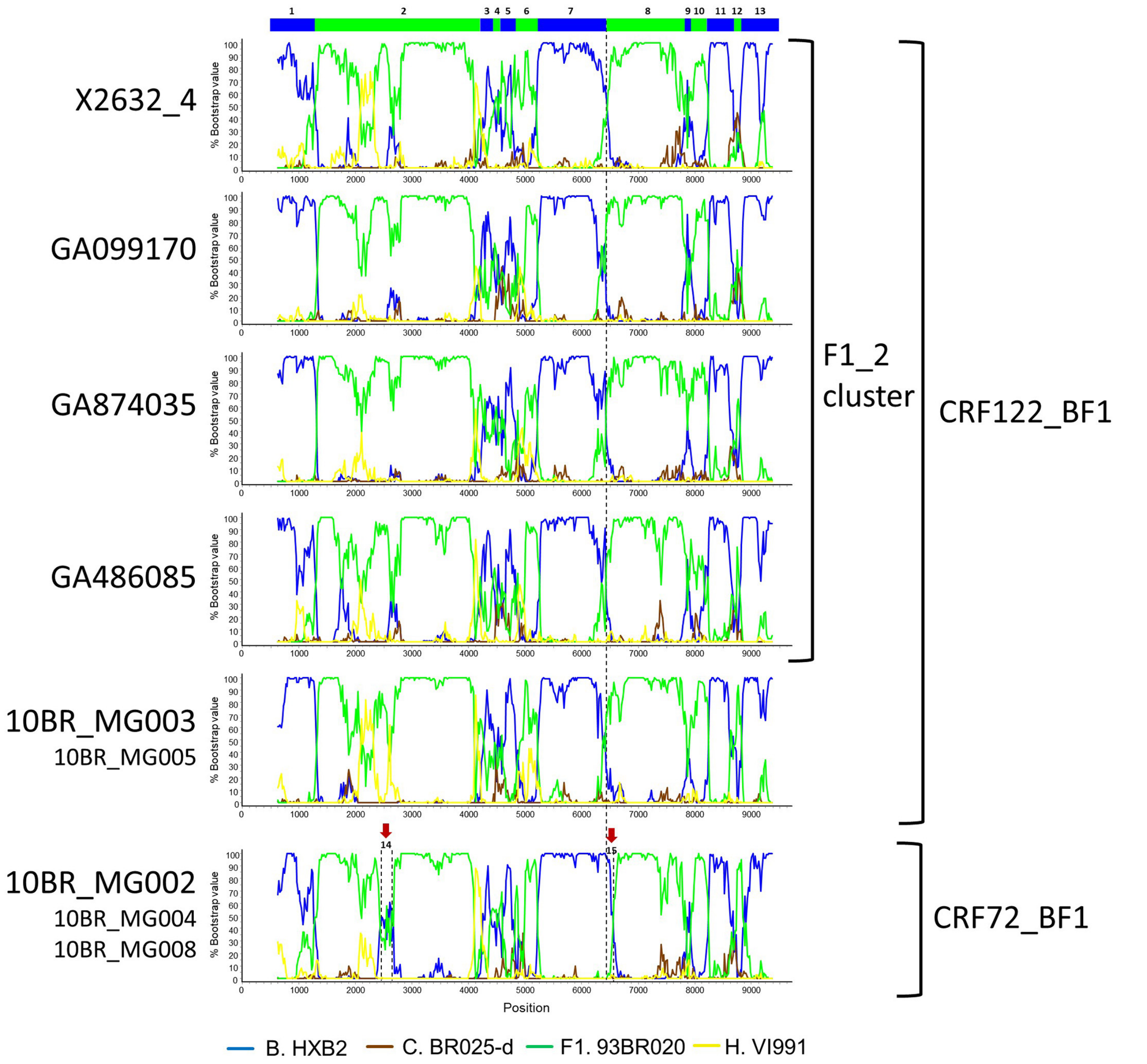


**FIGURE 3 |** Bootscan analyses of NFLG sequences of viruses of the F1_2 cluster compared to those of viruses previously classified as CRF72_BF1. Bootscan plots of all four F1_2 viruses are shown, together with those of 10BR_MG003 and 10BR_MG002. The bootscan plot of 10BR_MG005 is almost identical to that of 10BR_MG003, and those of 10BR_MG004 and 10BR_MG008 are almost identical to that of 10BR_MG002, and are shown in **Supplementary Figure 1**. The horizontal axis represents the position in the HXB2 genome of the midpoint of a 250-nt window moving in 20-nt increments, and the vertical axis represents bootstrap values supporting clustering with subtype reference sequences. The vertical dashed lines indicate BF1 breakpoints differing between the CRF72_BF1 viruses 10BR_MG002, 10BR_MG004, and 10BR_MG008, on the one hand, and viruses of the F1_2 cluster and 10BR_MG003 and 10BR_MG005 (newly identified as CRF122_BF1), on the other. The bar on the top indicates the segments that were further analyzed with ML and Bayesian trees (**Figure 4**). Two genome segments that appear to group with different subtypes in 10BR_MG002, 10BR_MG004, and 10BR_MG008 relative to 10BR_MG003, 10BR_MG005, and the F1_2 viruses are signaled with arrows above the bootscan plot of 10BR_MG002 and were also analyzed *via* ML and Bayesian inference (**Figure 4**).

three other Brazilian viruses previously classified as CRF72_BF1 keep their original designation. Three other BF1 viruses analyzed in NFLGs originally classified as URFs, two from Brazil and one from Italy, were also classified on the basis of phylogenetic and bootscan analyses as CRF72_BF1. The close relationship between CRF122_BF1 and CRF72_BF1 is one more example of closely related CRFs, with precedents in South America. Other examples are CRF12_BF, CRF17_BF, and CRF89_BF (and more distantly related, CRF38_BF and CRF44_BF) (Delgado et al., 2021); CRF28 and CRF29_BF (De Sá Filho et al., 2006); and CRF70_BF and CRF71_BF (Pessôa et al., 2014a).

Failure to realize that the five viruses previously identified as CRF72_BF1 represent two different CRFs may derive from the short segments in which both CRFs differ in subtypes. These

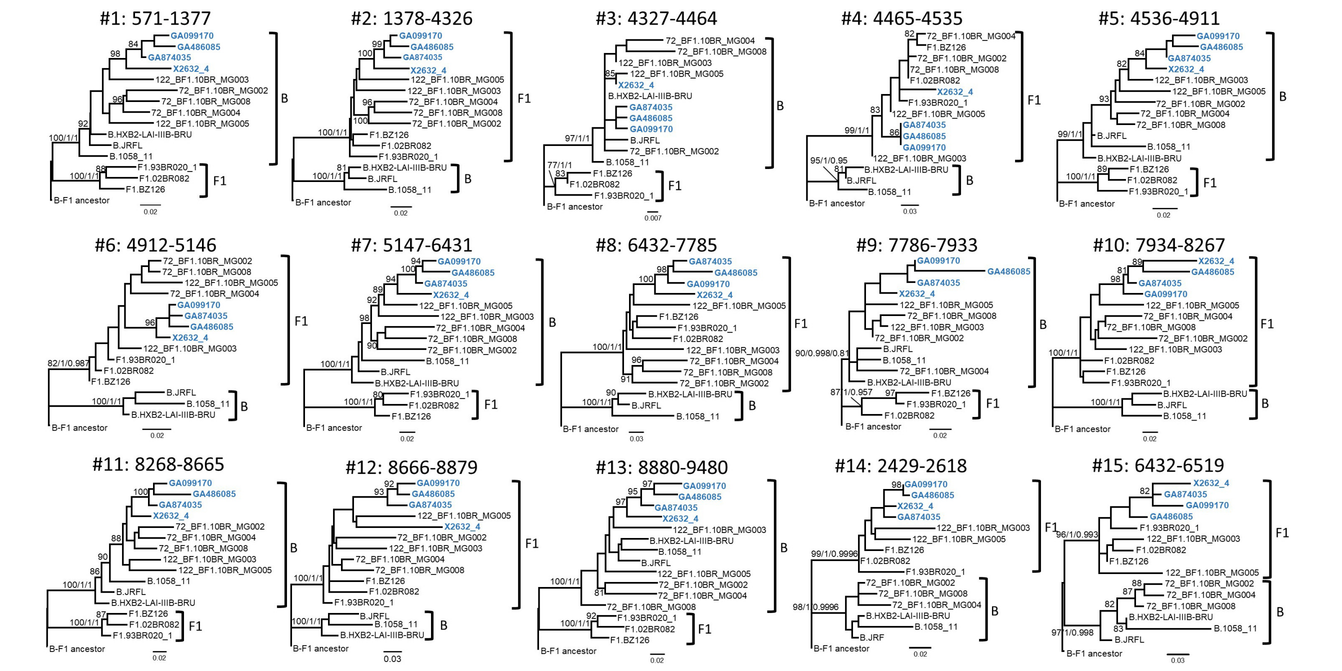


**FIGURE 4 |** Phylogenetic trees of interbreakpoint genome segments of F1_2 viruses. Breakpoints were defined according to the bootscan analyses and to midpoints of transitions between subtype-discriminating nucleotides, here defined as those where the 75% consensus of subtype B and of the Brazilian variant of subsubtype F1 differ. HXB2 positions delimiting the analyzed segments and their numbers as indicated in **Figure 3** are indicated on top of the trees. Sequence names of F1_2 viruses are in blue. Names of subtype reference sequences are preceded by the corresponding subtype name. Sequences of viruses previously classified as CRF72_BF1 were also included, with those reclassified in the present study as CRF122_BF1 (10BR_MG003 and 10BR_MG005) labeled with the new CRF designation. The trees are rooted with a reconstructed BF1 ancestor. Node supports for B and F1 clades are indicated, in this order, as ultrafast bootstrap value/aLRT SH–like support/posterior probability, which were obtained with IQ-Tree, PhyML, and MrBayes programs, respectively. For the other nodes, only ultrafast bootstrap values ≥ 80% are indicated. Trees #14 and #15 (segments 2429–2618 and 6432–6519, respectively) correspond to the segments indicated with arrows in **Figure 3**, where F1_2 viruses and 10BR_MG003 and 10BR_MG005, on the one hand, and 10BR_MG002, 10BR_MG004, and 10BR_MG008, on the other, appeared to differ in subtype affiliations in the bootscan analyses. The scales indicate substitutions/site.

differences may be missed if bootscan analyses are performed using window widths much greater than the length of the recombinant fragment. We have also noticed that jpHMM (Schultz et al., 2009), used in a previous study to analyze CRF72_BF1 genomes (Pessôa et al., 2016), often fails to detect short recombinant segments (Delgado et al., 2021).

Given the close relationship and partial coincidence in mosaic structures of CRF72_BF1 and CRF122_BF1, we propose that they are members of a CRF family, similar to the CRF family of BF1 recombinant viruses from South America comprising CRFs numbers 12, 17, 38, 44, and 89 (Delgado et al., 2021). The grouping of some closely related HIV-1 recombinants derived from a common recombinant ancestor in families was proposed by Thomson and Nájera (2005) and Zhang et al. (2010). The proposed families of CRFs with close phylogenetic relations, shared parental strains, and partially coincident breakpoints are indicated in the phylogenetic tree shown in **Supplementary Figure 6**.

It is interesting to note that, in the Pr–RT tree, viruses from the Spanish CRF122_BF1 (“F1_2”) cluster fail to group with the Brazilian CRF122_BF1 viruses. A similar phenomenon is observed with CRF66, CRF70, and CRF71_BF1 references, that fail to group in distinct clades with other references of the same CRF. This may be due to the relatively short length and high sequence conservation of this segment, together with the fact that 35 references from the Brazilian F1 strain or from CRF_BFs derived from it are included in the tree. This shows that exclusive phylogenetic analysis of Pr–RT may not be sufficient to phylogenetically classify an F1 sequence of the Brazilian strain as belonging or not to a given CRF_BF.

The estimated origin of CRF122_BF1 around 1987 is consistent with the estimated origin of the Brazilian F1 strain (around 1977) (Bello et al., 2007) and similar to those of other South American CRF_BFs (CRF12, CRF28/29, CRF38, CRF89, and CRF90) reported in the literature (Bello et al., 2010; Ristic et al., 2011; Reis et al., 2017; Delgado et al., 2021) but younger than some other estimates for CRF12_BF in the 1970s (Dilernia et al., 2011; Delgado et al., 2021) and older than the estimates for CRF99_BF, around 1993 (Reis et al., 2017).

The correct classification of HIV-1 genetic forms is important, since even relatively minor genetic differences in viral genomes may result in important biological differences. Examples in HIV-1 are frequent CXCR4 coreceptor usage in CRF14_BG, which is associated with only four amino acid residues in the Env V3 loop (Pérez-Álvarez et al., 2014), all or most of which are absent in viruses of the closely related CRF73_BG

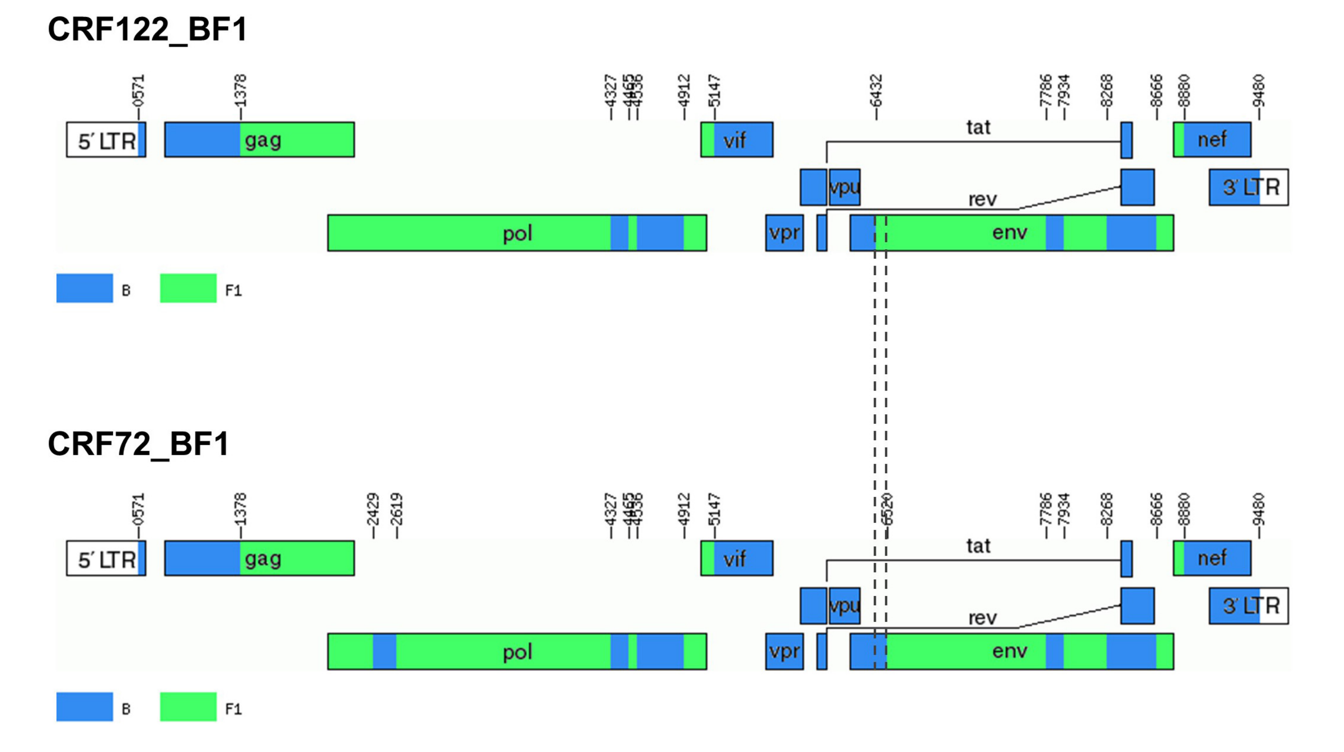


**FIGURE 5 |** Mosaic structures of CRF72_BF1 and CRF122_BF1. Breakpoint positions are numbered as in the HXB2 genome. Vertical dashed lines indicate the BF1 breakpoint positions in the 5′ segment of gp120$^{env}$ differing between CRF122_BF1 and CRF72_BF1. The drawing was made using the Recombinant HIV-1 Drawing Tool (https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html).

**TABLE 2 |** Differences in amino acid residues between CRF72_BF1 and CRF122_BF1*.

| | | p17$^{gag}$ | PR | | RT | IN | Vif | Tat | Rev | Vpu | gp120$^{env}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **61** | **61** | **89** | **399** | **84** | **151** | **68** | **22** | **80** | **69** |
| B | HXB2 | L | Q | L | E | I | A | S | L | D | D |
| CRF72_BF | 10BR_MG002 | I | Q | L | E | L | V | S | L | D | D |
| | 10BR_MG004 | I | Q | L | E | L | A | S | L | D | D |
| | 10BR_MG008 | I | Q | L | D | L | A | S | L | D | D |
| | 99UFRJ-2 | I/M | Q | L | E | L | A | P | L | D | D |
| | BREPM1029 | L | N | L | E | L | A | S | L | D | D |
| | IT_PRIN_454 | I | Q | L | E | L | A | S | L | D | D |
| CRF122_BF1 | 10BR_MG003 | L | N | M | D | I | T | D | I | N | N |
| | 10BR_MG005 | L | N | M | D | I | T | D | I | N | N |
| | X2632_4 | L | N | M | D | L | T | D | I | N | N |
| | GA099170 | L | N | M | D | I | T | D | I | N | N |
| | GA486085 | L/I | N | M | D | I/M | T | D | I | N | D |
| | GA874035 | L | N | M | D | I | T | D | I | N | N |

**Only amino acid residues conserved in at least five of six CRF72_BF1 viruses and in both Brazilian and at least three of four Spanish CRF122_BF1 viruses are included in the table.*

(Fernández-García et al., 2016), which has a very similar mosaic structure, and differences in pathogenic potential or therapeutic response associated with clusters within HIV-1 CRF01_AE (Song et al., 2019) and F1 subsubtype (Cid-Silva et al., 2018). Here, we show that CRF122_BF1, but not CRF72_BF1, has the protease L89M substitution, that has been reported to contribute, together with other protease mutations, to resistance to tipranavir/ritonavir (Marcelin et al., 2008) and, within an F1 subsubtype background, to other protease inhibitor drugs (Calazans et al., 2005).

CRF122_BF1 represents one more example of a CRF of South American ancestry first identified in Western Europe. Others are CRF42_BF (Struck et al., 2015), CRF47_BF (Fernández-García et al., 2010), CRF60_BC (Simonetti et al., 2014), CRF66_BF (Bacqué et al., 2021), and CRF89_BF (Delgado et al., 2021). This may derive from the increasing migratory

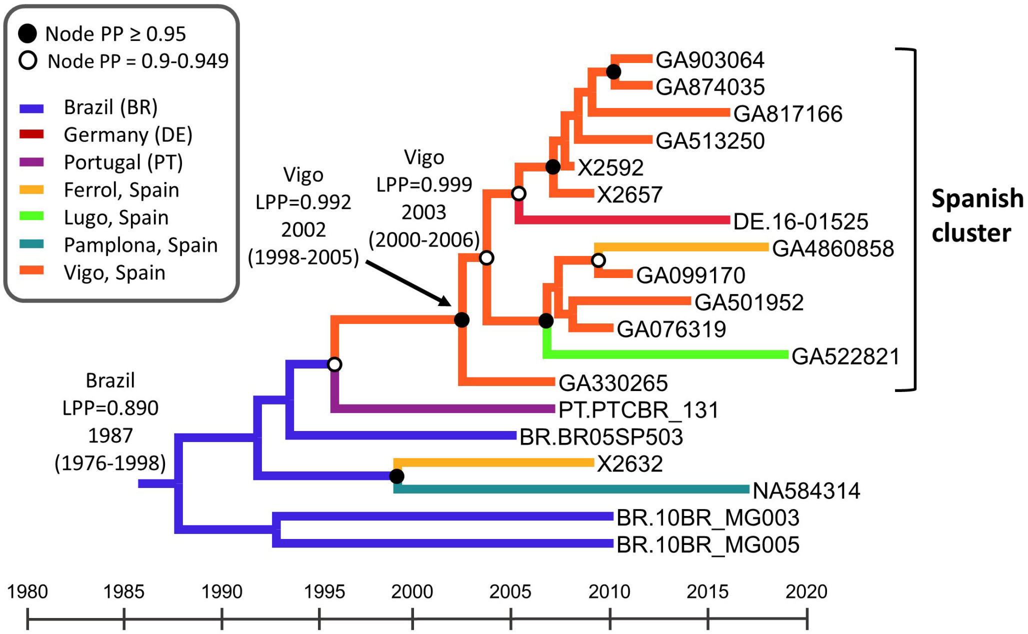


**FIGURE 6 |** Maximum clade credibility tree of CRF122_BF Pr–RT sequences. Branch colors indicate, for terminal branches, the place of sample collection and, for internal branches, the most probable location of the subtending node, according to the legend on the upper left. Nodes supported by PP ≥ 0.95 and PP 0.9–0.949 are indicated with filled and unfilled circles, respectively. The most probable locations at the root of the tree and at the node of the Spanish cluster are indicated, together with the PPs supporting each location (LPPs) and the year estimated for the MRCAs (mean value, with 95% HPD interval in parentheses). The scale under the tree represents calendar years.

flows from South America to Europe and from the relatively low number of HIV-1 sequences available in some South American countries (Bacqué et al., 2021). Therefore, HIV-1 molecular epidemiological studies in Europe may contribute to a better knowledge of the HIV-1 epidemics in South America.

In summary, we show that viruses of a BF1 recombinant cluster of Brazilian ancestry circulating in Spain exhibit a mosaic structure that is fully coincident with that of two Brazilian viruses previously classified as CRF72_BF1 and is highly similar, but not identical, to that of three other Brazilian viruses also classified as CRF72_BF1. Therefore, we propose a new CRF designation, CRF122_BF1, for the viruses of the Spanish cluster and the two Brazilian viruses with coincident structures, which together with CRF72_BF1 would constitute a CRF family. The accurate genetic characterization of HIV-1 variants is important to determine their associated biological features and to track their epidemic spread.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. Newly derived sequences are deposited in GenBank under accessions OL982311–OL982317 and OL982320–OL982323. Phylogenetic trees and alignments used for their construction have been deposited in TreeBase, with accession URL http://purl.org/phylo/treebase/phylows/study/TB2:S29595.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Committee of Research Ethics of Instituto de Salud Carlos III, Majadahonda, Madrid, Spain. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

MT and ED conceived the study and supervised the experimental work. JC-G, ED, and MT processed the sequences and performed phylogenetic and phylodynamic analyses. HG performed data curation. JC-G, SB, MS, and EG-B performed experimental work. AO, JC, CM, AM, PO, MG, and CE obtained the samples and epidemiological data from patients. MT, JC-G, ED, and HG wrote the manuscript. All authors read and approved the manuscript.

## FUNDING

This work was funded through Acción Estratégica en Salud Intramural (AESI), Instituto de Salud Carlos III, projects "Estudios Sobre Vigilancia Epidemiológica Molecular del VIH-1 en España", PI16CIII/00033, and "Epidemiología Molecular del VIH-1 en España y su Utilidad Para Investigaciones

Biológicas y en Vacunas", PI19CIII/00042, and through scientific agreement with Consellería de Sanidade, Government of Galicia (MVI 1004/16).

## ACKNOWLEDGMENTS

We would thank José Antonio Taboada from Consellería de Sanidade, Xunta de Galicia, for his support of this study, and the personnel at the Genomic Unit, Instituto de Salud Carlos III, for technical assistance in sequencing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.863084/full#supplementary-material

## REFERENCES

Aulicino, P. C., Gómez-Carrillo, M., Bello, G., Rocco, C., Mangano, A., Carr, J., et al. (2012). Characterization of full-length HIV-1 CRF17_BF genomes and comparison to the prototype CRF12_BF strains. *Infect. Genet. Evol.* 12, 443–447. doi: 10.1016/j.meegid.2012.01.003

Bacqué, J., Delgado, E., Benito, S., Moreno-Lorenzo, M., Montero, V., Gil, H., et al. (2021). Identification of CRF66_BF, a new HIV-1 circulating recombinant form of South American origin. *Front. Microbiol.* 12:774386. doi: 10.3389/fmicb.2021.774386

Bello, G., Aulicino, P. C., Ruchansky, D., Guimarães, M. L., López-Galíndez, C., Casado, C., et al. (2010). Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. *Retrovirology* 7:22. doi: 10.1186/1742-4690-7-22

Bello, G., Eyer-Silva, W. A., Couto-Fernandez, J. C., Guimarães, M. L., Chequer-Fernandez, S. L., Teixeira, S. L., et al. (2007). Demographic history of HIV-1 subtypes B and F in Brazil. *Infect. Genet. Evol.* 7, 263–270. doi: 10.1016/j.meegid.2006.11.002

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics* 21, 163–193. doi: 10.1111/j.1096-0031.2005.00059.x

Bruselles, A., Rozera, G., Bartolini, B., Prosperi, M., Del Nonno, F., Narciso, P., et al. (2009). Use of massive parallel pyrosequencing for near full-length characterization of a unique HIV Type 1 BF recombinant associated with a fatal primary infection. *AIDS Res. Hum. Retroviruses* 25, 937–942. doi: 10.1089/aid.2009.0083

Calazans, A., Brindeiro, R., Brindeiro, P., Verli, H., Arruda, M. B., Gonzalez, L. M., et al. (2005). Low accumulation of L90M in protease from subtype F HIV-1 with resistance to protease inhibitors is caused by the L89M polymorphism. *J. Infect. Dis.* 191, 1961–1970. doi: 10.1086/430002

Cañada, J. E., Delgado, E., Gil, H., Sánchez, M., Benito, S., García-Bodas, E., et al. (2021). Identification of a new HIV-1 BC intersubtype circulating recombinant form (CRF108_BC) in Spain. *Viruses* 13:93. doi: 10.3390/v13010093

Carr, J. K., Avila, M., Gomez Carrillo, M., Salomon, H., Hierholzer, J., Watanaveeradej, V., et al. (2001). Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America. *AIDS* 15, F41–F47. doi: 10.1097/00002030-200110190-00002

Cid-Silva, P., Margusino-Framiñán, L., Balboa-Barreiro, V., Martín-Herranz, I., Castro-Iglesias, Á, Pernas-Souto, B., et al. (2018). Initial treatment response among HIV subtype F infected patients who started antiretroviral therapy based on integrase inhibitors. *AIDS* 32, 121–125. doi: 10.1097/QAD.0000000000001679

De Sá Filho, D. J., Sucupira, M. C., Caseiro, M. M., Sabino, E. C., Diaz, R. S., and Janini, L. M. (2006). Identification of two HIV type 1 circulating recombinant forms in Brazil. *AIDS Res. Hum. Retroviruses* 22, 1–13. doi: 10.1089/aid.2006.22.1

Delgado, E., Benito, S., Montero, V., Cuevas, M. T., Fernández-García, A., Sánchez-Martínez, M., et al. (2019). Diverse large HIV-1 non-subtype B clusters are spreading among men who have sex with men in Spain. *Front. Microbiol.* 10:655. doi: 10.3389/fmicb.2019.00655

Delgado, E., Cuevas, M. T., Domínguez, F., Vega, Y., Cabello, M., Fernández-García, A., et al. (2015). Phylogeny and phylogeography of a recent HIV-1 subtype F outbreak among men who have sex with men in Spain deriving from a cluster with a wide geographic circulation in Western Europe. *PLoS One* 10:e0143325. doi: 10.1371/journal.pone.0143325

Delgado, E., Fernández-García, A., Pérez-Losada, M., Moreno-Lorenzo, M., Fernández-Miranda, I., Benito, S., et al. (2021). Identification of CRF89_BF, a new member of an HIV-1 circulating BF intersubtype recombinant form family widely spread in South America. *Sci. Rep.* 11:11442. doi: 10.1038/s41598-021-90023-x

Delgado, E., Ríos, M., Fernández, J., Pérez-Alvarez, L., Nájera, R., and Thomson, M. M. (2010). Identification of a new HIV type 1 BF intersubtype circulating recombinant form (CRF44_BF) in Chile. *AIDS Res. Hum. Retroviruses* 26, 821–826. doi: 10.1089/aid.2010.0006

Delgado, E., Thomson, M. M., Villahermosa, M. L., Sierra, M., Ocampo, A., Miralles, C., et al. (2002). Identification of a newly characterized HIV-1 BG intersubtype circulating recombinant form in Galicia, Spain, which exhibits a pseudotype-like virion structure. *J. Acquir. Immune Defic. Syndr.* 29, 536–543. doi: 10.1097/00126334-200204150-00016

Dilernia, D. A., Jones, L. R., Pando, M. A., Rabinovich, R. D., Damilano, G. D., Turk, G., et al. (2011). Analysis of HIV type 1 BF recombinant sequences from South America dates the origin of CRF12_BF to a recombination event in the 1970s. *AIDS Res. Hum. Retroviruses* 27, 569–578. doi: 10.1089/AID.2010.0118

Fernández-García, A., Delgado, E., Cuevas, M. T., Vega, Y., Montero, V., Sánchez, M., et al. (2016). Identification of an HIV-1 BG intersubtype recombinant form (CRF73_BG), partially related to CRF14_BG, which is circulating in Portugal and Spain. *PLoS One* 11:e0148549. doi: 10.1371/journal.pone.0148549

Fernández-García, A., Pérez-Alvarez, L., Cuevas, M. T., Delgado, E., Muñoz-Nieto, M., Cilla, G., et al. (2010). Identification of a new HIV type 1 circulating BF intersubtype recombinant form (CRF47_BF) in Spain. *AIDS Res. Hum. Retroviruses* 26, 827–832. doi: 10.1089/aid.2009.0311

Gil, H., Delgado, E., Benito, S., Georgalis, L., Montero, V., Sánchez, M., et al. (2022). Transmission clusters, predominantly associated with men who have sex with men, play a main role in the propagation of HIV-1 in northern Spain (2013-2018). *Front. Microbiol.* 13:782609. doi: 10.3389/fmicb.2022.782609

Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B., and Suchard, M. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* 30, 713–724. doi: 10.1093/molbev/mss265

Guimarães, M. L., Eyer-Silva, W. A., Couto-Fernandez, J. C., and Morgado, M. G. (2008). Identification of two new CRF_BF in Rio de Janeiro State, Brazil. *AIDS* 22, 433–435. doi: 10.1097/QAD.0b013e3282f47ad0

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., et al. (2020). Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *Lancet HIV* 7, e772–e781. doi: 10.1016/S1473-3099(18)30647-9

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kiwanuka, N., Laeyendecker, O., Robb, M., Kigozi, G., Arroyo, M., McCutchan, F., et al. (2008). Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. *J. Infect. Dis.* 197, 707–713. doi: 10.1086/527416

Kouri, V., Khouri, R., Alemán, Y., Abrahantes, Y., Vercauteren, J., Pineda-Peña, A. C., et al. (2015). CRF19_cpx is an evolutionary fit HIV-1 variant strongly associated with rapid progression to AIDS in Cuba. *EBioMedicine* 2, 244–254. doi: 10.1016/j.ebiom.2015.01.015

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lefort, V., Longueville, J. E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424. doi: 10.1093/molbev/msx149

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152–160. doi: 10.1128/JVI.73.1.152-160.1999

Los Alamos National Laboratory (2021). *HIV Sequence Database*. Availble online at: https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html (accessed August 17, 2021).

Marcelin, A. G., Masquelier, B., Descamps, D., Izopet, J., Charpentier, C., Alloui, C., et al. (2008). Tipranavir-ritonavir genotypic resistance score in protease inhibitor-experienced patients. *Antimicrob. Agents Chemother.* 52, 3237–3243. doi: 10.1128/AAC.00133-08

Nájera, R., Delgado, E., Pérez-Álvarez, L., and Thomson, M. M. (2002). Genetic recombination and its role in the development of the HIV-1 pandemic. *AIDS* 16, S3–S16. doi: 10.1097/00002030-200216004-00002

Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I., and Hatzakis, A. (2016). The application of HIV molecular epidemiology to public health. *Infect. Genet. Evol.* 46, 159–168. doi: 10.1016/j.meegid.2016.06.021

Pérez-Álvarez, L., Delgado, E., Vega, Y., Montero, V., Cuevas, T., Fernández-García, A., et al. (2014). Predominance of CXCR4 tropism in HIV-1 CRF14_BG strains from newly diagnosed infections. *J. Antimicrob. Chemother.* 69, 246–253. doi: 10.1093/jac/dkt305

Pessôa, R., Watanabe, J. T., Calabria, P., Felix, A. C., Loureiro, P., Sabino, E. C., et al. (2014a). Deep sequencing of HIV-1 near full-length proviral genomes identifies high rates of BF1 recombinants including two novel circulating recombinant forms (CRF) 70_BF1 and a disseminating 71_BF1 among blood donors in Pernambuco, Brazil. *PLoS One* 9:e112674. doi: 10.1093/jac/dkt30510.1371/journal.pone.0112674

Pessôa, R., Carneiro Proietti, A. B., Busch, M. P., and Sanabani, S. S. (2014b). Identification of a novel HIV-1 circulating recombinant form (CRF72_BF1) in deep sequencing data from blood donors in Southeastern Brazil. *Genome Announc.* 2:e00386-14. doi: 10.1128/genomeA.00386-14

Pessôa, R., Loureiro, P., Esther Lopes, M., Carneiro-Proietti, A. B., Sabino, E. C., Busch, M. P., et al. (2016). Ultra-deep sequencing of HIV-1 near full-length and partial proviral genomes reveals high genetic diversity among Brazilian blood donors. *PLoS One* 11:e0152499. doi: 10.1371/journal.pone.0152499

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rambaut, A., Lam, T. T., Max, C. L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

Reis, M. N. G., Bello, G., Guimarães, M. L., and Stefani, M. M. A. (2017). Characterization of HIV-1 CRF90_BF1 and putative novel CRFs_BF1 in Central West, North and Northeast Brazilian regions. *PLoS One* 12:e0178578. doi: 10.1371/journal.pone.0178578

Reis, M. N. G., Guimarães, M. L., Bello, G., and Stefani, M. M. A. (2019). Identification of new HIV-1 circulating recombinant forms CRF81_cpx and CRF99_BF1 in Central Western Brazil and of unique BF1 recombinant forms. *Front. Microbiol.* 10:97. doi: 10.3389/fmicb.2019.00097

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Rife, B. D., Mavian, C., Chen, X., Ciccozzi, M., Salemi, M., Min, J., et al. (2017). Phylodynamic applications in 21$^{st}$ century global infectious disease research. *Glob. Health Res. Policy* 2:13. doi: 10.1186/s41256-017-0034-y

Ristic, N., Zukurov, J., Alkmim, W., Diaz, R. S., Janini, L. M., and Chin, M. P. (2011). Analysis of the origin and evolutionary history of HIV-1 CRF28_BF and CRF29_BF reveals a decreasing prevalence in the AIDS epidemic of Brazil. *PLoS One* 6:e17485. doi: 10.1371/journal.pone.0017485

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Ruchansky, D., Casado, C., Russi, J. C., Arbiza, J. R., and López-Galíndez, C. (2009). Identification of a new HIV type 1 circulating recombinant form (CRF38_BF1) in Uruguay. *AIDS Res. Hum. Retroviruses* 25, 351–356. doi: 10.1089/aid.2008.0248

Sa-Filho, D., Kallas, E. G., Sanabani, S., Sabino, E., Sucupira, M. C., Sanchez-Rosa, A. C., et al. (2007). Characterization of the full-length human immunodeficiency virus-1 genome from recently infected subjects in Brazil. *AIDS Res. Hum. Retroviruses* 23, 1087–1094. doi: 10.1089/aid.2006.0173

Salminen, M. O., Carr, J. K., Burke, D. S., and McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res. Hum. Retroviruses* 11, 1423–1425. doi: 10.1089/aid.1995.11.1423

Sanabani, S. S., Pastena, E. R., Neto, W. K., Martinez, V. P., and Sabino, E. C. (2010). Characterization and frequency of a newly identified HIV-1 BF1 intersubtype circulating recombinant form in São Paulo, Brazil. *Virol. J.* 7:74. doi: 10.1186/1743-422X-7-74

Sanchez, A. M., DeMarco, C. T., Hora, B., Keinonen, S., Chen, Y., and Brinkley, C. (2014). Development of a contemporary globally diverse HIV viral panel by the EQAPOL program. *J. Immunol. Methods* 409, 117–130. doi: 10.1016/j.jim.2014.01.004

Schultz, A. K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., et al. (2009). jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res.* 37, W647–W651. doi: 10.1093/nar/gkp371

Seager, I., Travers, S. A., Leeson, M. D., Crampin, A. C., French, N., Glynn, J. R., et al. (2014). Coreceptor usage, diversity, and divergence in drug-naive and drug-exposed individuals from Malawi, infected with HIV-1 subtype C for more than 20 years. *AIDS Res. Hum. Retroviruses* 30, 975–983. doi: 10.1089/aid.2013.0240

Shafer, R. W. (2006). Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* 194(Suppl. 1), S51–S58. doi: 10.1086/505356

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* 23, 7–9. doi: 10.1093/molbev/msj021

Sierra, M., Thomson, M. M., Ríos, M., Casado, G., Ojea de Castro, R., Delgado, E., et al. (2005). The analysis of near full-length genome sequences of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Chile, Venezuela and Spain reveals their relationship to diverse lineages of recombinant viruses related to CRF12_BF. *Infect. Genet. Evol.* 5, 209–217. doi: 10.1016/j.meegid.2004.07.010

Simonetti, F. R., Lai, A., Monno, L., Binda, F., Brindicci, G., Punzi, G., et al. (2014). Identification of a new HIV-1 BC circulating recombinant form (CRF60_BC) in Italian young men having sex with men. *Infect. Genet. Evol.* 23, 176–181. doi: 10.1016/j.meegid.2014.02.007

Song, H., Ou, W., Feng, Y., Zhang, J., Li, F., Hu, J., et al. (2019). Disparate impact on CD4 T cell count by two distinct HIV-1 phylogenetic clusters from the same clade. *Proc. Natl. Acad. Sci. U.S.A.* 116, 239–244. doi: 10.1073/pnas.1814714116

Struck, D., Roman, F., De Landtsheer, S., Servais, J. Y., Lambert, C., Masquelier, C., et al. (2015). Near full-length characterization and population dynamics of the human immunodeficiency virus type 1 circulating recombinant form 42 (CRF42_BF) in Luxembourg. *AIDS Res. Hum. Retroviruses* 31, 554–558. doi: 10.1089/AID.2014.0364

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 4:vey016. doi: 10.1093/ve/vey016

Thomson, M. M., and Fernández-García, A. (2011). Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning. *Virology* 415, 30–38. doi: 10.1016/j.virol.2011.03.021

Thomson, M. M., and Nájera, R. (2005). Molecular epidemiology of HIV-1 variants in the global AIDS pandemic: an update. *AIDS Rev.* 7, 210–224.

Thomson, M. M., Fernández-García, A., Delgado, E., Vega, Y., Díez-Fuertes, F., Sánchez-Martínez, M., et al. (2012). Rapid expansion of a HIV-1 subtype F cluster of recent origin among men who have sex with men in Galicia, Spain. *J. Acquir. Immune Defic. Syndr.* 59, e49–e51. doi: 10.1097/QAI.0b013e3182400fc4

Thomson, M. M., Herrero, I., Villahermosa, M. L., Vázquez de Parga, E., Cuevas, M. T., Carmona, R., et al. (2002). Diversity of mosaic structures and common ancestry of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Argentina revealed by analysis of near full-length genome sequences. *J. Gen. Virol.* 83, 107–119. doi: 10.1099/0022-1317-83-1-107

Thomson, M. M., Sierra, M., Tanuri, A., May, S., Casado, G., Manjón, N., et al. (2004). Analysis of near full-length genome sequences of HIV type 1 BF intersubtype recombinant viruses from Brazil reveals their independent origins and their lack of relationship to CRF12_BF. *AIDS Res. Hum. Retroviruses* 20, 1126–1133. doi: 10.1089/aid.2004.20.1126

Thomson, M. M., Villahermosa, M. L., Vázquez de Parga, E., Cuevas, M. T., Delgado, E., Manjón, N., et al. (2000). Widespread circulation of a B/F intersubtype recombinant form among HIV-1-infected individuals in Buenos Aires, Argentina. *AIDS* 14, 897–899. doi: 10.1097/00002030-200005050-00020

Travers, S. A., Clewley, J. P., Glynn, J. R., Fine, P. E., Crampin, A. C., Sibande, F., et al. (2004). Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. *J. Virol.* 78, 10501–10506. doi: 10.1128/JVI.78.19.10501-10506.2004

Trifinopoulos, J., Nguyen, L. T., von Haeseler, A., and Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44, W232–W235. doi: 10.1093/nar/gkw256

Vasylyeva, T. I., Zarebski, A., Smyrnov, P., Williams, L. D., Korobchuk, A., Liulchuk, M., et al. (2020). Phylodynamics helps to evaluate the impact of an HIV prevention intervention. *Viruses* 12:469. doi: 10.3390/v12040469

Venner, C. M., Nankya, I., Kyeyune, F., Demers, K., Kwok, C., Chen, P. L., et al. (2016). Infecting HIV-1 subtype predicts disease progression in women of Sub-Saharan Africa. *EBioMedicine* 13, 305–314. doi: 10.1016/j.ebiom.2016.10.014

Wensing, A. M., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Günthard, H. F., Paredes, R., et al. (2019). 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 27, 111–121.

Zhang, M., Foley, B., Schultz, A. K., Macke, J. P., Bulla, I., Stanke, M., et al. (2010). The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M., and Ho, D. D. (1998). An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* 391, 594–597. doi: 10.1038/35400

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fmicb.2022.866705

# Genetic Diversity and Low Therapeutic Impact of Variant-Specific Markers in HIV-1 Pol Proteins

*Paloma Troyano-Hernáez, Roberto Reinosa and Africa Holguín**

*HIV-1 Molecular Epidemiology Laboratory, Department of Microbiology, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP), Red en Investigación Translacional en Infecciones Pediátricas (RITIP), Madrid, Spain*

The emergence and spread of new HIV-1 variants pose a challenge for the effectiveness of antiretrovirals (ARV) targeting Pol proteins. During viral evolution, non-synonymous mutations have fixed along the viral genome, leading to amino acid (aa) changes that can be variant-specific (V-markers). Those V-markers fixed in positions associated with drug resistance mutations (DRM), or R-markers, can impact drug susceptibility and resistance pathways. All available HIV-1 Pol sequences from ARV-naïve subjects were downloaded from the United States Los Alamos HIV Sequence Database, selecting 59,733 protease (PR), 6,437 retrotranscriptase (RT), and 6,059 integrase (IN) complete sequences ascribed to the four HIV-1 groups and group M subtypes and circulating recombinant forms (CRFs). Using a bioinformatics tool developed in our laboratory (EpiMolBio), we inferred the consensus sequences for each Pol protein and HIV-1 variant to analyze the aa conservation in Pol. We analyzed the Wu–Kabat protein variability coefficient (WK) in PR, RT, and IN group M to study the susceptibility of each site to evolutionary replacements. We identified as V-markers the variant-specific aa changes present in >75% of the sequences in variants with >5 available sequences, considering R-markers those V-markers that corresponded to DRM according to the IAS-USA2019 and Stanford-Database 9.0. The mean aa conservation of HIV-1 and group M consensus was 82.60%/93.11% in PR, 88.81%/94.07% in RT, and 90.98%/96.02% in IN. The median group M WK was 10 in PR, 4 in RT, and 5 in IN. The residues involved in binding or catalytic sites showed a variability <0.5%. We identified 106 V-markers: 31 in PR, 28 in RT, and 47 in IN, present in 11, 12, and 13 variants, respectively. Among them, eight (7.5%) were R-markers, present in five variants, being minor DRM with little potential effect on ARV susceptibility. We present a thorough analysis of Pol variability among all HIV-1 variants circulating to date. The relatively high aa conservation observed in Pol proteins across HIV-1 variants highlights their critical role in the viral cycle. However, further studies are needed to understand the V-markers' impact on the Pol proteins structure, viral cycle, or treatment strategies, and periodic variability surveillance studies are also required to understand PR, RT, and IN evolution.

**Keywords: HIV-1, Pol, protease, integrase, conservation, variants, resistance, reverse transcriptase**

OPEN ACCESS

***Edited by:***
*Joris Hemelaar,*
*University of Oxford, United Kingdom*

***Reviewed by:***
*Diako Ebrahimi,*
*Texas Biomedical Research Institute, United States*
*Yuta Hikichi,*
*National Cancer Institute at Frederick (NIH), United States*

****Correspondence:***
*Africa Holguín*
*africa.holguin@salud.madrid.org*

***Specialty section:***
*This article was submitted to Virology, a section of the journal Frontiers in Microbiology*

***Received:*** *31 January 2022*
***Accepted:*** *06 June 2022*
***Published:*** *14 July 2022*

***Citation:***
*Troyano-Hernáez P, Reinosa R and Holguín A (2022) Genetic Diversity and Low Therapeutic Impact of Variant-Specific Markers in HIV-1 Pol Proteins. Front. Microbiol. 13:866705. doi: 10.3389/fmicb.2022.866705*

## INTRODUCTION

HIV is one of the most genetically diverse pathogens due to its high recombination and mutation rates and its rapid replication rate (Hemelaar, 2012; Hemelaar et al., 2019). HIV mutations during replication are favored by the error-prone polymerization by the HIV reverse transcriptase (RT) that lacks proofreading exonuclease activity (Roberts et al., 1988; Bebenek et al., 1993). HIV-1 is responsible for most HIV infections worldwide. It is divided into four groups according to genetic homology: M (major or main), N (non-M, non-O) (Simon et al., 1998), O (outlier) (De Leys et al., 1990), and P (Plantier et al., 2009). Group M is the main HIV group related to the current HIV global pandemic (Hemelaar et al., 2019). This group is subdivided into 10 subtypes (A–D, F–H, and J–L) and 8 sub-subtypes (A1, A2, A3, A4, A5, A6, F1, and F2) (Robertson et al., 2000; Salminen et al., 2000; Yamaguchi et al., 2020), at least 118 circulating recombinant forms (CRFs) (Los Alamos National Laboratory, 2021), and countless unique recombinant forms (URF). The Pol protein has been associated with differences in the replication capacity and disease progression of the different subtypes (Nagata et al., 2017).

The HIV-1 *pol* gene encodes the three enzymes needed for viral replication: protease (PR), RT, and integrase (IN). These proteins have essential roles in the viral cycle and are the main targets of antiretroviral drugs (ARV) (Huff, 1991; Eron, 2000; El Safadi et al., 2007; Gu et al., 2020; Jóźwik et al., 2020). Molecular detection of Pol mutations associated with ARV resistance has enabled resistance monitoring and individualization of antiretroviral treatment (ART) regimens in HIV-positive subjects (Clarke, 2002). This approach is well extended in middle- and high-income countries, where clinicians often use online resistance interpretation algorithms, such as Stanford HIVdb Program[1], to detect drug resistance mutations (DRM) in *pol* sequences and for HIV fast subtyping. HIV-1 *pol* diversity within HIV-1 variants is high and could impact ARV susceptibility (Holguín et al., 2006b). Surveillance of DRM in non-B subtypes and recombinants is essential (Holguín and Soriano, 2002; Holguín et al., 2004; Kantor, 2006; Llacer Delicado et al., 2016), as most studies focus on HIV-1 subtype B, more prevalent in Western Europe and the United States (Hemelaar et al., 2019).

The PR (99 aa) is responsible for processing the Gag and Gag–Pol precursors into mature Gag and Pol viral proteins by site-specific cleavage to produce the matrix, capsid, nucleocapsid, P1 and P2 spacer segments and P6 proteins of Gag, and the PR, RT, and IN proteins of Pol (Frankel and Young, 1998; Konvalinka et al., 2015). Variability in specific cleavage sites has been detected across HIV-1 groups, subtypes, and recombinants (Torrecilla et al., 2014). This could affect Gag and Pol proteins' processing, viral budding, restore viral fitness, and influence the virological outcome of specific ARV (Goodenow et al., 2002; Myint et al., 2004; Holguín et al., 2006a; Dam et al., 2009). PR functions as a dimer with flexible flaps that close down on the active site upon substrate binding. This site resembles other aspartyl proteases with the conserved triad sequence Asp25-Thr26-Gly27 (Navia et al., 1989; Frankel and Young, 1998). There are five FDA-approved protease inhibitors currently recommended in the HHS HIV/AIDS medical practice guidelines (NIH FDA-Approved HIV Medicines, 2022).

The RT catalyzes RNA-dependent and DNA-dependent DNA polymerization reactions (Hu and Hughes, 2012). It is a heterodimer containing subunits p51 (440 aa) and p66 (560 aa), each with a polymerase domain composed of four subdomains (fingers, palm, thumb, and connection) and identical sequences, except for p66 additional RNase H domain (Rodgers et al., 1995). The polymerase active site contains the catalytic triad Asp110, Asp185, and Asp186, conserved in many polymerases (Frankel and Young, 1998). There are two classes of RT inhibitors: nucleoside RT inhibitors (NRTI) and non-nucleoside RT inhibitors (NNRTI), with a total of 10 FDA-approved RT inhibitors currently recommended in the HHS HIV/AIDS medical practice guidelines (NIH FDA-Approved HIV Medicines, 2022).

The IN (288 aa) catalyzes a series of reactions to integrate the viral genome into the host chromosome (Frankel and Young, 1998; Engelman and Singh, 2018). The N-terminal domain (aa 1–55) is dimeric and contains a zinc-binding site: His12, His16, Cys40, and Cys43 (Cai et al., 1997). The catalytic domain (aa 50–212) contains a D–D–E motif (Asp64, Asp116, and Glu152) conserved among integrases, essential for the processing and joining reactions (Dyda et al., 1994; Rice et al., 1996). Finally, the C-terminal domain has non-specific DNA-binding activity (Eijkelenboom et al., 1999). Integrase strand transfer inhibitors (INSTIs) are the most recently developed ARV drugs. There are three INSTIs approved by the FDA and currently recommended in the HHS HIV/AIDS medical practice guidelines (NIH FDA-Approved HIV Medicines, 2022). The three drugs bind to a common D–D–E motif in the IN catalytic domain, causing it to disengage (Sharma et al., 2014).

Since ARV development, more than 100 DRM have been described and classified as primary or secondary according to their effect on ARV efficacy (Shafer and Schapiro, 2008). During viral evolution, non-synonymous nucleotide mutations have been fixed along the viral genome, leading to aa changes; some of them are variant-specific (V-markers). Some V-markers can be related to drug resistance (R-markers) when fixed in positions associated with ARV resistance in the absence of antiretroviral therapy and may impact drug susceptibility and resistance pathways (Holguín et al., 2006b).

HIV Pol protein has an essential functional role in the viral cycle, being the main target for ARV and often used by clinicians to classify HIV-1 variants. The emergence of new HIV-1 variants and the spread of HIV-1 non-B subtypes and recombinants worldwide pose a challenge for the accuracy and efficiency of ARV, DRM detection, and fast subtyping online tools.

This descriptive study presents a thorough analysis of Pol diversity among HIV-1 variants circulating to date using ARV-naïve *pol* sequences available in Los Alamos National Laboratory HIV Sequence Database (LANL). We provide the aa conservation rate per residue within variants in PR, RT, and IN proteins. We also identify the V-markers and the R-markers across HIV-1 variants, analyzing the mean conservation of the consensus

[1] https://hivdb.stanford.edu/hivdb

sequences of each HIV-1 variant and HIV-1 group in the three Pol proteins.

## MATERIALS AND METHODS

In January 2022, we downloaded from the LANL database[2] all the available *pol* HIV-1 sequences from drug-naïve subjects carrying different HIV-1 variants (groups, subtypes, sub-subtypes, and CRFs), selecting the corresponding genome region (PR, RT, and IN). Before the downloading process, we selected only drug-naïve sequences and only one sequence per patient in the LANL (Los Alamos National Laboratory, 2022b) platform. We also considered as INSTI naïve all participants with IN sequences sampled before 2007, the year of marketing authorization of the first INSTI, raltegravir, and the start date of the first clinical study with an authorized INSTI both in Europe and in the United States. URF sequences and incomplete PR, RT, and IN sequences were not included in this study. The sequences were also sorted by country of origin and organized in geographic regions according to the United Nations geoscheme[3] joining the regions of Central America and The Caribbean and the regions of Southern Asia and Southeastern Asia for practical purposes. The maps for **Figures 1**, **2** were created using MapChart[4].

A sequence analysis was performed with an in-house bioinformatics tool (EpiMolBio) previously designed and used in our laboratory for HIV genetic variability analysis and recently updated for SARS-CoV-2 sequences study (Burgos et al., 2019; Troyano-Hernáez et al., 2019, 2020, 2021a,b, 2022). This tool is programmed in JAVA OpenJDK version 11.0.9.1 using IDE NetBeans version 12.2. Among other functions, this tool calculates the conservation of a sequence set compared with a reference sequence and the rate of aa changes for each position within the studied protein. Furthermore, it can infer a consensus from a group of sequences or previously calculated consensuses considering the total number of sequences and the frequency of any specific aa residue per position, avoiding the overestimation of polymorphisms present in variants with a small number of available sequences. We used HIV-1 reference sequence HXB2 (NCBI accession number K03455.1) for the sequences' alignment and EpiMolBio functions that required a reference sequence, such as conservation analysis and V-marker detection.

We inferred the PR, RT, and IN consensus sequence for HIV-1, each HIV-1 group (M, N, O, and P), and each HIV-1 group M variant (subtype, sub-subtype, CRF) using all downloaded LANL sequences. Group M consensus was generated from the consensus of group M subtypes, sub-subtypes, and CRF. HIV-1 consensus was inferred considering the consensuses of the four groups (M, N, O, and P). We calculated the mean aa conservation of group M and HIV-1 consensus sequences for the three Pol proteins and the variability of the residues involved in binding or catalytic sites. We also studied the average aa conservation of the PR, RT, and IN group M variants with $>5$ available sequences compared with HXB2 HIV-1 reference sequence.

We identified the presence of single variant markers or V-markers, defined as the natural aa changes specific for each variant and present in $>75\%$ of the sequence set for a given position in variants with $>5$ sequences available in LANL, to avoid biases due to a low number of sequences. We considered as R-markers, the V-markers coinciding with major or minor DRM to the four main ARV families (PI, NRTI, NNRTI, and INSTI) according to the updated version of two sources: Stanford HIV Drug Resistance Database v9.0[5] and IAS-USA 2019 (Wensing et al., 2019). The sub-classification of DRM into major or minor DRM was done following the Stanford Database 9.0 criteria, which considers the effect on *in vitro* drug susceptibility, the frequency among patients with virological failure, the presence among untreated persons, and the location of the mutation within the 3D structure protein. Deletions and insertions were not included in this study.

In the *pol* variants with $>5$ available sequences, besides detecting V- and R-markers, we also checked for the presence of drug resistance mutations present in the WHO 2009 list for transmitted mutations or TDR (Bennett et al., 2009) and of major and minor DRM present in at least 25% of the sequences for each variant. To study the effect of these DRM and R-markers on ARV susceptibility, we analyzed them with the online resistance interpretation algorithm Stanford HIVdb Program v9.0[6], which infers susceptibility to 25 ARV from PI, NRTI, NNRTI, and INSTI drug families.

We calculated the Wu–Kabat protein variability coefficient (WK) for group M using all available PR, RT, and IN sequences belonging to this group. WK coefficient allows studying the susceptibility of an aa position to evolutionary replacements (Kabat et al., 1977). It was calculated using the following formula: variability = $N \times k/n$, where $N$ is the number of sequences in the alignment, $k$ is the number of different amino acids at a given position, and $n$ is the absolute frequency of the most common amino acid at that position. Therefore, a WK of 1 indicates the same aa was found for that position in all the sequence set, whereas a WK $>1$ indicates the relative variability of the respective site, with greater diversity as the WK value increases.

## RESULTS

### Analyzed Pol Sequences and Inferred Consensus Sequences

A total of 59,733 PR (32,745 group M non-B subtypes and CRF or non-B variants), 6,437 RT (4,393 non-B variants), and 6,059 IN (4,552 non-B variants) sequences were included in this study (**Supplementary Table 1**). Subtypes with the greatest sequence representation in group M were subtype B (45.1% in PR, 31.6% in RT, and 23.8% in IN), followed by subtype C (16.6% in PR, 25.0% in RT, and 22.3% in IN). The most represented CRF was recombinant 01_AE (15.2% in PR, 19.0% in RT, and 19.7% in IN).

The country of origin of the LANL available sequences for each Pol protein is illustrated in **Figure 1** (complete information in

[2] https://www.hiv.lanl.gov

[3] https://unstats.un.org

[4] https://www.mapchart.net

[5] https://hivdb.stanford.edu/

[6] https://hivdb.stanford.edu/hivdb/by-mutations/

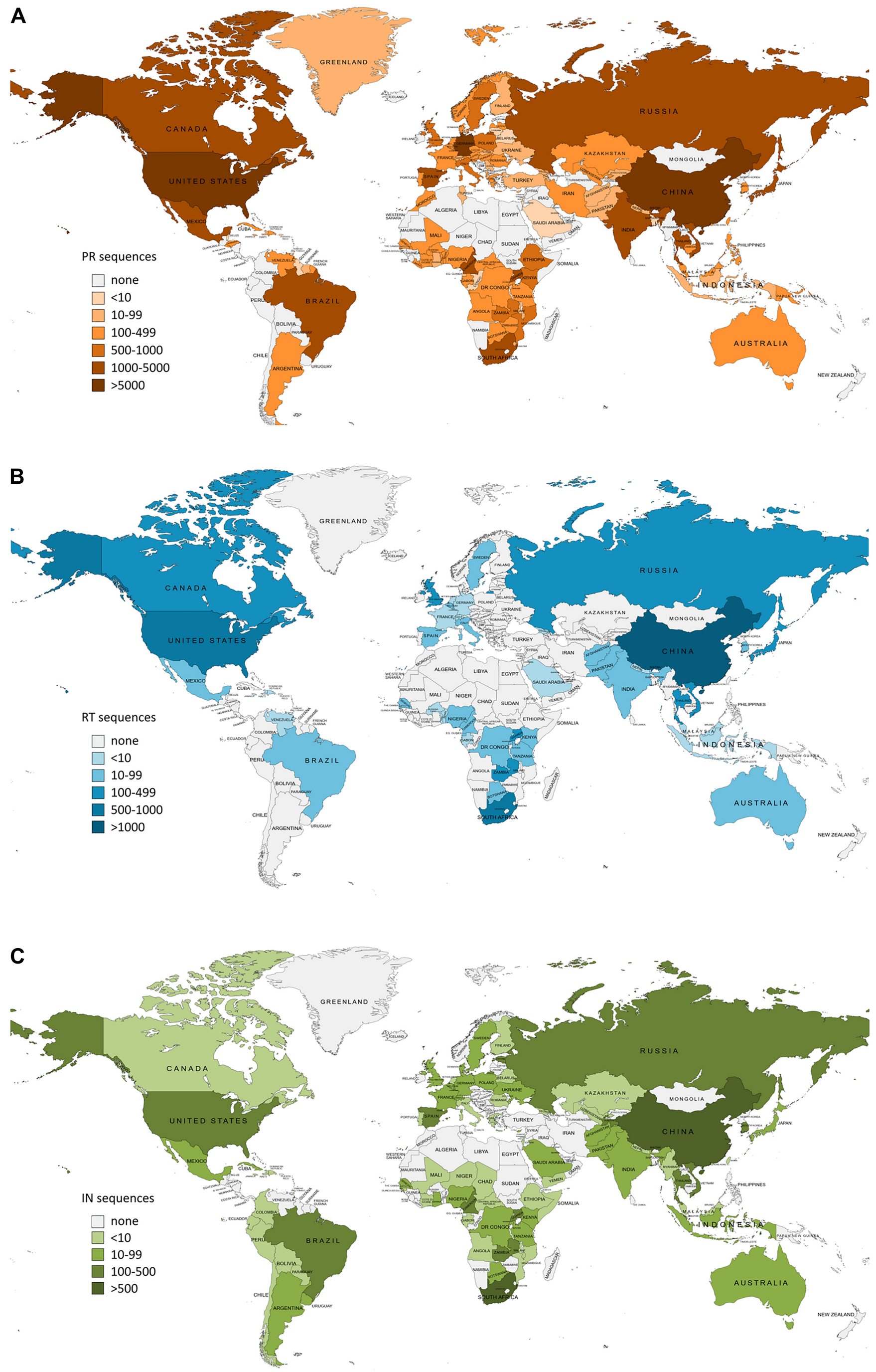


**FIGURE 1 |** Number of HIV-1 Pol sequences per country included in this study as available in Los Alamos HIV sequence database (LANL) in January 2022. PR, protease; RT, reverse transcriptase; IN, integrase. **(A)** Protease sequences per country. Total LANL sequences: 59.733. **(B)** Reverse transcriptase sequences per country. Total LANL sequences: 6.437. **(C)** Integrase sequences per country. Total LANL sequences: 6.059. Eleven integrase sequences had no record of the country of origin.

**Supplementary Table 2**). The geographic distribution by regions of the HIV-1 variants with available *pol* sequence in LANL is illustrated in **Figure 2** (described in **Supplementary Table 3**). PR sequences showed the highest diversity regarding the number of countries of origin of the sequenced samples (118 countries) compared with RT (50 countries) and IN (85 countries). In the three Pol proteins, China was the country that contributed the most sequences to the LANL database (11% in PR, 22% in RT, and 16% in IN), followed by South Africa in RT (15%) and IN (11%), and the United States in PR (10%), which was

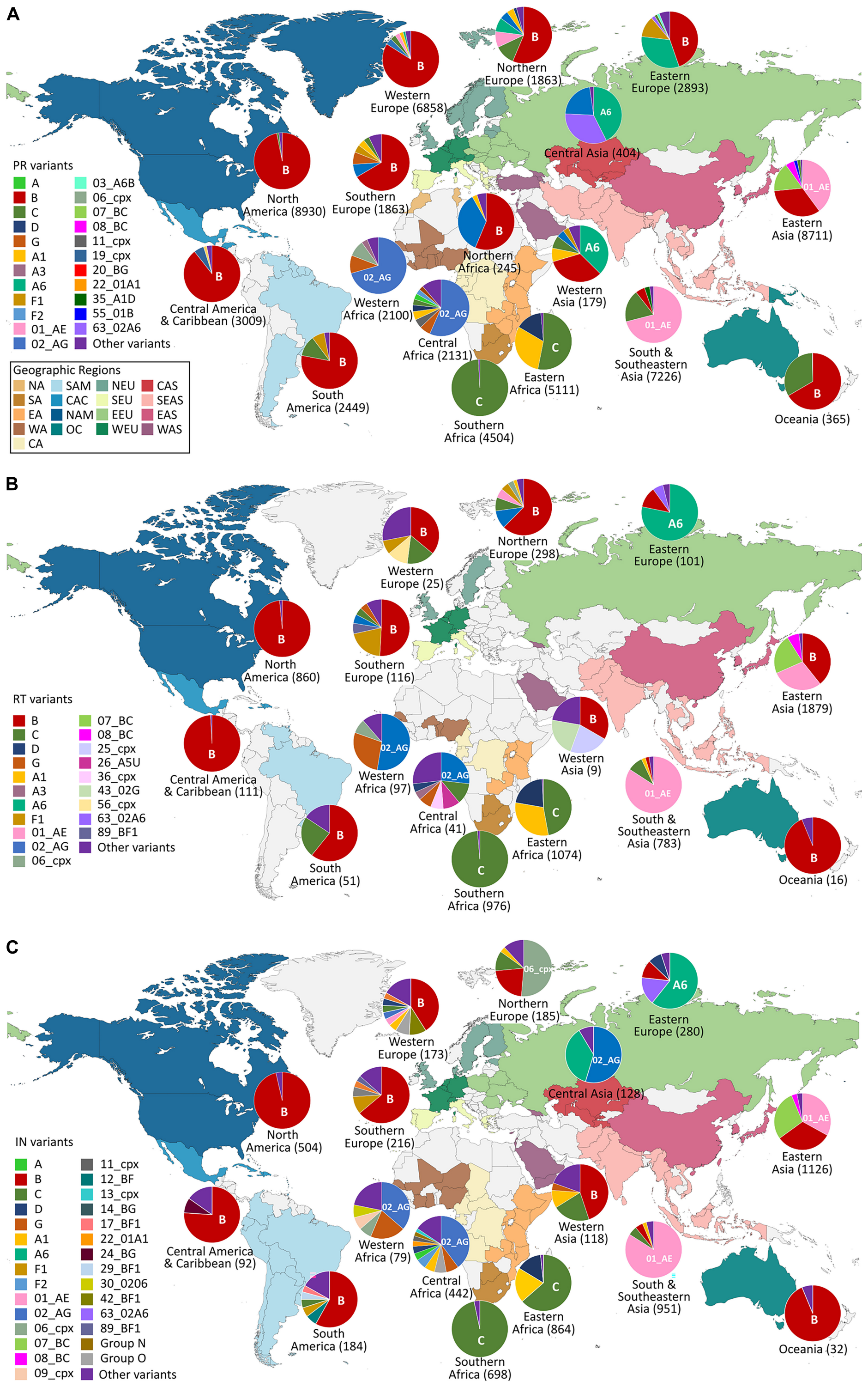


**FIGURE 2 |** Geographic distribution by regions of HIV-1 Pol variants available in Los Alamos HIV sequence database (LANL) in January 2022. HIV-1 variant distribution within regions in PR **(A)**, RT **(B)**, and IN **(C)**. PR, Protease; RT, reverse transcriptase; IN, integrase. Countries are colored by regions according to the United Nations geoscheme (https://unstats.un.org). Geographic regions color code inside the box in **(A)**. Pie graphs show the percentage of the HIV-1 variants per region as available in LANL in January 2022 and the most frequent variant per region. The total number of available LANL sequences per region is in brackets beside the region name. NA, Northern Africa; SA, Southern Africa; EA, Eastern Africa; WA, Western Africa; CA, Central Africa; SAM, South America; CAC, Central America and The Caribbean; NAM, North America; OC, Oceania; NEU, Northern Europe; SEU, Southern Europe; EEU, Eastern Europe; WEU, Western Europe; CAS, Central Asia; SEAS, Southern and Southeastern Asia; EAS, Eastern Asia; WAS, Western Asia.

in third place in RT (11%) and IN (8%). Thus, the countries present in each geographic region were not homogeneous in the three Pol proteins, and some geographic regions presented different HIV-1 main variants between Pol proteins. However, subtype B was the main variant for PR, RT, and IN in America, Western and Southern Europe, and Oceania, while there was a greater percentage of subtype A6 in Eastern Europe, CRF 01_AE in Southern, Southeastern, and Eastern Asia, and subtype C in Southern and Western Africa. The regions with more variant diversity in the three proteins were Central and Western Africa (**Figure 2**).

**Supplementary Table 4** reports the inferred consensus sequences for HIV-1, each HIV-1 group and group M variant in the three Pol proteins, showing the most frequent aa found per residue. Group M consensus was inferred using 84 variants' consensus in PR, 52 in RT, and 86 in IN. HIV-1 consensus and group M consensus sequences of PR, RT, and IN are displayed in **Figures 3–5**, respectively. The percentage of conservation of the most prevalent aa in each residue is indicated with a color code: dark green (100%), green (≥90–<100%), light green (>75–<90%), yellow (>50–≤75%), and gray (≤50%). The HXB2 reference sequence was included for further guidance. These figures also indicate the positions where major DRM to the four main drug families are located according to Stanford v9.0 and the location of PR, RT, and IN catalytic sites.

The mean aa conservation of HIV-1 and group M consensus sequences was 82.60 and 93.11% in PR, 88.81 and 94.07% in RT, and 90.98 and 96.02% in IN, respectively. **Table 1** describes the variability of PR, RT, and IN active sites and the mean residue variability in HIV-1 consensus for each protein. All the residues involved in binding or catalytic sites showed a variability below 0.5%, indicating a conservation over 95%. The mean aa conservation percentage in Pol residues of PR, RT, and IN is shown in **Figures 3–5**, respectively.

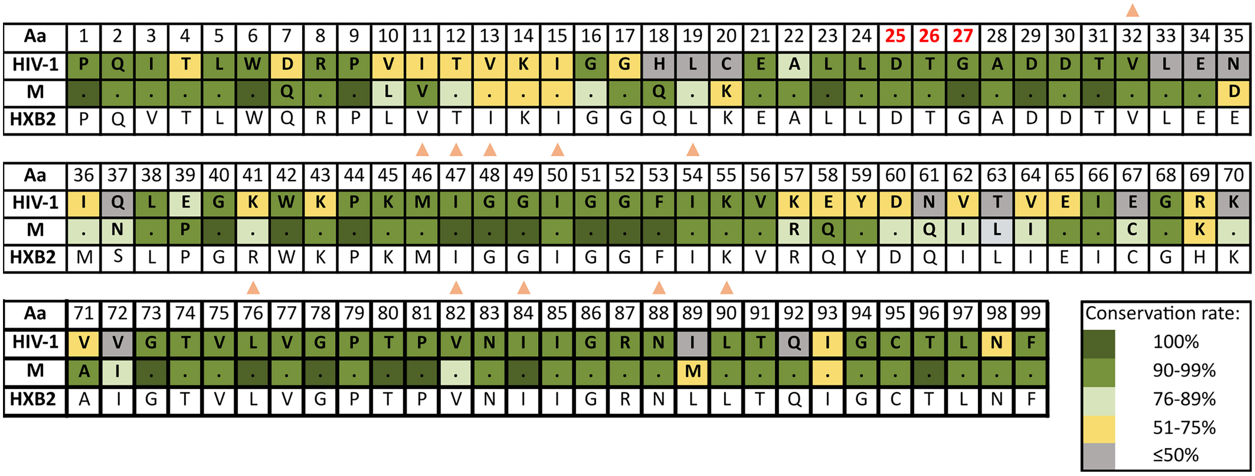
## Protease, Retrotranscriptase, and Integrase aa Conservation Across HIV-1 M Variants

We included in this analysis 46 variants in PR, 16 in RT, and 36 in IN, all with >5 sequences in LANL. Subtype B, the variant with the highest number of Pol sequences at LANL, was the most conserved variant in the three proteins (93.22% in PR, 96.02% in RT, and 92.09% in IN). The most conserved CRFs were 51_01B in PR (92.80%), 89_BF in RT (94.62%), and 42_BF in IN (94.31%). The least conserved variants were CRF13_cpx in PR (83.5%), subtype G in RT (90.44%), and CRF06_cpx in IN (92.09%). In RT and IN, all the variants had a conservation >90%, whereas, in PR, only 24% of the variants with >5 sequences had a conservation above 90% (**Figure 6**).

## V-Markers, R-Markers, and Other Drug Resistance Mutations

Among the variants with >5 available sequences, we found a total of 106 unique single V-markers and 8 R-markers present in >75% sequences (>75% conservation in their respective variants) across PR (**Table 2**), RT (**Table 3**), and IN (**Table 4**) variants. The analysis was performed in group O and 46 group M variants (7 subtypes, 6 sub-subtypes, and 33 CRF) of PR; in 16 group M variants (4 subtypes, 3 sub-subtypes, and 9 CRF) of RT; and in groups N and O, and 36 group M variants (7 subtypes, 5 sub-subtypes, and 24 CRF) of IN.

We detected 31 V-markers in PR (6 of them being R-markers), 28 V-markers in RT (1 R-marker), and 47 in IN (1 R-marker).



| Aa | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | P | Q | I | T | L | W | D | R | P | V | I | T | V | K | I | G | G | H | L | C | E | A | L | L | D | T | G | A | D | D | T | V | L | E | N |
| M | . | . | . | . | . | . | Q | . | . | L | V | . | . | . | . | . | . | Q | . | K | . | . | . | . | . | . | . | . | . | . | . | . | . | . | D |
| HXB2 | P | Q | V | T | L | W | Q | R | P | L | V | T | I | K | I | G | G | Q | L | K | E | A | L | L | D | T | G | A | D | D | T | V | L | E | E |

| Aa | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | I | Q | L | E | G | K | W | K | P | K | M | I | G | G | I | G | G | F | I | K | V | K | E | Y | D | N | V | T | V | E | I | E | G | R | K |
| M | . | N | . | P | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | R | Q | . | . | Q | I | L | I | . | . | C | . | K | . |
| HXB2 | M | S | L | P | G | R | W | K | P | K | M | I | G | G | I | G | G | F | I | K | V | R | Q | Y | D | Q | I | L | I | E | I | C | G | H | K |

| Aa | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | V | V | G | T | V | L | V | G | P | T | P | V | N | I | I | G | R | N | I | L | T | Q | I | G | C | T | L | N | F |
| M | A | I | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | M | . | . | . | . | . | . | . | . | . | . |
| HXB2 | A | I | G | T | V | L | V | G | P | T | P | V | N | I | I | G | R | N | L | L | T | Q | I | G | C | T | L | N | F |

Conservation rate:

100%

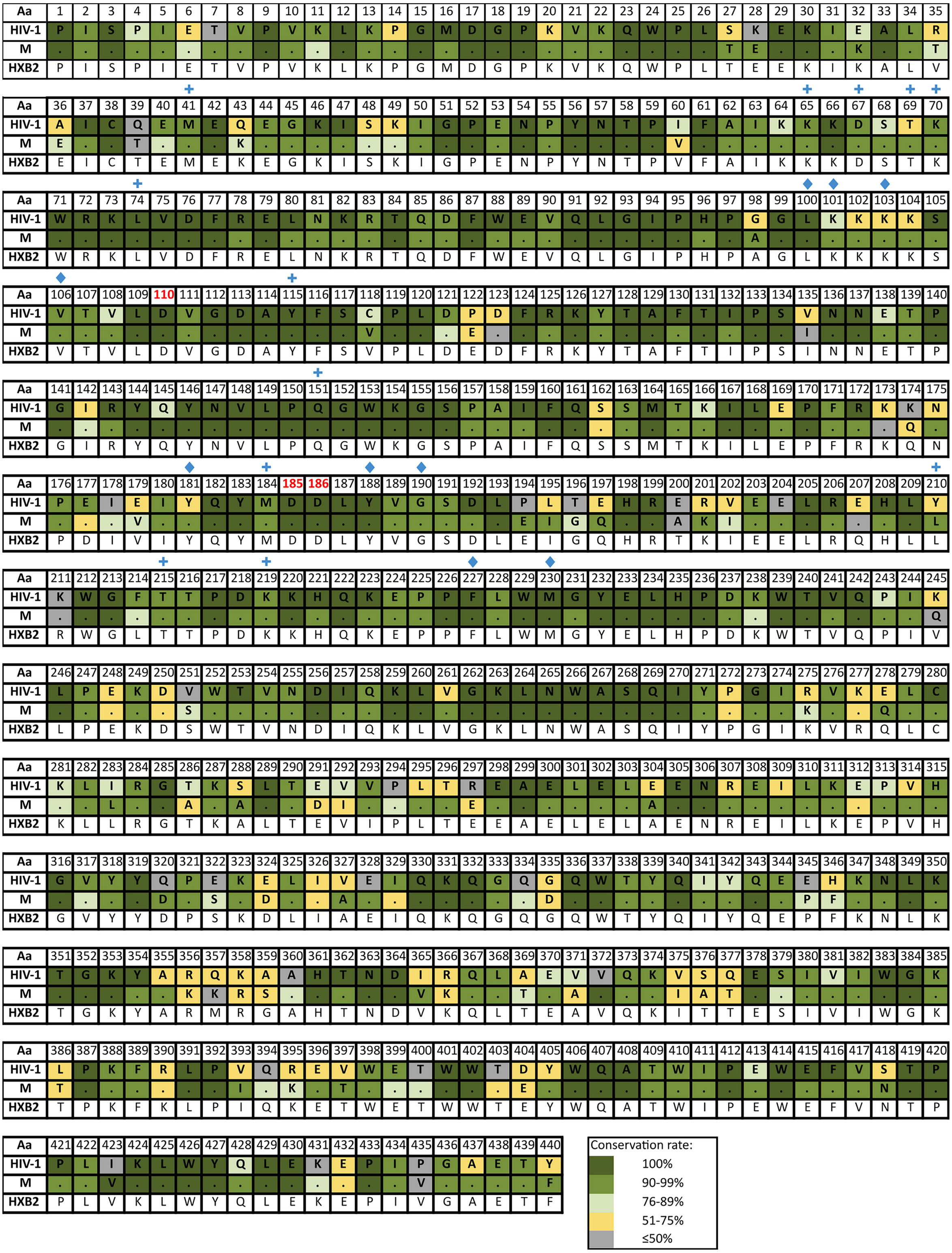
90-99%

76-89%

51-75%

≤50%

**FIGURE 3 |** Amino acid conservation rate along PR in HIV-1 and group M consensus. aa, amino acid; M, group M consensus. PR, protease (99 aa). Dots in group M represent the same aa as in HIV-1 consensus for that position. HXB2 reference sequence is described below the groups for further guidance. Colors represent the conservation rate. Residues of PR active site (triad Asp25-Thr26-Gly27, conserved among aspartyl proteases) are highlighted in red font. Orange triangles indicate positions where major DRM to PI are located according to Stanford v9.0 (Release Notes - HIV Drug Resistance Database, 2020) and summarized in https://cms.hivdb.org/prod/downloads/resistance-mutation-handout/resistance-mutation-handout.pdf. Aa code: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.



| Aa | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | P | I | S | P | I | E | T | V | P | V | K | L | K | P | G | M | D | G | P | K | V | K | Q | W | P | L | S | K | E | K | I | E | A | L | R |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | T | E | . | . | . | K | . | . | T |
| HXB2 | P | I | S | P | I | E | T | V | P | V | K | L | K | P | G | M | D | G | P | K | V | K | Q | W | P | L | T | E | E | K | I | K | A | L | V |

|  |  |  |  |  |  | + |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | + |  | + |  | + | + |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aa | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 |
| HIV-1 | A | I | C | Q | E | M | E | Q | E | G | K | I | S | K | I | G | P | E | N | P | Y | N | T | P | I | F | A | I | K | K | K | D | S | T | K |
| M | E | . | . | T | . | . | . | K | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | V | . | . | . | . | . | . | . | . | . | . |
| HXB2 | E | I | C | T | E | M | E | K | E | G | K | I | S | K | I | G | P | E | N | P | Y | N | T | P | V | F | A | I | K | K | K | D | S | T | K |

|  |  |  |  | + |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | ◆ | ◆ |  | ◆ |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aa | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 |
| HIV-1 | W | R | K | L | V | D | F | R | E | L | N | K | R | T | Q | D | F | W | E | V | Q | L | G | I | P | H | P | G | G | L | K | K | K | K | S |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | A | . | . | . | . | . | . | . |
| HXB2 | W | R | K | L | V | D | F | R | E | L | N | K | R | T | Q | D | F | W | E | V | Q | L | G | I | P | H | P | A | G | L | K | K | K | K | S |

|  | ◆ |  |  |  |  |  |  |  |  | + |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aa | 106 | 107 | 108 | 109 | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 |
| HIV-1 | V | T | V | L | D | V | G | D | A | Y | F | S | C | P | L | D | P | D | F | R | K | Y | T | A | F | T | I | P | S | V | N | N | E | T | P |
| M | . | . | . | . | . | . | . | . | . | . | . | . | V | . | . | . | E | . | . | . | . | . | . | . | . | . | . | . | . | I | . | . | . | . | . |
| HXB2 | V | T | V | L | D | V | G | D | A | Y | F | S | V | P | L | D | E | D | F | R | K | Y | T | A | F | T | I | P | S | I | N | N | E | T | P |

|  |  |  |  |  |  |  |  |  |  |  | + |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aa | 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| HIV-1 | G | I | R | Y | Q | Y | N | V | L | P | Q | G | W | K | G | S | P | A | I | F | Q | S | S | M | T | K | I | L | E | P | F | R | K | K | N |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | Q | . |
| HXB2 | G | I | R | Y | Q | Y | N | V | L | P | Q | G | W | K | G | S | P | A | I | F | Q | S | S | M | T | K | I | L | E | P | F | R | K | Q | N |

|  |  |  |  |  |  | ◆ |  |  | + |  |  |  | ◆ |  | ◆ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | + |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aa | 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 | 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 | 208 | 209 | 210 |
| HIV-1 | P | E | I | E | I | Y | Q | Y | M | D | D | L | Y | V | G | S | D | L | P | L | T | E | H | R | E | R | V | E | E | L | R | E | H | L | Y |
| M | . | . | . | V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | E | I | G | Q | . | . | A | K | I | . | . | . | . | . | . | . | L |
| HXB2 | P | D | I | V | I | Y | Q | Y | M | D | D | L | Y | V | G | S | D | L | E | I | G | Q | H | R | T | K | I | E | E | L | R | Q | H | L | L |

|  |  |  |  |  | + |  |  |  | + |  |  |  |  |  |  |  | ◆ |  |  | ◆ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aa | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 | 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 | 240 | 241 | 242 | 243 | 244 | 245 |
| HIV-1 | K | W | G | F | T | T | P | D | K | K | H | Q | K | E | P | P | F | L | W | M | G | Y | E | L | H | P | D | K | W | T | V | Q | P | I | K |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | Q |
| HXB2 | R | W | G | L | T | T | P | D | K | K | H | Q | K | E | P | P | F | L | W | M | G | Y | E | L | H | P | D | K | W | T | V | Q | P | I | V |

| Aa | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 | 256 | 257 | 258 | 259 | 260 | 261 | 262 | 263 | 264 | 265 | 266 | 267 | 268 | 269 | 270 | 271 | 272 | 273 | 274 | 275 | 276 | 277 | 278 | 279 | 280 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | L | P | E | K | D | V | W | T | V | N | D | I | Q | K | L | V | G | K | L | N | W | A | S | Q | I | Y | P | G | I | R | V | K | E | L | C |
| M | . | . | . | . | . | S | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | K | . | . | Q | . | . |
| HXB2 | L | P | E | K | D | S | W | T | V | N | D | I | Q | K | L | V | G | K | L | N | W | A | S | Q | I | Y | P | G | I | K | V | R | Q | L | C |

| Aa | 281 | 282 | 283 | 284 | 285 | 286 | 287 | 288 | 289 | 290 | 291 | 292 | 293 | 294 | 295 | 296 | 297 | 298 | 299 | 300 | 301 | 302 | 303 | 304 | 305 | 306 | 307 | 308 | 309 | 310 | 311 | 312 | 313 | 314 | 315 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | K | L | I | R | G | T | K | S | L | T | E | V | V | P | L | T | R | E | A | E | L | E | L | E | E | N | R | E | I | L | K | E | P | V | H |
| M | . | . | L | . | . | A | . | A | . | . | D | I | . | . | . | . | E | . | . | . | . | . | . | A | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | K | L | L | R | G | T | K | A | L | T | E | V | I | P | L | T | E | E | A | E | L | E | L | A | E | N | R | E | I | L | K | E | P | V | H |

| Aa | 316 | 317 | 318 | 319 | 320 | 321 | 322 | 323 | 324 | 325 | 326 | 327 | 328 | 329 | 330 | 331 | 332 | 333 | 334 | 335 | 336 | 337 | 338 | 339 | 340 | 341 | 342 | 343 | 344 | 345 | 346 | 347 | 348 | 349 | 350 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | G | V | Y | Y | Q | P | E | K | E | L | I | V | E | I | Q | K | Q | G | Q | G | Q | W | T | Y | Q | I | Y | Q | E | E | H | K | N | L | K |
| M | . | . | . | . | D | . | S | . | D | . | . | A | . | . | . | . | . | . | . | D | . | . | . | . | . | . | . | . | . | P | F | . | . | . | . |
| HXB2 | G | V | Y | Y | D | P | S | K | D | L | I | A | E | I | Q | K | Q | G | Q | G | Q | W | T | Y | Q | I | Y | Q | E | P | F | K | N | L | K |

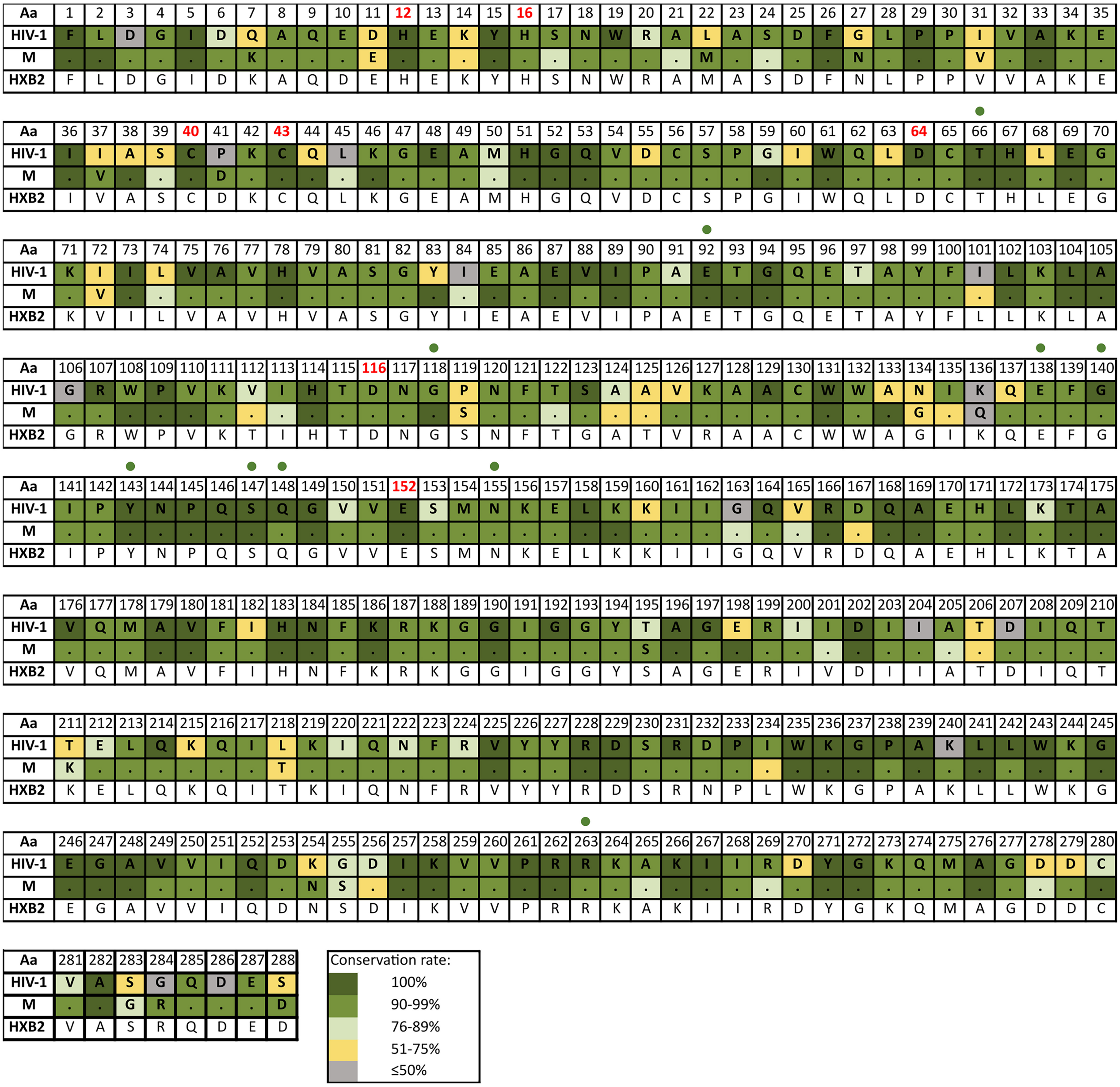
| Aa | 351 | 352 | 353 | 354 | 355 | 356 | 357 | 358 | 359 | 360 | 361 | 362 | 363 | 364 | 365 | 366 | 367 | 368 | 369 | 370 | 371 | 372 | 373 | 374 | 375 | 376 | 377 | 378 | 379 | 380 | 381 | 382 | 383 | 384 | 385 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | T | G | K | Y | A | R | Q | K | A | A | H | T | N | D | I | R | Q | L | A | E | V | V | Q | K | V | S | Q | E | S | I | V | I | W | G | K |
| M | . | . | . | . | . | K | K | R | S | . | . | . | . | . | V | K | . | . | T | . | A | . | . | . | I | A | T | . | . | . | . | . | . | . | . |
| HXB2 | T | G | K | Y | A | R | M | R | G | A | H | T | N | D | V | K | Q | L | T | E | A | V | Q | K | I | T | T | E | S | I | V | I | W | G | K |

| Aa | 386 | 387 | 388 | 389 | 390 | 391 | 392 | 393 | 394 | 395 | 396 | 397 | 398 | 399 | 400 | 401 | 402 | 403 | 404 | 405 | 406 | 407 | 408 | 409 | 410 | 411 | 412 | 413 | 414 | 415 | 416 | 417 | 418 | 419 | 420 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | L | P | K | F | R | L | P | V | Q | R | E | V | W | E | T | W | W | T | D | Y | W | Q | A | T | W | I | P | E | W | E | F | V | S | T | P |
| M | T | . | . | . | . | . | . | I | . | K | . | T | . | . | . | . | . | . | E | . | . | . | . | . | . | . | . | . | . | . | . | . | N | . | . |
| HXB2 | T | P | K | F | K | L | P | I | Q | K | E | T | W | E | T | W | W | T | E | Y | W | Q | A | T | W | I | P | E | W | E | F | V | N | T | P |

| Aa | 421 | 422 | 423 | 424 | 425 | 426 | 427 | 428 | 429 | 430 | 431 | 432 | 433 | 434 | 435 | 436 | 437 | 438 | 439 | 440 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | P | L | I | K | L | W | Y | Q | L | E | K | E | P | I | P | G | A | E | T | Y |
| M | . | . | V | . | . | . | . | . | . | . | . | . | . | . | V | . | . | . | . | F |
| HXB2 | P | L | V | K | L | W | Y | Q | L | E | K | E | P | I | V | G | A | E | T | F |

Conservation rate:
100%
90-99%
76-89%
51-75%
≤50%

**FIGURE 4 |** Amino acid conservation rate along RT in HIV-1 and group M consensus. Aa, amino acid; M, group M consensus. RT, reverse transcriptase (440 aa). Dots in group M represent the same aa as in HIV-1 consensus for that position. HXB2 reference sequence is described below the groups for further guidance. Colors represent the conservation rate. Residues of the catalytic triad (Asp110, Asp185, and Asp186) are highlighted in red font. Blue crosses and blue diamonds indicate positions where DRM to NRTI and to NNRTI, respectively, are located according to Stanford v9.0 (Release Notes - HIV Drug Resistance Database, 2020). Aa code according to **Figure 3**.



| Aa | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | F | L | D | G | I | D | Q | A | Q | E | D | H | E | K | Y | H | S | N | W | R | A | L | A | S | D | F | G | L | P | P | I | V | A | K | E |
| M | . | . | . | . | . | . | K | . | . | . | E | . | . | . | . | . | . | . | . | . | . | M | . | . | . | . | N | . | . | . | V | . | . | . | . |
| HXB2 | F | L | D | G | I | D | K | A | Q | D | E | H | E | K | Y | H | S | N | W | R | A | M | A | S | D | F | N | L | P | P | V | V | A | K | E |

| Aa | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 ● | 67 | 68 | 69 | 70 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | I | I | A | S | C | P | K | C | Q | L | K | G | E | A | M | H | G | Q | V | D | C | S | P | G | I | W | Q | L | D | C | T | H | L | E | G |
| M | . | V | . | . | . | D | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | I | V | A | S | C | D | K | C | Q | L | K | G | E | A | M | H | G | Q | V | D | C | S | P | G | I | W | Q | L | D | C | T | H | L | E | G |

| Aa | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 ● | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | K | I | I | L | V | A | V | H | V | A | S | G | Y | I | E | A | E | V | I | P | A | E | T | G | Q | E | T | A | Y | F | I | L | K | L | A |
| M | . | V | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | K | V | I | L | V | A | V | H | V | A | S | G | Y | I | E | A | E | V | I | P | A | E | T | G | Q | E | T | A | Y | F | L | L | K | L | A |

| Aa | 106 | 107 | 108 | 109 | 110 | 111 | 112 | 113 | 114 | 115 | 116 | 117 | 118 ● | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 ● | 139 | 140 ● |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | G | R | W | P | V | K | V | I | H | T | D | N | G | P | N | F | T | S | A | A | V | K | A | A | C | W | W | A | N | I | K | Q | E | F | G |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | S | . | . | . | . | . | . | . | . | . | . | . | . | . | . | G | . | Q | . | . | . | . |
| HXB2 | G | R | W | P | V | K | T | I | H | T | D | N | G | S | N | F | T | G | A | T | V | R | A | A | C | W | W | A | G | I | K | Q | E | F | G |

| Aa | 141 | 142 | 143 ● | 144 | 145 | 146 | 147 ● | 148 ● | 149 | 150 | 151 | 152 | 153 | 154 | 155 ● | 156 | 157 | 158 | 159 | 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | I | P | Y | N | P | Q | S | Q | G | V | V | E | S | M | N | K | E | L | K | K | I | I | G | Q | V | R | D | Q | A | E | H | L | K | T | A |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | I | P | Y | N | P | Q | S | Q | G | V | V | E | S | M | N | K | E | L | K | K | I | I | G | Q | V | R | D | Q | A | E | H | L | K | T | A |

| Aa | 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 | 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 | 208 | 209 | 210 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | V | Q | M | A | V | F | I | H | N | F | K | R | K | G | G | I | G | G | Y | T | A | G | E | R | I | I | D | I | I | A | T | D | I | Q | T |
| M | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | S | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | V | Q | M | A | V | F | I | H | N | F | K | R | K | G | G | I | G | G | Y | S | A | G | E | R | I | V | D | I | I | A | T | D | I | Q | T |

| Aa | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 | 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 | 240 | 241 | 242 | 243 | 244 | 245 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | T | E | L | Q | K | Q | I | L | K | I | Q | N | F | R | V | Y | Y | R | D | S | R | D | P | I | W | K | G | P | A | K | L | L | W | K | G |
| M | K | . | . | . | . | . | . | T | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | K | E | L | Q | K | Q | I | T | K | I | Q | N | F | R | V | Y | Y | R | D | S | R | N | P | L | W | K | G | P | A | K | L | L | W | K | G |

| Aa | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 | 256 | 257 | 258 | 259 | 260 | 261 | 262 | 263 ● | 264 | 265 | 266 | 267 | 268 | 269 | 270 | 271 | 272 | 273 | 274 | 275 | 276 | 277 | 278 | 279 | 280 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV-1 | E | G | A | V | V | I | Q | D | K | G | D | I | K | V | V | P | R | R | K | A | K | I | I | R | D | Y | G | K | Q | M | A | G | D | D | C |
| M | . | . | . | . | . | . | . | . | N | S | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| HXB2 | E | G | A | V | V | I | Q | D | N | S | D | I | K | V | V | P | R | R | K | A | K | I | I | R | D | Y | G | K | Q | M | A | G | D | D | C |

| Aa | 281 | 282 | 283 | 284 | 285 | 286 | 287 | 288 |
|---|---|---|---|---|---|---|---|---|
| HIV-1 | V | A | S | G | Q | D | E | S |
| M | . | . | G | R | . | . | . | D |
| HXB2 | V | A | S | R | Q | D | E | D |

Conservation rate:

| | |
|---|---|
| 🟩 | 100% |
| 🟩 | 90-99% |
| 🟩 | 76-89% |
| 🟨 | 51-75% |
| ⬜ | ≤50% |

**FIGURE 5 |** Amino acid conservation rate along IN in HIV-1 and group M consensus. Aa, amino acid; M, group M consensus. IN, integrase (288 aa). Dots in group M represent the same aa as in HIV-1 consensus for that position. HXB2 reference sequence is described below the groups for further guidance. Colors represent the conservation rate. Residues of the zinc-binding site (His12, His16, Cys40, and Cys43) and the D–D–E motif of the catalytic domain (Asp64, Asp116, and Glu152) are highlighted in red font. Green circles indicate positions where major INSTIs DRM are located according to Stanford v9.0 (Release Notes - HIV Drug Resistance Database, 2020). Aa code according to **Figure 3**.

None of the R-markers corresponded to major DRM, being all of them minor DRM according to Stanfordv 9.0. No V-markers were observed in the PR active site (Asp25, Thr26, and Gly27), in the RT catalytic triad (Asp110, Asp185, and Asp186), the IN zinc-binding site (His12, His16, Cys40, and Cys43), or the D–D–E motif (Asp64, Asp116, and Glu152) of the IN catalytic domain.

The 31 V-markers in PR were present in 11 variants (9 CRF, subtype J, and group O), being group O the variant with most V-markers: 15 (**Table 2**). Six (19.3%) of the 31 PR V-markers corresponded to R-markers: Four were detected in group O (K43T/Q58E/H69R/A71V) and two in group M complex recombinants CRF13_cpx (V77I) and CRF60_BC (L10V). The R-markers in group O were present in 83.3% (H69R), 87.5% (K43T), and 100% (Q58E, A71V) of this group's PR sequences. L10V was found in 95.5% of CRF60_BC PR sequences and V77I in 96.6% of CRF13_cpx sequences. In RT, we detected a total of 28 V-markers in 12 group M variants (5 subtypes and 7 CRF). Only one (3.6%) of them was an R-marker: V179E, found in 100%

**TABLE 1 |** Variability in protease (PR), retrotranscriptase (RT), and integrase (IN) active sites.

| Protein | Sites | | Variability |
|---|---|---|---|
| PR(99 aa) | Complete PR | | 17.40%* |
| | Active site triad | Asp25 | 0.11% |
| | | Thr26 | 0.04% |
| | | Gly27 | 0.01% |
| RT(440 aa) | Complete RT | | 11.19%* |
| | Catalytic triad | Asp110 | 0.03% |
| | | Asp185 | 0.04% |
| | | Asp186 | 0.06% |
| IN(288 aa) | Complete IN | | 9.02%* |
| | Zinc-binding site | His12 | 0.08% |
| | | His16 | 0.15% |
| | | Cys40 | 0.13% |
| | | Cys43 | 0.05% |
| | D–D–E motif | Asp64 | 0.07% |
| | | Asp116 | 0.08% |
| | | Glu152 | 0.23% |

*PR, protease; RT, reverse transcriptase; IN, integrase; with an asterisk, mean residue variability for each protein calculated from their respective HIV-1 consensus.*

of CRF55_01B RT sequences (**Table 3**). The largest number of V-markers was found in IN: 47 total V-markers in two non-M groups and 11 group M variants (3 subtypes and 8 CRF) (**Table 4**). Group O presented most V-markers (25/47, 53.2%). The only R-marker detected was E157Q in 100% of CRF03_A6B sequences.

No TDR present in the WHO list or major DRM to PI, NNRTI, NRTI, or INSTI were found in ≥25% of sequences belonging to PR, RT, or IN variants with >5 available sequences. However, we found seven minor DRM in the three Pol proteins. Two minor DRM to PI were detected in PR: Q58E and K43T, coinciding with two of group O R-markers in PR (**Table 2**). Another two minor DRM to NNRTI were found in RT: V179E in CRF06_cpx (29.4% sequences), coinciding with the only R-marker found in RT in CRF55_01B (**Table 2**), and V106I in sub-subtype F1 (42.5%). We found one minor DRM to NRTI, A62V, present in 47% of RT sequences of sub-subtype A6. Finally, two minor DRM to INSTI were detected in IN: G163R in CRF17_BF1 (28.6% sequences) and CRF89_BF1 (50%), and M50I in group O (49%), subtype A (60%), subtype C (48%), CRFs 11_cpx (31.2%), CRF22_01A1 (78.6%), and CRF63_02A6 (76.6%).

## Wu–Kabat Pol Variability Coefficient in Protease, Retrotranscriptase, and Integrase Group M

**Figure 7** describes the group M variability WK coefficient plot in the three *pol* proteins using all available LANL sequences for this group (26,988 PR, 2,044 RT, and 1,507 IN), including 84/52/86 HIV-1 group M variants in PR/RT/IN: 9/5/9 subtypes, 8/6/7 sub-subtypes, and 67/41/70 CRF. The WK values for each residue and studied protein are described in **Supplementary Table 5**.

The median variability coefficient in PR group M sequences was 10.26. The highest WK coefficient was 52 in residue 63, followed by WK 35 in site 69 (**Figure 7A**). The lowest WK was 3 in site 27, part of the triad Asp25, Thr26, and Gly27 in the PR active site. The other two residues of this triad (Asp25 and Thr26) presented a WK of 9 and 7, respectively. None of the 99 residues along PR were completely conserved (WK 1). Most PR residues had a WK between 10 and 20 (**Figure 8**).

The median WK along RT in group M sequences was 4.03. Site 245 presented the highest WK coefficient (WK 47.61), followed by site 207 (WK 46.84) and site 211 (WK 37.2) (**Figure 7B**). The smallest WK was 1, present in 20 of the 440 RT residues. The RT catalytic triad Asp110, Asp185, and Asp186 had a WK of 2, 3, and 4, respectively. Most RT residues had a WK between 1 and 5 (**Figure 8**).

IN median variability coefficient in group M sequences was 5. The highest WK coefficient was 29.35, located in residue 136, followed by site 125 (WK 16.6) and site 234 (WK 16.22) (**Figure 7C**). The smallest WK was 1, present in sites 131 and 235. The residues of the IN zinc-binding site, namely, His12, His16, and Cys40, showed a WK coefficient of 4, 7, and 5, respectively. The IN catalytic domain's Asp64, Asp116, and Glu152 motif presented a WK of 4, 3, and 6, respectively. Most IN residues had a WK between 1 and 5 (**Figure 8**).

# DISCUSSION

This descriptive study analyzes the Pol diversity among HIV-1 variants, providing the aa conservation rate per residue and HIV-1 variant in PR, RT, and IN proteins. A better understanding of HIV variability is important, since it has been reported that HIV-1 transmissibility, replication, and disease progression can differ between HIV-1 variants (Renjifo et al., 2004; Baeten et al., 2007; Bennett et al., 2009; Ng et al., 2014). The HIV-1 Pol proteins PR, RT, and IN are essential for viral replication and are the main targets of ARV (Huff, 1991; Eron, 2000; El Safadi et al., 2007; Gu et al., 2020; Jóźwik et al., 2020). Since Pol variability can impact ARV monitoring and efficacy, conservation studies must consider all circulating HIV-1 variants worldwide.

The consensus sequences of HIV proteins and their conservation studies allow a better understanding of structural, functional, and immunogenic potential differences across HIV-1 groups, subtypes, sub-subtypes, and recombinants and have been previously analyzed in other HIV-1 proteins (Li et al., 2013; Sliepen et al., 2019; Zhang et al., 2021). A recent work by Linchangco et al. reconstructed 90 HIV-1 subtype and CRF consensus sequences from 3,470 full HIV genomes downloaded from LANL (Linchangco et al., 2021). Our study updates and expands the knowledge regarding HIV Pol variability, including 59,733 PR, 6,437 RT, and 6,059 IN sequences from more than 100 different variants, including all the currently available HIV-1 groups, subtypes, and CRF in LANL. Moreover, **Supplementary Table 4** summarizes the aa conservation in each Pol protein and each variant to help identify the conservation or consensus aa in any Pol residue and HIV-1 variant of interest.

The sequences were processed by an in-house bioinformatics tool (EpiMolBio) developed for HIV and SARS-CoV-2 variability analysis. In the most extensive panel of HIV-1 variants analyzed to date, we have also identified the natural polymorphisms that

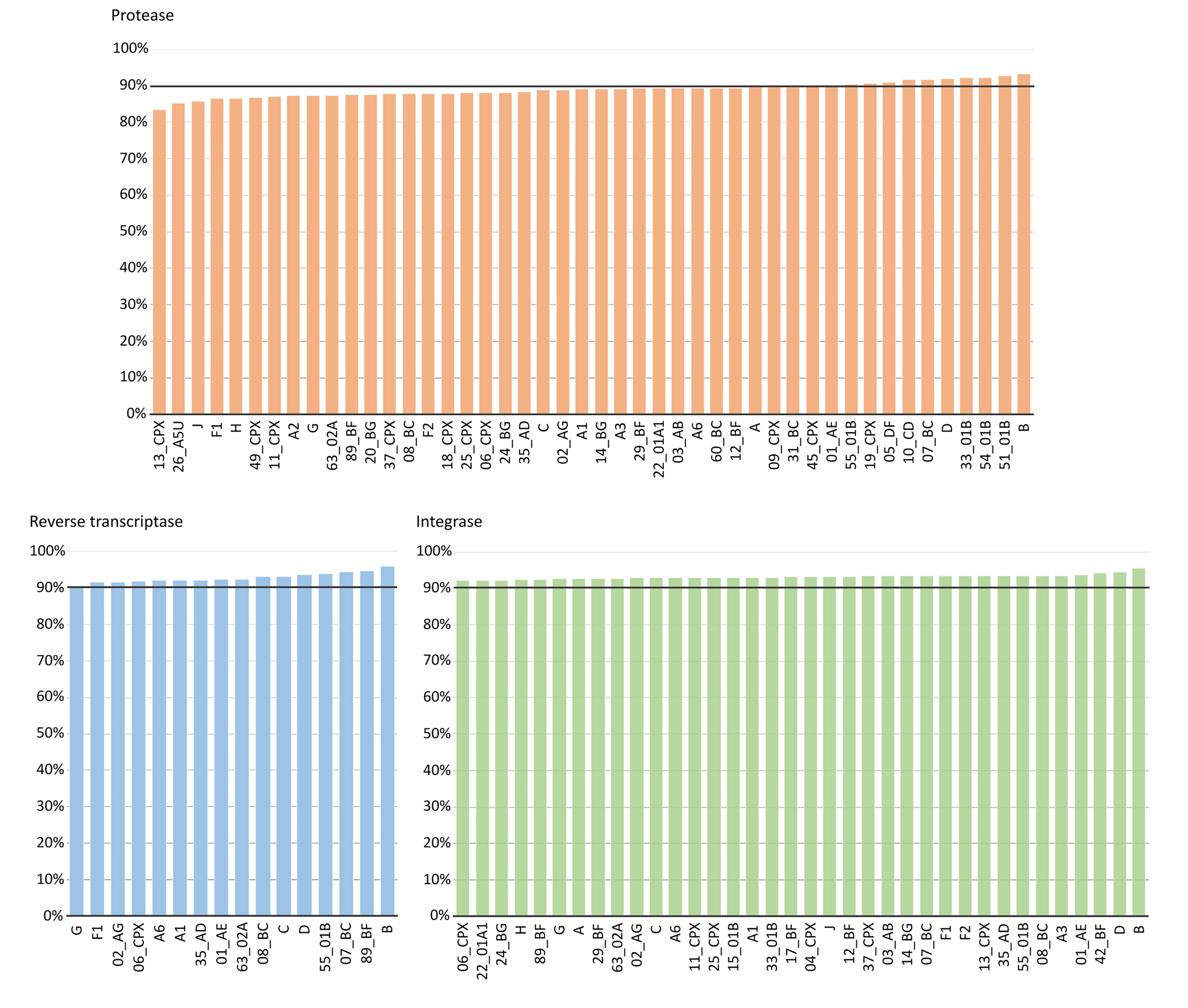


**FIGURE 6 |** Percentage of aa conservation of PR, RT, and IN across the HIV-1 group M variants with >5 sequences at LANL. *X*-axis: HIV-1 group M variants with >5 available sequences at LANL (46 in PR, 16 in RT, and 36 in IN). *Y*-axis: conservation rate for each variant included in this analysis. The horizontal line represents 90% conservation.

can be considered as genetic markers of each HIV-1 variant (V-markers) and those that correspond to major or minor DRM (R-markers) across HIV-1 groups, and group M subtypes and recombinants. We also present the consensus PR, RT, and IN sequences for HIV-1, HIV-1 groups, and variants, and the Wu–Kabat variability coefficient for group M in the three studied Pol proteins. This information is helpful to improve the understanding of structural, functional, and immunogenic differences across HIV-1 groups, subtypes, sub-subtypes, and recombinants and their impact on drug susceptibility and resistance pathways (Nagata et al., 2017; Sliepen et al., 2019; Zhang et al., 2021).

In previous studies, the variability in Pol proteins was low but slightly higher in PR compared with RT and IN (Turner et al., 2004; Rhee et al., 2016). In this study, the mean aa conservation of group M consensus sequences was high (>90%) in the three studied proteins, being slightly lower for PR (93% vs. 94% in RT and 96% in IN). As expected, HIV-1 consensus sequences were slightly less conserved (>80%) as non-M groups were included in the consensus. Still, the conservation rate followed the same order as in group M consensus, with less conservation in PR (83% vs. 88% in RT and 96% in IN). We also observed a low variability (below 0.5%) in the residues involved in binding or catalytic sites after testing all the available Pol sequences, highlighting the fragility of these important functional sites (**Table 1**).

HIV-1 variants have different global prevalence (Hemelaar et al., 2019) and levels of HIV-1 genetic diversity (Abecasis et al., 2009). HIV-1 group M subtype C is the most prevalent variant in the ongoing HIV pandemic, causing around 50% of worldwide infections (Hemelaar et al., 2019). In addition, subtype

**TABLE 2 |** Single V-markers and R-markers in protease found across HIV-1 variants with >5 LANL sequences.

| HIV-1 variant | Countries | V-markers and R-markers (bold red) |
|---|---|---|
| Group O (24) | Belgium (2), Cameroon (12), Senegal (3), Spain (3), United Kingdom (2), United States (2) | T4P (100%), Q7D (100%), I13A (100%), Q18H (92%), K20C (100%), A22V (92%), E35N (100%), S37Q (92%), P39E (100%), **K43T (88%)**, **Q58E (100%)**, **H69R (83%)**, K70E (83%), **A71V (100%)**, Q92G (100%) |
| Subtype J (20) | Republic of Angola (2), The Democratic Republic of the Congo (2), Central African Republic (3), Congo (1), Cameroon (7), Gabon (1), Spain (1), Senegal (1), Belgium (2) | Q61E (83%) |
| CRF08_BC (431) | China (430), India (1) | T12S (87%) |
| CRF13_cpx (42) | Belgium (2), Burkina Faso (1), Cameroon (22), Central African Republic (5), Germany (5), Greenland (1), Poland (1), Spain (2), The Democratic Republic of the Congo (3) | G17E (76%), E34K (90%), I62M (76%), **V77I (98%)** |
| CRF19_cpx (178) | Cuba (172), Spain (4), Tunisia (1), United Kingdom (1) | H69Q (75%) |
| CRF35_A1D (216) | Afghanistan (9), China (1), Iran (205), Romania (1) | L19Q (76%) |
| CRF49_cpx (10) | Botswana (1), Gambia (3), Germany (4), Nigeria (1), Senegal (1) | D60N (90%), Q61D (90%), I66V (80%) |
| CRF51_01B (8) | Singapore (8) | L63S (100%) |
| CRF60_BC (25) | Brazil (1), Germany (2), Italy (22) | **L10V (95%)**, M36T (91%) |
| CRF63_02A6 (193) | Kyrgyzstan (2), Russian Federation (59), Uzbekistan (132) | K14R (79%) |
| CRF89_BF1 (22) | Argentina (1), Spain (20), Sweden (1) | T12E (91%) |

*In brackets, number of sequences in variant or country and conservation percentage in markers; in bold red font, V-markers that are R-markers. None of the R-markers corresponded to major DRM to PI according to Stanfordv 9.0.*

**TABLE 3 |** Single V-markers and R-markers in reverse transcriptase found across HIV-1 variants with >5 LANL sequences.

| HIV-1 variant | Countries | V-markers and R-markers (bold red) |
|---|---|---|
| Sub-subtype A6 (85) | United Kingdom (4), Georgia (1), Italy (1), Russian Federation (79) | K11T (85%), E36D (76%) |
| Subtype C (1607) | Belgium (4), Brazil (12), Botswana (30), The Democratic Republic of the Congo (4), China (11), Spain (5), United Kingdom (14), Georgia (1), Equatorial Guinea (1), India (65), Kenya (4), Nigeria (1), Nepal (2), Pakistan (1), Rwanda (3), Sweden (9), Senegal (2), Thailand (1), United Republic of Tanzania (37), Uganda (7), United States (5), South Africa (933), Zambia (455) | T39E (78%) |
| Subtype D (238) | The Democratic Republic of the Congo (2), United Kingdom (2), Kenya (1), South Korea (1), Nigeria (1), United Republic of Tanzania (1), Uganda (229), United States (1) | L282C (91%), P345Q (89%), T377Q (92%), S379C (82%) |
| Sub-subtype F1 (41) | Germany (1), Spain (22), France (1), United Kingdom (14), Italy (2), United States (1) | D123E (88%), I178L (85%) |
| Subtype G (40) | The Democratic Republic of the Congo (2), Cameroon (1), Spain (5), United Kingdom (2), Kenya (1), Nigeria (27), Russian Federation (1), South Africa (1) | M357R (93%), Q394R (90%), T400V (83%), F440Y (93%) |
| CRF01_AE (1225) | Afghanistan (1), China (547), Sweden (3), Thailand (221), United Kingdom (13), United States (3), Viet Nam (437) | V245E (91%) |
| CRF02_AG (110) | Belgium (1), Benin (2), Cameroon (5), China (2), Spain (5), Gabon (1), United Kingdom (29), Equatorial Guinea (5), Italy (1), South Korea (1), Mexico (1), Nigeria (33), Pakistan (3), Russian Federation (1), Sweden (2), Senegal (9), Togo (7), Thailand (1), United States (1) | S162A (96%) |
| CRF06_cpx (19) | Burkina Faso (3), China (1), United Kingdom (10), Nigeria (4), Senegal (1) | F346H (84%), R358K (89%) |
| CRF08_BC (129) | China (129) | E53D (98%), D324E (86%) |
| CRF35_A1D (9) | Afghanistan (9) | L283I (78%) |
| CRF55_01B (11) | China (11) | T39K (90%), K43E (100%), I135R (100%), **V179E (100%)**, V245A (82%), K388R (91%) |
| CRF89_BF (7) | Spain (7) | Q394L (86%), E399D (100%) |

*In brackets, number of sequences in variant or country and conservation percentage in markers; in bold red font, V-markers that are R-markers. None of the R-markers corresponded to major DRM to NRTI or NNRTI according to Stanford v9.0.*

**TABLE 4** | Single V-markers and R-markers in integrase found across HIV-1 variants with >5 LANL sequences.

| HIV-1 variant | Countries | V-markers and R-markers (bold red) |
|---|---|---|
| Group N (12) | Cameroon (11), France (1) | D55N (100%), V165I (92%), K215T (100%), T218L (100%), I220V (83%), D279G (100%) |
| Group O (50) | Cameroon (30), Belgium (1), France (13), United States (3), Senegal (3) | D3E (90%), K7Q (100%), M22L (100%), N27G (100%), D41P (100%), Q44H (98%), L45I (88%), G59E (96%), Y83F (98%), G106A (100%), V126M (100%), Q137H (94%), S153A (96%), K160S (98%), G163Q (90%), I182V (90%), I204L (100%), D207Q (80%), K211T (100%), K240Q (100%), N254K (98%), C280S (86%), V281M (84%), D286T (92%), D288S (94%) |
| Subtype C (1353) | Cameroon (1), Ethiopia (2), Kenya (6), Malawi (1), Mozambique (1), Rwanda (3), Somalia (1), United Republic of Tanzania (57), Uganda (7), Zambia (471), Poland (1), Belgium (7), Denmark (1), Spain (5), Sweden (20), United States (2), Botswana (52), South Africa (623), Argentina (1), Brazil (7), Uruguay (1), Senegal (2), China (4), Cyprus (4), Georgia (1), India (43), Myanmar (1), Israel (5), Nepal (3), Pakistan (1), Saudi Arabia (14), Tajikistan (2), Thailand (1), Unknown (1), Yemen (1) | D25E (79%) |
| Subtype H (11) | Belgium (3), Cameroon (1), Central African Republic (2), The Democratic Republic of the Congo (4), United Kingdom (1) | N222K (91%) |
| Subtype J (10) | Republic of Angola (1), Belgium (3), Cameroon (1), The Democratic Republic of the Congo (3), Sweden (2) | Y99F (80%) |
| CRF03_A6B (8) | Belarus (1), Russian Federation (3), Tajikistan (4) | **E157Q (100%)**, K160Q (100%) |
| CRF06_cpx (103) | Australia (1), Cameroon (2), Estonia (95), Ghana (1), Mali (2), Russian Federation (1), Senegal (1) | L63I (91%) |
| CRF07_BC (324) | China (321), South Korea (1), Taiwan (1), Viet Nam (1) | I84M (91%) |
| CRF08_BC (36) | China (36) | K211R (92%) |
| CRF22_01A1 (14) | Cameroon (14) | A23V (100%) |
| CRF33_01B (6) | Indonesia (2), Malaysia (4) | L63V (83%) |
| CRF35_A1D (13) | Afghanistan (13) | I60M (100%), V126F (100%), G134S (77%) |
| CRF42_BF1 (17) | Luxembourg (17) | L28I (100%), S39C (100%), G163E (100%) |

*In brackets, number of sequences in variant or country and conservation percentage in markers; in bold red font, V-markers that are R-markers. None of the R-markers corresponded to major DRM to INSTI according to Stanford v9.0.*

C is the most prevalent variant in Southern Africa and India; subtype A in some countries of Eastern Africa, Russia, and Eastern Europe; subtype B in the rest of Europe, the Americas, and Oceania; CRF01_AE in Asia; and CRF02_AG in Western Africa (Bbosa et al., 2019). However, HIV genomic sequencing is more widespread in economically developed nations, which explains that in our Pol dataset the most represented HIV-1 variant was subtype B, despite the fact that this variant only causes around 12% of the 38 million infections globally (Hemelaar et al., 2019), followed by the most abundant variant subtype C and recombinant CRF01_AE with the highest number of sequences belonging to China and the United States, according to the sequence availability in LANL. The main limitation in this study is the low number of sequences available in LANL for some non-B subtypes and CRF (Hemelaar et al., 2019), due to their low prevalence in the pandemic or because they are circulating in areas with none or scarce HIV sequencing.

Across group M variants with >5 available sequences in LANL, subtype B was the most conserved variant (>92%) in the three Pol proteins as expected, since the reference strain for the alignments was the subtype B HXB2 isolate. Again, PR showed slightly greater variability: While in RT and IN all the included variants had a conservation >90%, in PR only 24% of the variants were conserved in >90% of their sequences.

The Wu–Kabat protein variability coefficient (WK) was analyzed in PR, RT, and IN group M to study the susceptibility of each aa position to evolutionary replacements. The median variability coefficient in PR (WK10) was higher than that in IN (WK5) and RT (WK4). All 99 PR residues presented some degree of variability as none had a coefficient of 1. PR also presented the site with the highest WK value (52 in residue 63). Most IN (92%) and RT (88%) sites showed a WK below 10, while almost half (48%) of PR residues had a WK between 11 and 20, being the Pol protein with more sites prone to evolutionary replacements.

Although similar mutations occur in subtype B and non-subtype-B viruses and drug resistance evolution is comparable in both groups, subtype-specific mutation rates have been identified, with differences that could affect genotypic interpretation and DRM monitoring (Kantor et al., 2005; Kantor, 2006; Yebra et al., 2010). We found 31 total V-markers in PR, 28 V-markers in RT,

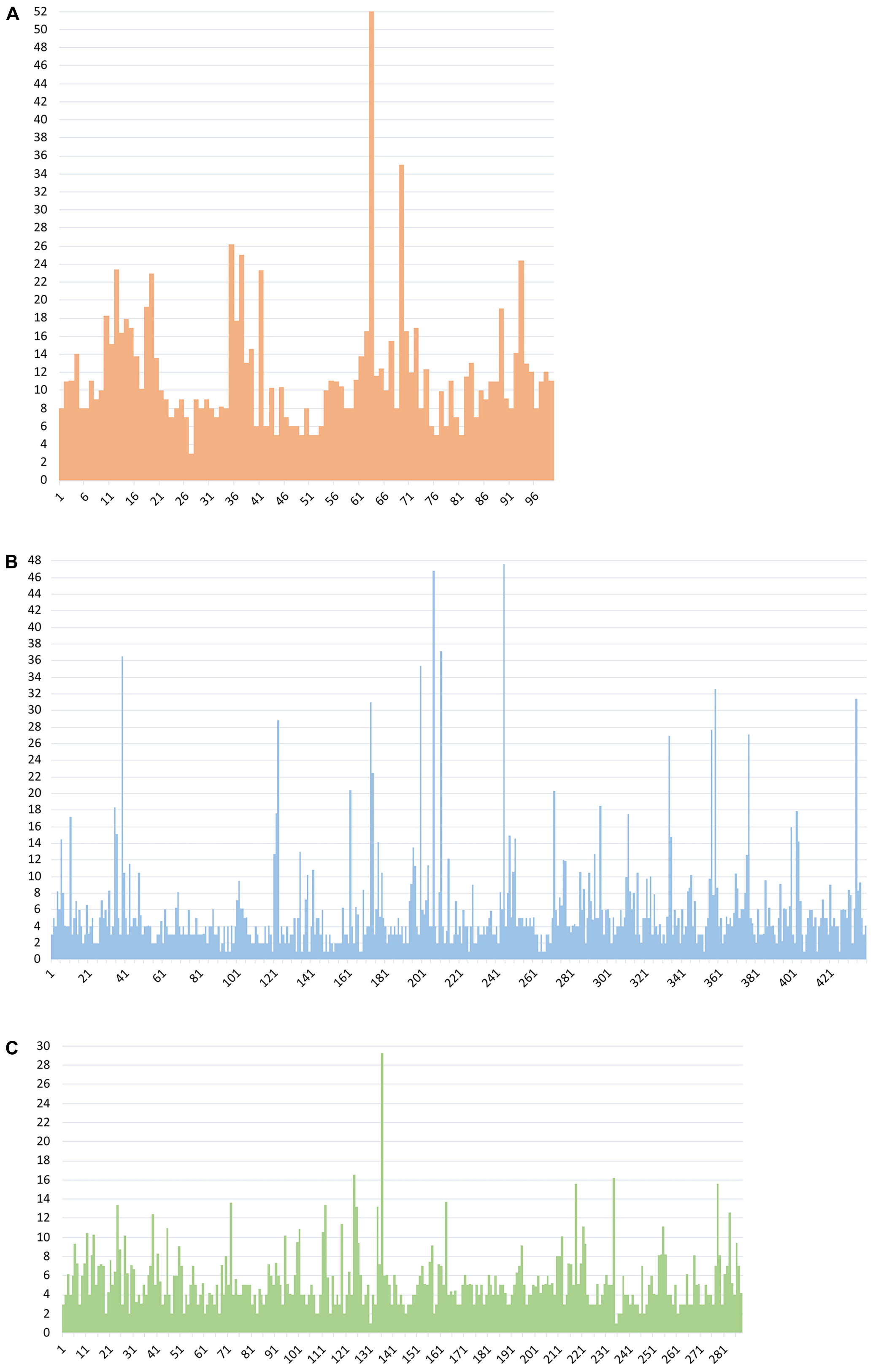


**FIGURE 7 |** Wu–Kabat variability coefficient plot of PR, RT, and IN group M sequences. **(A)** Wu–Kabat variability coefficient plot of PR (99 aa). **(B)** Wu–Kabat variability coefficient plot of RT (440 aa). **(C)** Wu–Kabat variability coefficient plot of IN (288 aa). *X*-axis, amino acid position; *Y*-axis, WK variability coefficient.

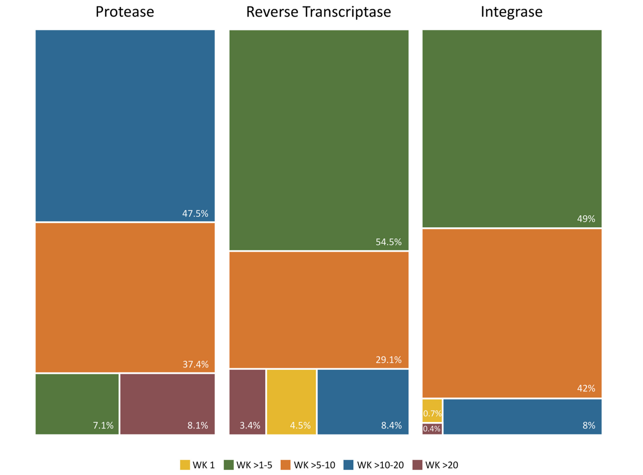


**FIGURE 8 |** Proportion of Wu–Kabat variability coefficient values in PR, RT, and IN residues. Each box represents the proportion of residues within each protein that present a Wu–Kabat coefficient value within the range indicated beneath the figure and colored accordingly. Protease (99 aa), reverse transcriptase (440 aa), integrase (288 aa). WK, WK variability coefficient.

and 47 in IN. Only eight were R-markers; none considered major DRM, being minor DRM with low impact in ARV susceptibility. In a previous study on HIV-2 variability (Troyano-Hernáez et al., 2021a), the R-markers corresponding to DRM to PI, NRTI, and INSTIs appeared not to have a significant impact on ARV susceptibility as well. However, HIV-2 presents natural polymorphisms related to drug resistance that make it naturally resistant to NNRTI, certain PI, and fusion inhibitor enfuvirtide (Tuaillon et al., 2004; Desbois et al., 2008; Menéndez-Arias and Álvarez, 2014).

Six R-markers were found in PR (K43T/Q58E/H69R/A71V in group O, V77I in CRF13_cpx, and L10V in CRF60_BC), one in RT (V179E in CRF55_01B), and one in IN (E157Q in CRF03_A6B). None of the R-markers in PR's group O conferred intermediate or high-level resistance to PI alone or combined. K43T and Q58E are accessory non-polymorphic mutations that confer potential or low-level resistance to nelfinavir (NFV) and tipranavir/ritonavir (TPV/r) and other PIs (Rhee et al., 2003, 2010; Baxter et al., 2006; Bennett et al., 2009). H69R is a minor mutation affecting TPV/r according to IAS (Wensing et al., 2019) (not included in Stanford). A71V is a polymorphic accessory mutation associated with an increase of viral replication in the presence of other PI resistance mutations (Nijhuis et al., 1999; Rhee et al., 2003). V77I was present in 96.6% of CRF13_cpx PR sequences and is considered a minor mutation affecting indinavir/ritonavir according to IAS2019 (Wensing et al., 2019) (not included in Stanford). L10V was found in 95.5% of CRF60_BC PR sequences, being a polymorphic accessory mutation that may reduce PI susceptibility or increase the replication of viruses containing PI resistance mutations (Stanford University, 2022). Regarding the R-marker found in RT, V179E is a DRM to NNRTI, considered a non-polymorphic accessory mutation associated with potential low-level resistance to efavirenz (EFV), etravirine (ETR), nevirapine (NVP), and rilpivirine (RPV) (Rhee et al., 2003; Tambuyzer et al., 2009). As for the R-marker present in IN, E157Q is an accessory mutation with little effect by itself on the response to INSTI therapy, conferring potential low-level resistance to elvitegravir (EVG) and raltegravir (RAL) (Anstett et al., 2016; Stanford University, 2022; Charpentier et al., 2018).

Similarly, when analyzing the seven DRM found in ≥25% of the sequences in variants with >5 available PR, RT, or IN sequences, none corresponded to major DRM. Three were R-markers: the previously described accessory DRM to PI, Q58E and K43T, and the accessory DRM to NNRTI V179E. In RT, we identified another accessory DRM to NNRTI, V106I, in sub-subtype F1. This mutation is present in 1–2% of naïve patients and contributes to reduced NNRTI susceptibility combined with other mutations, such as V179D, not found in the available F1 sequences (Rhee et al., 2003; Gatanaga et al., 2010). Alone it has little effect on NNRTI susceptibility conferring potential low-level resistance to doravirine (DOR), ETR, NVP, and RPV (Release Notes—HIV Drug Resistance Database). A62V (sub-subtype A6) is an accessory mutation that often occurs together with the multi-NRTI resistance mutations K65R or Q151M (Svarovskaia et al., 2008). However, these mutations were not found in this subtype among our sequence sets. A62V is widespread in subtype A viruses belonging to the former Soviet Union countries but is otherwise non-polymorphic (Carr et al., 2005). Two accessory DRM to INSTI were detected in IN: G163R, found in two CRFs, and M50I, in six variants. G163R is a non-polymorphic mutation that confers low-level resistance to EVG and RAL and usually appears in combination with N155H (Gatell et al., 2010; Stanford University, 2022), not found in these CRFs. M50I is a polymorphic mutation that may reduce dolutegravir (DTG) susceptibility in combination with R263K (Wares et al., 2014), absent in all the variants carrying M50I.

We present a thorough descriptive analysis of Pol variability among all HIV-1 variants circulating to date. The relatively high aa conservation observed in Pol proteins across HIV-1 variants highlights their critical role in the viral cycle. The variant-specific polymorphisms (V-markers) found in Pol presented little or no predicted impact on clinical ARV efficacy. Our data support previous studies reporting limited evidence of associations between HIV-1 subtypes and treatment failure (Rockstroh et al., 2011; Poon et al., 2019). However, it has been reported that some natural polymorphisms in Pol can promote alternative resistance pathways (Kantor and Katzenstein, 2003; Holguín et al., 2006b; Sanches et al., 2007; Sánchez et al., 2020), affect inhibitor binding (Tran et al., 2020), be present in Pol epitopes interacting with the immune system (Los Alamos National Laboratory, 2022a), or affect protein structure and conformation (Bandaranayake et al., 2008; Coman et al., 2008a,b; Kear et al., 2009). For example, some HIV-1 variants in our study (group O, subtype J, CRF13_cpx, 19_cpx, 49_cpx, and 51_01B, see **Table 2**) presented V-markers within the PR flaps (PR residues 37-71), the regions mediating accessibility of substrate to the PR active site (Hornak et al., 2006). However, the impact of these V-markers on the flap conformational changes of the corresponding variant is still unknown. Further research is required to evaluate the impact of the different levels of aa conservation in the PR, RT, and IN across HIV-1 variants and to evaluate the influence of each specific V-markers found at Pol in the viral replication cycle, protein structure, and function, as well as in the interactions with antiretroviral drugs or with the immune system.

## DATA AVAILABILITY STATEMENT

The original contributions presented in this study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author. The datasets analyzed for this study can be found in the Los Alamos National Laboratory database (https://www.hiv.lanl.gov).

## ETHICS STATEMENT

The viral sequences were retrieved from public databases, and no human studies or animal studies were performed in this manuscript.

## AUTHOR CONTRIBUTIONS

PT-H analyzed the HIV Pol LANL sequences, validated some EpiMolBio functions necessary for the sequences analyses, performed the computations, discussed results, and wrote the first version of the manuscript. RR downloaded and aligned the HIV Pol LANL sequences, developed the in-house EpiMolBio bioinformatics program, and validated the EpiMolBio functions necessary for the sequences analyses. AH designed and supervised the study, discussed results, reviewed and edited the manuscript, and applied for funding, being responsible for project administration. All authors approved the submitted final version.

## FUNDING

This research was supported by Instituto de Salud Carlos III (PI18/00904 Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016) and co-financed by the European Regional Development Fund "A way to achieve Europe" (ERDF). This study was also included in the "Subprograma de Inmigración y Salud" from CIBERESP (Spain). PT-H was funded by ISCIII-Programa Estatal de Promoción del Talento-AES Río Hortega exte. CM19/00057. RR was funded by FONDOS FUR 2020/0285. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.866705/full#supplementary-material

## REFERENCES

Abecasis, A. B., Vandamme, A.-M., and Lemey, P. (2009). Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J. Virol.* 83, 12917–12924. doi: 10.1128/JVI.01022-09

Anstett, K., Cutillas, V., Fusco, R., Mesplède, T., and Wainberg, M. A. (2016). Polymorphic substitution E157Q in HIV-1 integrase increases R263K-mediated dolutegravir resistance and decreases DNA binding activity. *J. Antimicrob. Chemother.* 71, 2083–2088. doi: 10.1093/jac/dkw109

Baeten, J. M., Chohan, B., Lavreys, L., Chohan, V., McClelland, R. S., Certain, L., et al. (2007). HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. *J. Infect. Dis.* 195, 1177–1180. doi: 10.1086/512682

Bandaranayake, R. M., Prabu-Jeyabalan, M., Kakizawa, J., Sugiura, W., and Schiffer, C. A. (2008). Structural analysis of human immunodeficiency virus type 1 CRF01_AE protease in complex with the substrate p1-p6. *J. Virol.* 82, 6762–6766. doi: 10.1128/JVI.00018-08

Baxter, J. D., Schapiro, J. M., Boucher, C. A. B., Kohlbrenner, V. M., Hall, D. B., Scherer, J. R., et al. (2006). Genotypic changes in human immunodeficiency virus type 1 protease associated with reduced susceptibility and virologic response to the protease inhibitor tipranavir. *J. Virol.* 80, 10794–10801. doi: 10.1128/JVI.00712-06

Bbosa, N., Kaleebu, P., and Ssemwanga, D. (2019). HIV subtype diversity worldwide. *Curr. Opin. HIV AIDS* 14, 153–160. doi: 10.1097/COH.0000000000000534

Bebenek, K., Abbotts, J., Wilson, S. H., and Kunkel, T. A. (1993). Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots. *J. Biol. Chem.* 268, 10324–10334.

Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., et al. (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* 4:e4724. doi: 10.1371/journal.pone.0004724

Burgos, M., Llácer, T., Reinosa, R., Rubio-Garrido, M., González, A., and Holguín, A. (2019). "Impaired genotypic resistance interpretation due to HIV-1 variant specific Markers," in *Proceedings of the 10th IAS Conference on HIV Science*, Ciudad de México.

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G. M., and Gronenborn, A. M. (1997). Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nat. Struct. Biol.* 4, 567–577. doi: 10.1038/nsb0797-567

Carr, J. K., Nadai, Y., Eyzaguirre, L., Saad, M. D., Khakimov, M. M., Yakubov, S. K., et al. (2005). Outbreak of a West African recombinant of HIV-1 in Tashkent, Uzbekistan. *J. Acquir. Immune Defic. Syndr.* 39, 570–575.

Charpentier, C., Malet, I., Andre-Garnier, E., Storto, A., Bocket, L., Amiel, C., et al. (2018). Phenotypic analysis of HIV-1 E157Q integrase polymorphism and impact on virological outcome in patients initiating an integrase inhibitor-based regimen. *J. Antimicrob. Chemother.* 73, 1039–1044. doi: 10.1093/jac/dkx511

Clarke, J. R. (2002). Molecular diagnosis of HIV. *Expert Rev. Mol. Diagn.* 2, 233–239. doi: 10.1586/14737159.2.3.233

Coman, R. M., Robbins, A. H., Fernandez, M. A., Gilliland, C. T., Sochet, A. A., Goodenow, M. M., et al. (2008a). The contribution of naturally occurring polymorphisms in altering the biochemical and structural characteristics of HIV-1 subtype C protease. *Biochemistry* 47, 731–743. doi: 10.1021/bi7018332

Coman, R. M., Robbins, A. H., Goodenow, M. M., Dunn, B. M., and McKenna, R. (2008b). High-resolution structure of unbound human immunodeficiency virus 1 subtype C protease: implications of flap dynamics and drug resistance. *Acta Crystallogr. D Biol. Crystallogr. D* 64, 754–763. doi: 10.1107/S090744490801278X

Dam, E., Quercia, R., Glass, B., Descamps, D., Launay, O., Duval, X., et al. (2009). Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog.* 5:e1000345. doi: 10.1371/journal.ppat.1000345

De Leys, R., Vanderborght, B., Vanden Haesevelde, M., Heyndrickx, L., van Geel, A., Wauters, C., et al. (1990). Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J. Virol.* 64, 1207–1216. doi: 10.1128/JVI.64.3.1207-1216.1990

Desbois, D., Roquebert, B., Peytavin, G., Damond, F., Collin, G., Bénard, A., et al. (2008). In vitro phenotypic susceptibility of human immunodeficiency virus type 2 clinical isolates to protease inhibitors. *Antimicrob. Agents Chemother.* 52, 1545–1548. doi: 10.1128/AAC.01284-07

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R., and Davies, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science* 266, 1981–1986. doi: 10.1126/science.7801124

Eijkelenboom, A. P., Sprangers, R., Hård, K., Puras Lutzke, R. A., Plasterk, R. H., Boelens, R., et al. (1999). Refined solution structure of the C-terminal DNA-binding domain of human immunovirus-1 integrase. *Proteins* 36, 556–564.

El Safadi, Y., Vivet-Boudou, V., and Marquet, R. (2007). HIV-1 reverse transcriptase inhibitors. *Appl. Microbiol. Biotechnol.* 75, 723–737. doi: 10.1007/s00253-007-0919-7

Engelman, A. N., and Singh, P. K. (2018). Cellular and molecular mechanisms of HIV-1 integration targeting. *Cell. Mol. Life Sci.* 75, 2491–2507. doi: 10.1007/s00018-018-2772-5

Eron, J. J. J. (2000). HIV-1 protease inhibitors. *Clin. Infect. Dis.* 30(Suppl. 2), S160–S170. doi: 10.1086/313853

Frankel, A. D., and Young, J. A. (1998). HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.* 67, 1–25. doi: 10.1146/annurev.biochem.67.1.1

Gatanaga, H., Ode, H., Hachiya, A., Hayashida, T., Sato, H., and Oka, S. (2010). Combination of V106I and V179D polymorphic mutations in human immunodeficiency virus type 1 reverse transcriptase confers resistance to efavirenz and nevirapine but not etravirine. *Antimicrob. Agents Chemother.* 54, 1596–1602. doi: 10.1128/AAC.01480-09

Gatell, J. M., Katlama, C., Grinsztejn, B., Eron, J. J., Lazzarin, A., Vittecoq, D., et al. (2010). Long-term efficacy and safety of the HIV integrase inhibitor raltegravir in patients with limited treatment options in a Phase II study. *J. Acquir. Immune Defic. Syndr.* 53, 456–463. doi: 10.1097/qai.0b013e3181c9c967

Goodenow, M. M., Bloom, G., Rose, S. L., Pomeroy, S. M., O'Brien, P. O., Perez, E. E., et al. (2002). Naturally occurring amino acid polymorphisms in human immunodeficiency virus type 1 (HIV-1) Gag p7(NC) and the C-cleavage site impact Gag-Pol processing by HIV-1 protease. *Virology* 292, 137–149. doi: 10.1006/viro.2001.1184

Gu, S.-X., Zhu, Y.-Y., Wang, C., Wang, H.-F., Liu, G.-Y., Cao, S., et al. (2020). Recent discoveries in HIV-1 reverse transcriptase inhibitors. *Curr. Opin. Pharmacol.* 54, 166–172. doi: 10.1016/j.coph.2020.09.017

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* 18, 182–192. doi: 10.1016/j.molmed.2011.12.001

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Fleminger, I., Kirtley, S., et al. (2019). Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. *Lancet Infect. Dis.* 19, 143–155. doi: 10.1016/S1473-3099(18)30647-9

Holguín, A., Alvarez, A., and Soriano, V. (2006a). Variability in the P6gag domains of HIV-1 involved in viral budding. *AIDS* 20, 624–627. doi: 10.1097/01.aids.0000210619.75707.21

Holguín, A., Ramirez de Arellano, E., Rivas, P., and Soriano, V. (2006b). Efficacy of antiretroviral therapy in individuals infected with HIV-1 non-B subtypes. *AIDS Rev.* 8, 98–107.

Holguín, A., Paxinos, E., Hertogs, K., Womac, C., and Soriano, V. (2004). Impact of frequent natural polymorphisms at the protease gene on the in vitro susceptibility to protease inhibitors in HIV-1 non-B subtypes. *J. Clin. Virol.* 31, 215–220. doi: 10.1016/j.jcv.2004.03.015

Holguín, A., and Soriano, V. (2002). Resistance to antiretroviral agents in individuals with HIV-1 non-B subtypes. *HIV Clin. Trials* 3, 403–411. doi: 10.1310/7bwp-0x7f-nxna-qrnp

Hornak, V., Okur, A., Rizzo, R. C., and Simmerling, C. (2006). HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* 103, 915–920. doi: 10.1073/pnas.0508452103

Hu, W.-S., and Hughes, S. H. (2012). HIV-1 reverse transcription. *Cold Spring Harb. Perspect. Med.* 2:a006882. doi: 10.1101/cshperspect.a006882

Huff, J. R. (1991). HIV protease: a novel chemotherapeutic target for AIDS. *J. Med. Chem.* 34, 2305–2314. doi: 10.1021/jm00112a001

Jóźwik, I. K., Passos, D. O., and Lyumkis, D. (2020). Structural biology of HIV integrase strand transfer inhibitors. *Trends Pharmacol. Sci.* 41, 611–626. doi: 10.1016/j.tips.2020.06.003

Kabat, E. A., Wu, T. T., and Bilofsky, H. (1977). Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *J. Biol. Chem.* 252, 6609–6616.

Kantor, R. (2006). Impact of HIV-1 pol diversity on drug resistance and its clinical implications. *Curr. Opin. Infect. Dis.* 19, 594–606. doi: 10.1097/QCO.0b013e3280109122

Kantor, R., and Katzenstein, D. (2003). Polymorphism in HIV-1 non-subtype B protease and reverse transcriptase and its potential impact on drug susceptibility and drug resistance evolution. *AIDS Rev.* 5, 25–35.

Kantor, R., Katzenstein, D. A., Efron, B., Carvalho, A. P., Wynhoven, B., Cane, P., et al. (2005). Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. *PLoS Med.* 2:e112. doi: 10.1371/journal.pmed.0020112

Kear, J. L., Blackburn, M. E., Veloro, A. M., Dunn, B. M., and Fanucci, G. E. (2009). Subtype polymorphisms among HIV-1 protease variants confer altered flap conformations and flexibility. *J. Am. Chem. Soc.* 131, 14650–14651. doi: 10.1021/ja907088a

Konvalinka, J., Kräusslich, H.-G., and Müller, B. (2015). Retroviral proteases and their roles in virion maturation. *Virology* 479-480, 403–417. doi: 10.1016/j.virol.2015.03.021

Li, G., Verheyen, J., Rhee, S.-Y., Voet, A., Vandamme, A.-M., and Theys, K. (2013). Functional conservation of HIV-1 Gag: implications for rational drug design. *Retrovirology* 10:126. doi: 10.1186/1742-4690-10-126

Linchangco, G. V. J., Foley, B., and Leitner, T. (2021). Updated HIV-1 consensus sequences change but stay within similar distance from worldwide samples. *Front. Microbiol.* 12:828765. doi: 10.3389/fmicb.2021.828765

Llacer Delicado, T., Torrecilla, E., and Holguin, A. (2016). Deep analysis of HIV-1 natural variability across HIV-1 variants at residues associated with integrase inhibitor (INI) resistance in INI-naive individuals. *J. Antimicrob. Chemother.* 71, 362–366. doi: 10.1093/jac/dkv333

Los Alamos National Laboratory (2021). *HIV Circulating Recombinant Forms (CRFs)*. Available Online at: https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html [accessed January 10, 2022].

Los Alamos National Laboratory (2022a). *Interactive Epitope Maps*. Available Online at: https://www.hiv.lanl.gov/content/immunology/maps/maps.html [accessed May 24, 2022].

Los Alamos National Laboratory (2022b). *Main Search Interface of HIV Sequence Database*. Available Online at: https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html [accessed May 24, 2022].

Menéndez-Arias, L., and Álvarez, M. (2014). Antiretroviral therapy and drug resistance in human immunodeficiency virus type 2 infection. *Antiviral Res.* 102, 70–86. doi: 10.1016/j.antiviral.2013.12.001

Myint, L., Matsuda, M., Matsuda, Z., Yokomaku, Y., Chiba, T., Okano, A., et al. (2004). Gag non-cleavage site mutations contribute to full recovery of viral fitness in protease inhibitor-resistant human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.* 48, 444–452. doi: 10.1128/AAC.48.2.444-452.2004

Nagata, S., Imai, J., Makino, G., Tomita, M., and Kanai, A. (2017). Evolutionary analysis of HIV-1 pol proteins reveals representative residues for viral subtype differentiation. *Front. Microbiol.* 8:2151. doi: 10.3389/fmicb.2017.02151

Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., et al. (1989). Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* 337, 615–620. doi: 10.1038/337615a0

Ng, O. T., Laeyendecker, O., Redd, A. D., Munshaw, S., Grabowski, M. K., Paquet, A. C., et al. (2014). HIV type 1 polymerase gene polymorphisms are associated with phenotypic differences in replication capacity and disease progression. *J. Infect. Dis.* 209, 66–73. doi: 10.1093/infdis/jit425

NIH FDA-Approved HIV Medicines (2022). Available Online at: https://hivinfo.nih.gov/understanding-hiv/fact-sheets/fda-approved-hiv-medicines [accessed January 20, 2022].

Nijhuis, M., Schuurman, R., de Jong, D., Erickson, J., Gustchina, E., Albert, J., et al. (1999). Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS* 13, 2349–2359. doi: 10.1097/00002030-199912030-00006

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nat. Med.* 15, 871–872. doi: 10.1038/nm.2016

Poon, A. F. Y., Ndashimye, E., Avino, M., Gibson, R., Kityo, C., Kyeyune, F., et al. (2019). First-line HIV treatment failures in non-B subtypes and recombinants: a cross-sectional analysis of multiple populations in Uganda. *AIDS Res. Ther.* 16:3. doi: 10.1186/s12981-019-0218-2

Release Notes - HIV Drug Resistance Database (2020). Available Online at: https://hivdb.stanford.edu/page/release-notes/ (accessed May 20, 2020).

Renjifo, B., Gilbert, P., Chaplin, B., Msamanga, G., Mwakagile, D., Fawzi, W., et al. (2004). Preferential in-utero transmission of HIV-1 subtype C as compared to

HIV-1 subtype A or D. *AIDS* 18, 1629–1636. doi: 10.1097/01.aids.0000131392.68597.34

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Rhee, S.-Y., Sankaran, K., Varghese, V., Winters, M. A., Hurt, C. B., Eron, J. J., et al. (2016). HIV-1 protease, reverse transcriptase, and integrase variation. *J. Virol.* 90, 6058–6070. doi: 10.1128/JVI.00495-16

Rhee, S.-Y., Taylor, J., Fessel, W. J., Kaufman, D., Towner, W., Troia, P., et al. (2010). HIV-1 protease mutations and protease inhibitor cross-resistance. *Antimicrob. Agents Chemother.* 54, 4253–4261. doi: 10.1128/AAC.00574-10

Rice, P., Craigie, R., and Davies, D. R. (1996). Retroviral integrases and their cousins. *Curr. Opin. Struct. Biol.* 6, 76–83. doi: 10.1016/s0959-440x(96)80098-4

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science* 242, 1171–1173. doi: 10.1126/science.2460925

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., et al. (2000). HIV-1 nomenclature proposal. *Science* 288, 55–56. doi: 10.1126/science.288.5463.55d

Rockstroh, J. K., Teppler, H., Zhao, J., Sklar, P., Miller, M. D., Harvey, C. M., et al. (2011). Clinical efficacy of raltegravir against B and non-B subtype HIV-1 in phase III clinical studies. *AIDS* 25, 1365–1369. doi: 10.1097/QAD.0b013e328348065a

Rodgers, D. W., Gamblin, S. J., Harris, B. A., Ray, S., Culp, J. S., Hellmig, B., et al. (1995). The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. U.S.A.* 92, 1222–1226. doi: 10.1073/pnas.92.4.1222

Salminen, M. O., Ehrenberg, P. K., Mascola, J. R., Dayhoff, D. E., Merling, R., Blake, B., et al. (2000). Construction and biological characterization of infectious molecular clones of HIV-1 subtypes B and E (CRF01_AE) generated by the polymerase chain reaction. *Virology* 278, 103–110. doi: 10.1006/viro.2000.0640

Sanches, M., Krauchenco, S., Martins, N. H., Gustchina, A., Wlodawer, A., and Polikarpov, I. (2007). Structural characterization of B and non-B subtypes of HIV-protease: insights into the natural susceptibility to drug resistance development. *J. Mol. Biol.* 369, 1029–1040. doi: 10.1016/j.jmb.2007.03.049

Sánchez, D., Arazi Caillaud, S., Zapiola, I., Fernandez Giuliano, S., Bologna, R., Mangano, A., et al. (2020). Impact of genotypic diversity on selection of subtype-specific drug resistance profiles during raltegravir-based therapy in individuals infected with B and BF recombinant HIV-1 strains. *J. Antimicrob. Chemother.* 75, 1567–1574. doi: 10.1093/jac/dkaa042

Shafer, R. W., and Schapiro, J. M. (2008). HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS Rev.* 10, 67–84.

Sharma, A., Slaughter, A., Jena, N., Feng, L., Kessl, J. J., Fadel, H. J., et al. (2014). A new class of multimerization selective inhibitors of HIV-1 integrase. *PLoS Pathog.* 10:e1004171. doi: 10.1371/journal.ppat.1004171

Simon, F., Mauclère, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M. C., Saragosti, S., et al. (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med.* 4, 1032–1037. doi: 10.1038/2017

Sliepen, K., Han, B. W., Bontjer, I., Mooij, P., Garces, F., Behrens, A.-J., et al. (2019). Structure and immunogenicity of a stabilized HIV-1 envelope trimer based on a group-M consensus sequence. *Nat. Commun.* 10:2355. doi: 10.1038/s41467-019-10262-5

Stanford University (2022). *HIV Drug Resistance Database*. Available online at: https://hivdb.stanford.edu/ (accessed June 16, 2022).

Svarovskaia, E. S., Feng, J. Y., Margot, N. A., Myrick, F., Goodman, D., Ly, J. K., et al. (2008). The A62V and S68G mutations in HIV-1 reverse transcriptase partially restore the replication defect associated with the K65R mutation. *J. Acquir. Immune Defic. Syndr.* 48, 428–436. doi: 10.1097/QAI.0b013e31817bbe93

Tambuyzer, L., Azijn, H., Rimsky, L. T., Vingerhoets, J., Lecocq, P., Kraus, G., et al. (2009). Compilation and prevalence of mutations associated with resistance to non-nucleoside reverse transcriptase inhibitors. *Antivir. Ther.* 14, 103–109.

Torrecilla, E., Llácer Delicado, T., and Holguín, Á. (2014). New findings in cleavage sites variability across groups, subtypes and recombinants of human immunodeficiency virus type 1. *PLoS One* 9:e88099. doi: 10.1371/journal.pone.0088099

Tran, T. T., Liu, Z., and Fanucci, G. E. (2020). Conformational landscape of non-B variants of HIV-1 protease: a pulsed EPR study. *Biochem. Biophys. Res. Commun.* 532, 219–224. doi: 10.1016/j.bbrc.2020.08.030

Troyano-Hernáez, P., Reinosa, R., Burgos, M. C., and Holguín, Á. (2021a). Short communication: update in natural antiretroviral resistance-associated mutations among HIV type 2 variants and discrepancies across HIV type 2 resistance interpretation tools. *AIDS Res. Hum. Retroviruses* 37, 793–795. doi: 10.1089/AID.2020.0180

Troyano-Hernáez, P., Reinosa, R., and Holguín, Á. (2021b). Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to September 2020: a global and regional approach by epidemiological week. *Viruses* 13:243. doi: 10.3390/v13020243

Troyano-Hernáez, P., Reinosa, R., and Holguín, Á. (2019). "Marcadores genéticos en la proteína de la Cápside p24 en los grupos, subtipos, sub-subtipos y recombinantes del VIH-1," in *Proceedings of the XI CONGRESO NACIONAL GeSIDA*, Toledo, 124–125.

Troyano-Hernáez, P., Reinosa, R., and Holguín, Á. (2020). "Mutaciones en la proteína Spike de SARS-CoV-2 por Comunidades Autónomas en secuencias españolas recogidas hasta junio 2020," in *Proceedings of the I Congreso Nacional COVID-19*, Toledo, 76.

Troyano-Hernáez, P., Reinosa, R., and Holguín, Á. (2022). HIV capsid protein genetic diversity across HIV-1 variants and impact on new capsid-inhibitor lenacapavir. *Front. Microbiol.* 13:854974. doi: 10.3389/fmicb.2022.854974

Tuaillon, E., Gueudin, M., Lemee, V., Gueit, I., Roques, P., Corrigan, G. E., et al. (2004). Phenotypic susceptibility to nonnucleoside inhibitors of virion-associated reverse transcriptase from different HIV types and groups. *J. Acquir. Immune Defic. Syndr.* 37, 1543–1549. doi: 10.1097/00126334-200412150-00001

Turner, D., Roldan, A., Brenner, B., Moisi, D., Routy, J.-P., and Wainberg, M. A. (2004). Variability in the PR and RT genes of HIV-1 isolated from recently infected subjects. *Antivir. Chem. Chemother.* 15, 255–259. doi: 10.1177/095632020401500504

Wares, M., Mesplède, T., Quashie, P. K., Osman, N., Han, Y., and Wainberg, M. A. (2014). The M50I polymorphic substitution in association with the R263K mutation in HIV-1 subtype B integrase increases drug resistance but does not restore viral replicative fitness. *Retrovirology* 11:7. doi: 10.1186/1742-4690-11-7

Wensing, A. M., Calvez, V., Ceccherini-Silberstein, F., Charpentier, C., Gunthard, H. F., Paredes, R., et al. (2019). 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 27, 111–121.

Yamaguchi, J., Vallari, A., McArthur, C., Sthreshley, L., Cloherty, G. A., Berg, M. G., et al. (2020). Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *J. Acquir. Immune Defic. Syndr.* 83, 319–322. doi: 10.1097/QAI.0000000000002246

Yebra, G., de Mulder, M., del Romero, J., Rodríguez, C., and Holguín, A. (2010). HIV-1 non-B subtypes: high transmitted NNRTI-resistance in Spain and impaired genotypic resistance interpretation due to variability. *Antiviral Res.* 85, 409–417. doi: 10.1016/j.antiviral.2009.11.010

Zhang, Y., Murakoshi, H., Chikata, T., Akahoshi, T., Van Tran, G., Nguyen, T. V., et al. (2021). Effect of difference in consensus sequence between HIV-1 subtype A/E and subtype B viruses on elicitation of gag-specific CD8(+) T cells and accumulation of HLA-associated escape mutations. *J. Virol.* 95:e02061-20. doi: 10.1128/JVI.02061-20

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read
for greatest visibility
and readership

**FAST PUBLICATION**
Around 90 days
from submission
to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative,
and constructive
peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers
acknowledged by name
on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data
and methods to enhance
research reproducibility

**DIGITAL PUBLISHING**
Articles designed
for optimal readership
across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics
track visibility across
digital media

**EXTENSIVE PROMOTION**
Marketing
and promotion
of impactful research

**LOOP RESEARCH NETWORK**
Our network
increases your
article's readership