IDbSV: An Open-Access Repository for Monitoring SARS-CoV-2 Variations and Evolution

Abdelmounim Essabbar[1], Souad Kartti[1], Tarek Alouane[1], Mohammed Hakmi[1], Lahcen Belyamani[2] and Azeddine Ibrahimi[1]*

[1] Medical Biotechnology Laboratory (MedBiotech), Bioinova Research Center, Rabat Medical and Pharmacy School, Mohammed Vth University, Rabat, Morocco, [2] Emergency Department, Military Hospital Mohammed V, Rabat Medical & Pharmacy School, Mohammed Vth University, Rabat, Morocco

Ending COVID-19 pandemic requires a collaborative understanding of SARS-CoV-2 and COVID-19 mechanisms. Yet, the evolving nature of coronaviruses results in a continuous emergence of new variants of the virus. Central to this is the need for a continuous monitoring system able to detect potentially harmful variants of the virus in real-time. In this manuscript, we present the International Database of SARS-CoV-2 Variations (IDbSV), the result of ongoing efforts in curating, analyzing, and sharing comprehensive interpretation of SARS-CoV-2's genetic variations and variants. Through user-friendly interactive data visualizations, we aim to provide a novel surveillance tool to the scientific and public health communities. The database is regularly updated with new records through a 4-step workflow (1—Quality control of curated sequences, 2—Call of variations, 3—Functional annotation, and 4—Metadata association). To the best of our knowledge, IDbSV provides access to the largest repository of SARS-CoV-2 variations and the largest analysis of SARS-CoV-2 genomes with over 60 thousand annotated variations curated from the 1,808,613 genomes alongside their functional annotations, first known appearance, and associated genetic lineages, enabling a robust interpretation tool for SARS-CoV-2 variations to help understanding SARS-CoV-2 dynamics across the world.

Keywords: SARS-CoV-2, COVID-19, genomic variations, database, mutation

INTRODUCTION

The Coronavirus Disease 2019 outbreak (COVID-19) caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has spread from Wuhan China in November 2019 to over 214 countries and territories around the world causing more than 4 million deaths (as August 2021) (1). Concerted efforts have been made in sequencing, analyzing and sharing SARS-CoV-2 genomes all around the world to control the spread of the virus and in particular to assess the virulence of the variants in circulation (2). In the absence of evidence of mutational escape from the currently developed treatments, one should continuously track all possible variations (3, 4). Monitoring SARS-CoV-2's variation dynamics is critical for the treatment of COVID-19 and ensuring the effectiveness of potential vaccines plays a central role in reinforcing international efforts to control the spread of viruses. So far, several databases have been

published focusing on the genetic variants of SARS-CoV-2. GISAID is a pathogenic virus database that provides options to search for SARS-CoV-2 sequences based on their location and date of collection alongside an analytical tool for sequence alignment and visualization (5). The abundance of sequencing data on GISAID and other databases such as NCBI Genbank (6), ViPR allowed the development of more specific tools for monitoring SARS-CoV-2 evolution (7). Pangolin (Phylogenetic Assignment of Named Global Outbreak Lineages) was developed to help the assignment of likeness between SARS-CoV-2 genomes according to a dynamic lineage nomenclature scheme (8). However, Pangolin were dedicated to the classification of SARS-CoV-2 by clades, which have been determined on the basis of several variants of current genetic markers instead of a systematic analysis of all individual variations. In the same context, after over a year of COVID-19, several tools have been developed including Nextstrain (9), BioAider (10), Coronapp (11), CoV-Seq (12), ViruSurf (13), NGDC (14), CoV-GLUE (15), Favicov (16), and IDP 2.0 (17) to provide analysis of SARS-CoV-2 sequences. Yet most of these tools either settle for the annotation of an input given sequence, or lack information associated with the genetic variations such as functional interpretation, location and date of appearance, and associated lineages which are essential for exploring the time course and potential routes of transmission of SARS-CoV-2. Likewise, GESS (18) provide information about single nucleotide variants (SNVs) within a chosen genomic region or protein, or in a certain country/area of interest, however, it misses information about the other types of variations (INDELS and MNV) that played a crucial role in the evolution of SARS-CoV-2 and enhancing its spreading capacities (19). The International Database of SARS-CoV-2 variations (IDbSV) was developed to close these gaps. IDbSV is an open repository, with monthly scheduled updates, hosting curated data about SARS-CoV-2 genetic variations identified from the analysis of high-quality SARS-CoV-2 genome sequences. In the next sections, we present a brief overview of the main genomic findings, with special focus on the most dominant variations, their first appearance and associated lineages well as the main functions implemented within IDbSV.

MATERIALS AND METHODS

Data Collection
Complete nucleotide sequences of SARS-CoV-2 genomes were collected from the GISAID EpiCov^{TM} (https://www.epicov.org/epi3/), belonging to 188 territories and distributed over five continents as follows: Africa (1.57%), Asia (8.62%), Europe (63.81%), North America (21.23%), Oceania (1.86%) and South America (2.91%) and the date of samples collection was between December 24 2019 to July 28th, 2021. (The list of genomes used to build the current version of IDbSV can be found as **Supplementary Table**).

Quality Control
Only high-quality complete genomes with available metadata were considered for the variations analysis. Genomes were first filtered considering genomes completeness (>29,000 bp),

coverage (<1%) and percentage of undefined bases (<5% Ns). The remaining sequences were selected according to the availability of their geographical and temporal metadata.

Variants Calling and Functional Annotation
High-quality sequences were mapped individually against the SARS-CoV-2 reference genome Wuhan-Hu-1/2019 (Genbank ID: NC_045512.2) using Minimap2-2.17 (20) to identify variants. The resulting SAM files were sorted and converted to BAM formats before calling the genetic variants in Variant Call Format (VCF) using multiple-sample pileup (mpileup) from the SAMtools suite (21). Variation's functional significance was predicted using snpEff 5.0e (22) based on each variant's relative location and nucleic acid alteration (**Figure 1A**). The variants identified in more than 1% of studied samples were considered as recurrent variations and the variations identified in more than 10% of the studied samples were considered as hotspot variations. The identified DNA variations and Amino acid mutations were represented in the HGVS standards and nomenclature to enable systematic exploration of our database via semantic web tools and APIs (23).

Metadata Annotation
The identified list of variants was first linked to their appropriate strain's contextual information according to GISAID geographical and temporal metadata. Then, based on their amino acid annotations, each variant was associated with appropriate lineages according to Rambaut's nomenclature proposal for SARS-CoV-2 lineages (pangolin.cog-uk.io/) (8).

Platform Architectural Design and Structure
Data processing and analysis were conducted using Python-3.8 and R-3.6, and the web platform was implemented using PHP 7 and a relational database connection.

An object-oriented architecture was designed and implemented in a relational database (MySQL) to store the annotated variants instead of the conventional spreadsheet file (CSV/VCF) to allow further flexibility when formulating search queries and alleviate database load by reducing data duplication. The database architecture and relationships between tables is shown in **Figure 1B**. The Human Genome Variation Society (HGVS) nomenclature (23) were used as primary keys for both nucleotide and amino acids variations to join tables.

RESULTS

Distribution of Variations
From over 2,683,000 genomic sequences available on the GISAID database on the 5th of August 2021, we selected 1,808,613 (67.8%) complete high-quality SARS-CoV-2 genomic sequences. Our analysis of these sequences revealed the presence of 60,148 distinct variations coding for 57,581 different amino acid mutations across the 11 SARS-CoV-2 genes. The accumulation of variations, especially in structural regions gives viruses a selective advantage for host invasion and adaptation, higher translatability

FIGURE 1 | Data analysis workflow and database schema. **(A)** Variations' extraction workflow: the flowchart resumes the procedure of variation extraction and annotation. The variations are identified using MINIMAP2, SAMTOOLS and BCFTOOLS and annotated using snpEff. These SNPs are then associated with appropriate strain's metadata and lineages according to GISAID (www.epicov.org/epi3/) and PANGOLIN (pangolin.cog-uk.io). The data is then exported from CSV files to relational tables as SQL files. Finally, the outputs are deployed online on a monthly basis. Data processing scripts are available openly in https://github.com/mouneem/IDbSV and the extracted list of strains can be found in **Supplementary Table 1**. **(B)** Database Schema. Using this object-oriented architecture, instead of the standard VCF table, allowed further queries and alleviation of search queries flexibility and reduced storage.

of more virulent strains, and drug resistance (24, 25). **Figure 2** shows the different types of variants identified, their positions, and their frequencies. We identified 27.2% of the variants in regions coding for structural proteins including spike (S): 14.8%, nucleocapsid (N): 4.8%, membrane (M): 2.1%, and envelope (E): 0.8%. while the remaining 72.2% were distributed over six Open-Reading Frame genes (ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10) as it is shown in **Figures 2A,E**. From an evolutionary perspective, the rate of variations can also be a key parameter to assess the speed of viral evolution. We found an evolution rate of 18.86 variations per genome on average and we estimated the increase in the cumulative count of variation by ∼0.08 [std error 0.001] additional variant each day as it is shown in **Figure 2F**. In addition, among the 60 thousand variations, 98.2% were located in coding regions of the genome and distributed as follows: 67.2%

missense variations and 30.8% synonymous variations, resulting in a 2.18 Non-synonymous/Synonymous variations ratio.

Frequency of Variations

Despite the low rate of recurrent variations, some variations were widely spread worldwide. **Figure 2E** shows the distribution of variations and their frequencies: 162 variations were identified with a frequency >1% while only 40 variations were identified as hotspot variations (frequency > 10% of the total samples). Expectedly, the two missense mutations 23403A > G and 14408C > T were identified in nearly 1.6 million genomic samples (93.8%), this mutation was linked to the B.1 lineages that spread from Europe to become the most prevalent form of the virus around the world. The frequency of the remaining variations changed according to geographic location as described in **Table 1**.
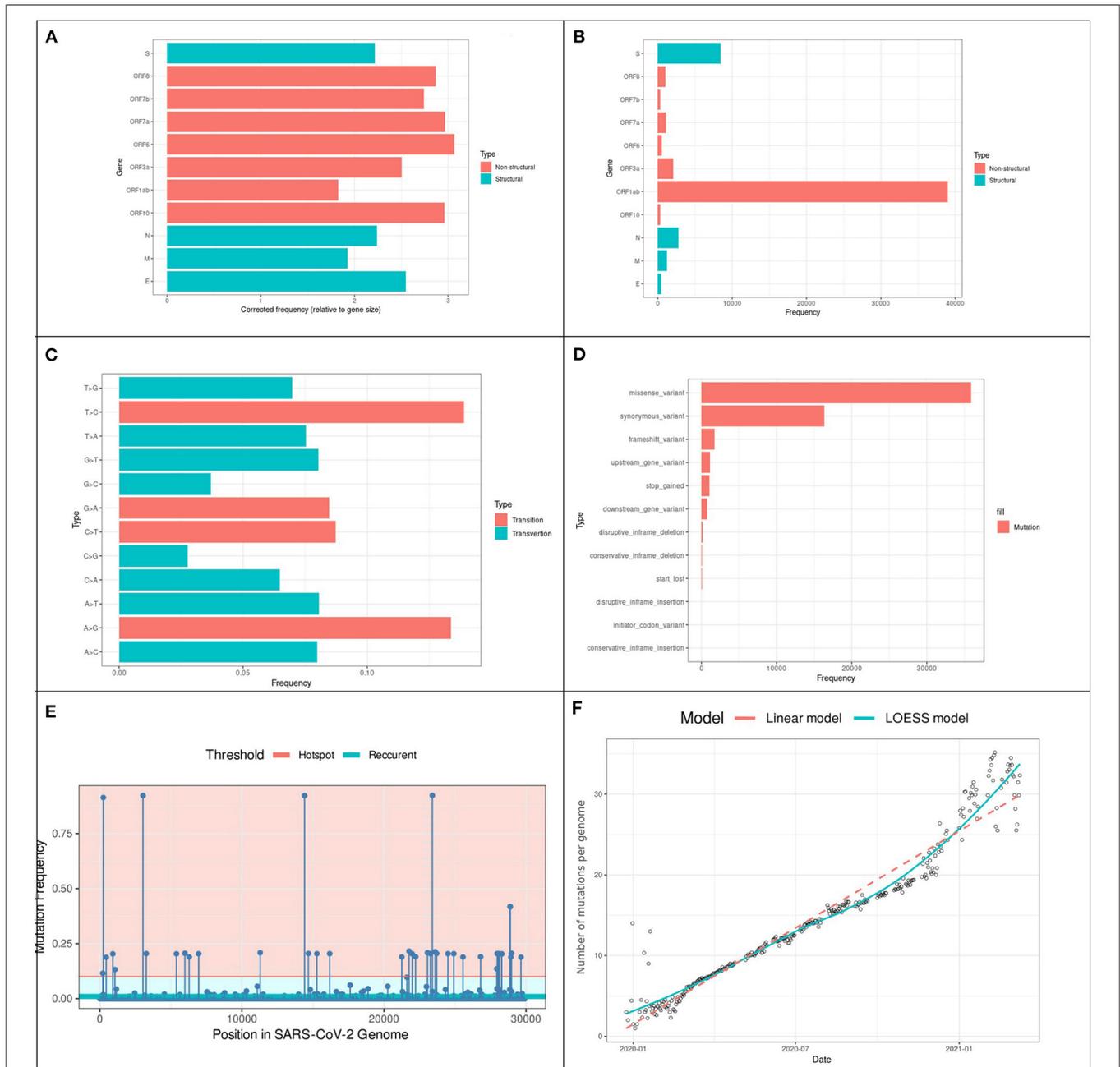
FIGURE 2 | Prevalence and distribution of types of variants in 201,951 SARS-CoV-2 genomes **(A)** variants locations: the y-axis represents the gene location of variants, and the x-axis the rate of each gene. **(B)** Gene length corrected frequency **(C)** variants type frequencies: the y-axis represents the type of variants, and the x-axis the rate of each variation. **(D)** Types frequency: The prevalence of each type of variation: the y-axis represents the type of variant, and the x-axis the rate of each type. **(E)** Distribution of the 36,967 variants across the SARS-CoV-2 genome. The Lollipop plot illustrates the location of variations. The horizontal lines represent the threshold for recurrent (light-blue) and hotspot (Red) variations. All types of variations are included (non-synonymous, synonymous, and intergenic). Forty two (hotspot) variations occurred in more than 10% of analyzed genomes from which 241C > T, 3037C > T, 14408C > T and 23403A > G were identified in more than 91% (N ≥ 184525). **(F)** Variations accumulative on SARS-CoV-2 genomes over 16 months. Points on the scatter-plot represent genomes, the x-axis represents the collection date and y-axis is the number of variations. Linear (red) and Local Polynomial (light-blue) regression models are plotted to visualize the trend of evolution over the past 16 months of the pandemic.

Other noteworthy lineages that spread to over 10% of the population are B.1.1, B.1.1.7 and B.1.617.2 which correspond to the current literature (3, 26).

Content and Features

The present findings in the previous sections were summarized in the interactive web-platform (accessible through http://IDbSV.

TABLE 1 | Top-20 recurrent variants and their frequencies in 5 different geographic regions.

SNP	Africa	Americas	Asia	Europe	Oceania
23403A > G	94.51%	90.63%	71.26%	94.38%	88.87%
3037C > T	94.40%	90.48%	71.81%	94.22%	88.90%
14408C > T	92.51%	90.50%	71.68%	94.11%	88.84%
241C > T	93.73%	89.48%	71.76%	93.84%	80.69%
28881G > A	49.86%	15.30%	39.72%	33.51%	74.85%
28883G > C	49.81%	15.19%	39.42%	33.42%	74.83%
28882G > A	49.81%	15.19%	39.48%	33.44%	74.82%
1163A > T	0.00%	0.03%	3.24%	0.17%	69.03%
18555C > T	0.00%	0.56%	0.30%	0.37%	68.34%
16647G > T	0.00%	0.09%	0.07%	0.08%	68.00%
23401G > A	0.00%	0.00%	0.00%	0.02%	67.98%
7540T > C	0.00%	0.00%	0.00%	0.00%	67.98%
22992G > A	1.39%	0.62%	0.32%	7.73%	66.40%
25563G > T	15.20%	59.33%	22.60%	16.34%	8.51%
1059C > T	8.26%	54.38%	5.92%	6.35%	6.34%
27964C > T	0.00%	22.74%	0.55%	0.42%	1.05%
22227C > T	0.94%	0.31%	0.46%	20.24%	0.46%
6286C > T	0.67%	0.16%	0.42%	20.21%	0.32%
28932C > T	1.44%	0.11%	0.37%	20.04%	0.28%
445T > C	0.67%	0.09%	0.36%	19.97%	0.29%
29645G > T	0.72%	0.12%	0.51%	19.82%	0.28%

medbiotech-lab.ma) to assist the navigation over thousands of annotated records via a user-friendly graphical user interface. A querying tool has been implemented in the platform to simplify the genome browsing by positions in genome and genes which allows the investigation of the variations occurring in specific regions or genes (**Figure 3**). Users may retrieve information about a specific variation by its position in the genome (**Figure 3**), by its position in a specific gene (**Figure 3**) or by the summarize table visible in the home page. User is automatically redirected to the page with functional annotation of the selected variation such as HGVS nomenclature, resulting Amino Acid mutation, position in the specific gene, type of mutation and the predicted impact of the mutation. For example, the screenshots in **Figure 3** provide a demonstration of the database functionalities using the substitution of Guanine (G) by Adenine (A) in the 23,012 positions of SARS-CoV-2 genome which led to a missense change of glutamic acid (E) by Lysine (K) in the position 484 of the Spike protein (S: p.Glu484Lys).

Furthermore, the annotated list variations were linked to metadata information about the countries and regions of origin and date of collection. Which allows the users of IDbSV to extract the contextual information about the first known appearance of each variant (**Figure 3**). Moreover, we identified variations to their appropriate lineages following (8) nomenclature to allow further understanding of the global spread patterns and determinants. As shown in the show case example in (**Figure 3**), according to the IDbSV 23012G > A was first identified in Japan on 1st February 2020 (GISAID ID: HCOV-19JAPANDONNER3802020), then subsequently spread around the world across more than 8 lineages including the lineages of interest B.1.351, A.23.1 and B.1.525. This example not only

showcases the use of the present database, but also highlights the importance of metadata such as appearance location and date and lineages association in the context of following the spread of SARS-CoV-2 and understanding the dynamics of the virus.

DISCUSSION

The global collaborative efforts have been, with no doubt, the key weapon in the fight against COVID-19. The massive efforts in sequencing and sharing SARS-CoV-2 genomes allowed investigators to reveal many previously unknown characteristics of COVID-19 in its diagnosis and treatment (27). IDbSV joins international efforts by providing comprehensive datasets on the genetic evolution of SARS-CoV-2 in time and space. The current version of IDbSV provides access to the result of analysis of over 1.8 million high quality complete genomes. The inclusion of a maximum number of genomic samples increases the statistical significance of our findings and allows the consideration of more recent variations and less pathogenic ones, which may raise more concerns in the future (28). In addition to the thousands of identified variations that can be browsed through the platform's GUI, the database provides access to their associated metadata. Furthermore, unlike the existing tools developed to assist monitoring SARS-CoV-2's evolution, IDbSV does not require any input file and/or computational knowledge to be used. Moreover, IDbSV can be used as an online annotation tool for the interpretation of mutations. These annotated variations are openly accessible using the GUI or API requests which enable the use of IDbSV for the development of other specific pipelines.

Until 5th August 2021, IDbSV hosted over 60 thousand variations extracted from the analysis of over 1.8 million

FIGURE 3 | Case study of the 23012G > A variant. Screenshots from IDbSV web-portal providing contextual information about IDbSV c.23012G > A coding for the variation E484K in spike glycoprotein region. The users can start by searching for a genetic variation using its position in the genome (1), or its amino acid position in a specific gene (2). The database automatically returns three main results: (1) the functional annotation of the selected mutation (Position in the genome, associated amino acid mutation, type of mutation, predicted effect), (2) geographic and time information about the first known appearance of the selection variation, (3) the associated lineages with the selected variation.

SARS-CoV-2 genomes. It is interesting to note that the analysis of these strains revealed consistent results with the finding of more specialized studies (29–35). Yet we revealed a median variation rate of 18.6 variations per genome with an increasing rate of one more variation every 12.5 days, which is expectedly higher than what identified in earlier studies (36, 37). Noteworthy, 27.8% of variations were identified in regions coding for structural proteins, this put more emphasis on the importance of monitoring SARS-CoV-2 variations especially in these regions, as these structural proteins are the main targets of the currently developed vaccines) (38, 39).

SUMMARY

Since November 2020, IDbSV provided a complete atlas of SARS-CoV-2 genetic changes, with particular emphasis on recurrent and potentially harmful mutations. To the best of our knowledge, the current version of IDbSV (August 2021) provides open access to the largest repository of SARS-CoV-2 variations, with 60.148 annotated genetic changes curated from 1.8 million selected samples representing different regions and countries. Given the

importance of monitoring the changes in virus transmissibility and severity, the goal of IDbSV is to provide an open-access and user-friendly platform for researchers and the public to browse SARS-CoV-2 variations in real-time. In addition to the functional annotation of the identified variations, IDbSV provides detailed information about the date of appearance, location of appearance and associated phylogenetic lineage of each variation. The results of these work produced an overview of circulating variations that provide guidance for public health measures to fight the pandemic. We plan to continuously update the platform monthly with new data and features as the fight against COVID-19 continues, to help researchers reveal and interpret new variations and potentially aid in drug and vaccine design.

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

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed. 2021.765249/full#supplementary-material

REFERENCES

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis.* (2020) 20:533–4. doi: 10.1016/S1473-3099(20)30120-1
2. Uddin M, Mustafa F, Rizvi TA, Loney T, Al Suwaidi H, Al-Marzouqi AHH. SARS-CoV-2/COVID-19: viral genomics, epidemiology, vaccines, and therapeutic interventions. *Viruses.* (2020) 12:526. doi: 10.3390/v12050526
3. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell.* (2020) 182:812–27. doi: 10.1016/j.cell.2020.06.043
4. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ.* (2020) 98:495. doi: 10.2471/BLT.20.253591
5. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance.* (2017) 22:30494. doi: 10.2807/1560-7917.ES.2017.22.13.30494
6. Benson DA, Cavanaugh M, Clark KK. (2012). "GenBank." *Nucleic Acids Res.* 41:D36–42. doi: 10.1093/nar/gks1195
7. Pickett BES. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* (2012) 40:D593–8. doi: 10.1093/nar/gkr859
8. Rambaut A, Holmes EC, O'Toole Á, Hill V. McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol.* (2020) 5:1403–7. doi: 10.1038/s41564-020-0770-5
9. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics.* (2018) 34:4121–3. doi: 10.1093/bioinformatics/bty407
10. Zhou ZJ, Qiu Y, Pu Y, Huang X, Ge XY. BioAider: an efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission. *Sustain Cities Soc.* (2020) 63:102466. doi: 10.1016/j.scs.2020.102466
11. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM. Coronapp: a web application to annotate and monitor SARS-CoV-2 mutations. *J Med Virol.* (2021) 93:3238–45. doi: 10.1002/jmv.26678
12. Liu B, Liu K, Zhang H, Zhang L, Bian Y, Huang L. CoV-Seq, a new tool for SARS-CoV-2 genome analysis and visualization: development and usability study. *J Med Internet Res.* (2020) 22:e22299. doi: 10.2196/22299
13. Canakoglu A, Pinoli P, Bernasconi A, Alfonsi T, Melidis DP, Ceri S. ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Res.* (2021) 49:D817–24. doi: 10.1093/nar/gkaa846
14. Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK. The 2019 novel coronavirus resource. *Hereditas.* (2020) 42:212–21. doi: 10.16288/j.yczz.20-030
15. Singer J, Gifford R, Cotten M, Robertson D. *CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation.* (2020). doi: 10.20944/preprints202006.0225.v1
16. Rophina M, Pandhare K, Mangla M, Shamnath A, Jolly B, Sethi M, et al. . *FaviCoV-A Comprehensive Manually Curated Resource for Functional Genetic Variants in SARS-CoV-2.* (2020). doi: 10.31219./osf.io/wp5tx
17. Desai S, Rane A, Joshi A, Dutt A. IPD 20: to derive insights from an evolving SARS-CoV-2 genome. *BMC Bioinform.* (2021) 22:1–9. doi: 10.1186/s12859-021-04172-x
18. Fang S, Li K, Shen J, Liu S, Liu J, Yang L, et al. GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. *Nucleic Acids Res.* (2021) 49:D706–14. doi: 10.1093/nar/gkaa808
19. Garry RF, Andersen KG, Gallaher WR, Lam TT, Gangaparapu K, Latif AA. Spike protein mutations in novel SARS-CoV-2 'variants of concern'commonly occur in or near indels. *Image.* (2021) 881:85.
20. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* (2018) 34:3094–100. doi: 10.1093/bioinformatics/bty191
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics.* 25, 2078–9. doi: 10.1093/bioinformatics/btp352
22. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly.* (2012) 6:80–92. doi: 10.4161/fly.19695
23. den Dunnen JT, Dalgleish R, Maglott R, Hart DR, Greenblatt RK. MS, McGowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat.* (2016) 37:564–9. doi: 10.1002/humu.22981
24. Omotoso OE. Contributory role of SARS-CoV-2 genomic variations and life expectancy in COVID-19 transmission and low fatality rate in Africa. *Egypt J Med Hum Genetics.* (2020) 21:1–6. doi: 10.1186/s43042-020-00116-x
25. Xiao Y, Rouzine IM, Bianco S, Acevedo A, Goldstein EF, Farkov M. RNA recombination enhances adaptability and is required for virus spread and virulence. Cell Host Microbe. (2016) 19:493–503. doi: 10.1016/j.chom.2016.03.009
26. Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. *Proc Natl Acad Sci.* (2021) 118:e2012008118. doi: 10.1073/pnas.2012008118
27. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol.* (2020) 11:1800. doi: 10.3389/fmicb.2020.01800
28. Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li, M. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr Biol.* (2021) 31:R918–29. doi: 10.1016/j.cub.2021.06.049
29. Harcourt J, Tamin A, Lu X, Kamili S, Sakthivel SK, Murray J. Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. *Emerg Infect Dis.* (2020) 26:1266. doi: 10.3201/eid2606.200516
30. Ferguson N, Laydon D, Nedjati Gilani G, Imai N, Ainslie K, Baguelin M. *Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID19 Mortality and Healthcare Demand.* London: Imperial College of London (2020).
31. Wang M, Li Y, HuangFu M, Xiao Y, Zhang T, Han M, et al. Pluronic-attached polyamidoamine dendrimer conjugates overcome drug resistance in breast cancer. *Nanomedicine.* (2016) 11:2917–34. doi: 10.2217/nnm-2016-0252
32. Grint DJ, Wing K, Williamson E, McDonald HI, Bhaskaran K, Evans D, et al. Case fatality risk of the SARS-CoV-2 variant of concern B. 11 7 in England, 16 November to 5 February. *Eurosurveillance.* (2021) 26:2100256. doi: 10.2807/1560-7917.ES.2021.26.11.210025

33. Williams C, Al-Bargash D, Macalintal C, Stuart R, Seth A, Latham J. (2021). COVID-19 outbreak associated with a SARS-CoV-2 P. 1 lineage in a long-term care home after implementation of a vaccination program–Ontario, April-May 2021. *Clin Infect Dis.* 2021:ciab617. doi: 10.1093/cid/ciab617

34. Alouane T, Laamarti M, Essabbar A, Hakmi M, Bouricha EM, Chemao-Elfihri MW. Genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes: moving toward a universal vaccine for the "confined virus"? *Pathogens.* (2020) 9:829. doi: 10.3390/pathogens9100829

35. Villoutreix BO, Calvez V, Marcelin AG, Khatib AM. In silico investigation of the new UK (B 11 7) and South African (501y v2) SARS-CoV-2 variants with a focus at the ace2–spike rbd interface. *Int J Mol Sci.* (2021) 22:1695. doi: 10.3390/ijms22041695

36. MacLean OA, Orton RJ, Singer JB, Robertson DL. (2020). No evidence for distinct types in the evolution of SARS-CoV-2. *Virus Evol.* 6:veaa034. doi: 10.1093/ve/veaa034

37. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. *PLoS ONE.* (2020) 15:e0240345. doi: 10.1371/journal.pone.0240345

38. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol Immunol.* (2020) 17:613–20. doi: 10.1038/s41423-020-0400-4

39. Sanami S, Alizadeh M, Nosrati M, Dehkordi KA, Azadegan-Dehkordi F, Tahmasebian S. Exploring SARS-COV-2 structural proteins to design a multi-epitope vaccine using immunoinformatics approach: an in silico study. *Comput Biol Med.* (2021) 133:104390. doi: 10.1016/j.compbiomed.2021.104390

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.