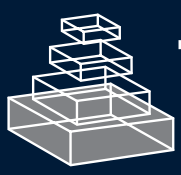
frontiers

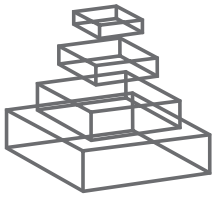
RESEARCH TOPICS

GENOMICS AND COMPUTATIONAL SCIENCE FOR VIRUS RESEARCH

Topic Editors
Hironori Sato, Masaru Yokoyama and Hiroyuki Toh



frontiers in MICROBIOLOGY



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-126-0
DOI 10.3389/978-2-88919-126-0

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

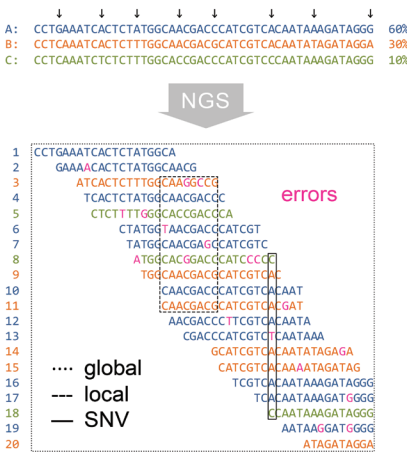
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - 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: researchtopics@frontiersin.org

GENOMICS AND COMPUTATIONAL SCIENCE FOR VIRUS RESEARCH

Topic Editors:
Hironori Sato, National Institute of Infectious Diseases, Japan
Masaru Yokoyama, National Institute of Infectious Diseases, Japan
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan



A biologically striking and clinically important feature of viruses is their rapid evolutionary dynamics in nature. The continual interactions between viruses and host organisms promote quick changes in virus populations, eventually leading to co-evolution of viruses and hosts for their survival. The structural and functional information on the interactions between viruses and hosts should provide a molecular and biological basis to understand infection, replication, cell/host-tropism, immune escape, pathogenesis, and direction of evolution of viruses. The information is also essential to develop methods to control transmission and replication of pathogenic viruses. However, the integrated information on the structure, function, and evolution of viruses and hosts has remained poorly accumulated, partly due to the limitation of analytical methods. Recent progress in genome science and computational approach may open up a new avenue of research of the interactions between viruses and hosts by integrating information on the structures, functions, and evolution. In this Research Topic, we welcome papers concerning the computer-assisted structural and functional studies based on genomic information, with theoretical or in combination with experimental approaches, for understanding molecules, infection, replication, cell/host-tropism, immune escape, pathogenesis, and evolution of viruses in nature.

Table of Contents

frontiers in
MICROBIOLOGY

EDITORIAL
published: 07 March 2013
doi: 10.3389/fmicb.2013.00042

Genomics and computational science for virus research

Hironori Sato[1]*, Masaru Yokoyama[1] and Hiroyuki Toh[2]

[1] *Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo, Japan*
[2] *Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan*
**Correspondence: hirosato@nih.go.jp*

Edited by:
Akio Adachi, The University of Tokushima Graduate School, Japan

Reviewed by:
Akio Adachi, The University of Tokushima Graduate School, Japan

RNA viruses are highly mutable, yet changes in genomes and proteins would be restricted by the functional and structural constraints inherent in the survival strategies of viruses in nature. Rapidly evolving technologies in genomics and computational science are now opening up a new avenue for elucidating the real picture of diversity of the organism in nature and for studying the principles underlying the maintenance and change of structures, interactions, and functions of biomolecules. The information is essential for understanding the evolutionary dynamics of virus-host interactions in virological, immunological, and epidemiological phenomena and for rationally developing methods to control RNA viruses. In this Research Topic, we present 17 timely articles, consisting of 5 reviews, 3 mini reviews, 7 original researches, 1 hypothesis & theory, and 1 perspective, all of which underscore the challenges and increasing importance of incorporating the new technologies to study RNA viruses and their impacts on hosts.

EXPLORATIONS OF VIRAL QUASISPECIES, MICRORNAs, AND ANTIBODYOMES IN NATURE

Beerenwinkel et al. (2012) reviewed the challenges and opportunities in inferring the diversity of intra-host virus populations using next-generation sequencing technologies. They discuss the wisdom of reducing artificial errors during sample preparation, existing approaches inferring local and global diversity from sequence data, and successful applications on basic and biomedical studies. Tan Gana et al. (2012) reviewed the latest articles describing cellular and viral microRNAs involved in HIV-1 infection. They describe recent advances in understanding of the biogenesis and functions of the microRNAs in the virus-cell battles and point out roles of the genomics and computational science in obtaining and integrating the information.

Prabakaran et al. (2012) reported on the antibodyomes of 10 healthy individuals obtained by 454 pyrosequencing and bioinformatics analyses. They showed genetic evidence that the antibody subsets with distinct diversity and related to the already-known neutralizing antibodies against the HIV-1, SARS coronavirus, and henipaviruses exist in human IgM repertoires of uninfected individuals. Zhu et al. (2012) reported an antibodyome of an HIV-1-infected individual who produced broadly neutralizing antibodies. Using 454 pyrosequencing, bioinformatics, and functional analyses, they suggested a role of somatic maturation in generating heavy- and light-chain sequences with varied neutralization phenotypes against HIV-1.

COMPUTATIONAL ANALYSES OF THE 3-D STRUCTURES, INTERACTIONS, AND EVOLUTION OF PROTEINS USING GENETIC INFORMATION

Ode et al. (2012) reviewed the results of molecular dynamic simulations to learn the structural dynamics of proteins in solution. They highlight studies on the structure and function of viral enzymes, virion structures, mechanisms of viral resistance against host immunities and anti-viral drugs, and the development of anti-viral agents. Franzosa et al. (2012) reviewed structural systems biology of interactomes in the host-pathogen relationships. They present existing experimental datasets of the host-pathogen interactome and discuss approaches to obtain structural interactome by integrating the biophysical, functional, and evolutionary information. Miki and Katayama (2012) presented a viewpoint on the *in silico* 3-D structural analysis in virus research. They describe importance of incorporating *in silico* modeling techniques into experimental studies to solve structural problems in their neutralization study of the norovirus.

Bozek et al. (2012) provided *in silico* structural models of capsid proteins of HIV-2 and SIV, which revealed marked differences in the electrostatic potential on the interaction surface and suggested a potential role of electrostatic interactions in the evasion of SIV from the rhesus restriction factor Trim5α. Daiyasu et al. (2012) reported a new application of information theory to the study of the divergent evolution of function of chemokine receptors and their homologs, such as decoy and viral receptors, in which both sequence and structural information are used to identify amino acid positions that might be responsible for evolving their distinct functions. Rusu et al. (2012) provided *in silico* structural models of the Rift Valley fever virus glycoproteins Gn and Gc. The models with the cryo-electron microscopy data allowed the authors to identify four possible arrangements of the glycoproteins in the virion envelope and to indicate how these proteins assemble to form the capsomer base and intercapsomer connections. Yokoyama et al. (2012) provided *in silico* structural models of sapovirus protease docked to its substrate peptides; these models described how this enzyme realizes the functional binding of cleavage sites with distinct sequences and allowed rational identification of the sapovirus protease inhibitors in combination with experimental approaches.

ANALYSES OF VIRUS-CELL INTERACTIONS, VIRAL REPLICATION, AND HOST IMMUNE RESPONSES

Iwami et al. (2012) reported a mathematical model to quantitatively characterize the viral replication in cell cultures. In their

study, the data from two time-course experiments of infections with a cell-free virus stock are used to estimate the half-life of infected cells, viral production rate of an infected cell, and the basic reproductive number. Takemura and Murakami (2012) reviewed structure and function of HIV-1 capsid proteins. They describe the capsid structure in relation to their abilities to form a conical core in a virion or to interact with various cellular proteins that promote or suppress viral replication. Nomura and Matano (2012) reviewed the critical roles of host HLA/MHC-I genotypes in disease progression in primate lentivirus infections. They highlighted studies showing the association of the HLA/MHC-I genotypes with rapid or slow AIDS progression during HIV/SIV persistent infections. Kuroki et al. (2012) reviewed the structural biology of the immunologically intriguing cell surface receptors termed paired receptors. By referencing recent studies of two major structural superfamilies, the immunoglobulin-like and the C-type lectin-like receptors, they described how these receptors discriminate self and non-self ligands to maintain homeostasis in the immune system.

ANALYSES OF VIRAL TRANSMISSION NETWORK, EVOLUTION, AND RE-EMERGING DISEASES

Shiino (2012) reviewed phylodynamic analyses of viral infections, transmission, and evolution in host populations. The author highlighted recent viral researches inferring epidemiologically important parameters, such as infectious clusters, transmission connections, basic reproductive number, and fluctuation of the viral population size. Sasaki et al. (2012) proposed a mathematical model that describes the temporal dynamics of the spread of poliovirus avirulent and virulent strains in a human population; this model was used to estimate the risk of outbreak following the cessation of administration of oral polio vaccine and to predict conditions that significantly increased the outbreak risk.

In summary, these articles present overviews and examples of "Frontiers in Virology" to elucidate viral diversity, protein 3-D structures, virus-host interactions, and evolution of RNA viruses using forefront technologies in genomics and computational science.

REFERENCES

Beerenwinkel, N., Gunthard, H. F., Roth, V., and Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front. Microbiol.* 3:329. doi: 10.3389/fmicb.2012.00329

Bozek, K., Nakayama, E. E., Kono, K., and Shioda, T. (2012). Electrostatic potential of human immunodeficiency virus type 2 and rhesus macaque simian immunodeficiency virus capsid proteins. *Front. Microbiol.* 3:206. doi: 10.3389/fmicb.2012.00206

Daiyasu, H., Nemoto, W., and Toh, H. (2012). Evolutionary analysis of functional divergence among chemokine receptors, decoy receptors, and viral receptors. *Front. Microbiol.* 3:264. doi: 10.3389/fmicb.2012.00264

Franzosa, E. A., Garamszegi, S., and Xia, Y. (2012). Toward a three-dimensional view of protein networks between species. *Front. Microbiol.* 3:428. doi: 10.3389/fmicb.2012.00428

Iwami, S., Sato, K., De Boer, R. J., Aihara, K., Miura, T., and Koyanagi, Y. (2012). Identifying viral parameters from *in vitro* cell cultures. *Front. Microbiol.* 3:319. doi: 10.3389/fmicb.2012.00319

Kuroki, K., Furukawa, A., and Maenaka, K. (2012). Molecular recognition of paired receptors in the immune system. *Front. Microbiol.* 3:429. doi: 10.3389/fmicb.2012.00429

Miki, M., and Katayama, K. (2012). *In silico* 3D structure analysis accelerates the solution of a real viral structure and antibodies docking mechanism. *Front. Microbiol.* 3:387. doi: 10.3389/fmicb.2012.00387

Nomura, T., and Matano, T. (2012). Association of MHC-I genotypes with disease progression in HIV/SIV infections. *Front. Microbiol.* 3:234. doi: 10.3389/fmicb.2012.00234

Ode, H., Nakashima, M., Kitamura, S., Sugiura, W., and Sato, H. (2012). Molecular dynamics simulation in virus research. *Front. Microbiol.* 3:258. doi: 10.3389/fmicb.2012.00258

Prabakaran, P., Zhu, Z., Chen, W., Gong, R., Feng, Y., Streaker, E., et al. (2012). Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. *Front. Microbiol.* 3:277. doi: 10.3389/fmicb.2012.00277

Rusu, M., Bonneau, R., Holbrook, M. R., Watowich, S. J., Birmanns, S., Wriggers, W., et al. (2012). An assembly model of rift valley Fever virus. *Front. Microbiol.* 3:254. doi: 10.3389/fmicb.2012.00254

Sasaki, A., Haraguchi, Y., and Yoshida, H. (2012). Estimating the risk of re-emergence after stopping polio vaccination. *Front. Microbiol.* 3:178. doi: 10.3389/fmicb.2012.00178

Shiino, T. (2012). Phylodynamic analysis of a viral infection network. *Front. Microbiol.* 3:278. doi: 10.3389/fmicb.2012.00278

Takemura, T., and Murakami, T. (2012). Functional constraints on HIV-1 capsid: their impacts on the viral immune escape potency. *Front. Microbiol.* 3:369. doi: 10.3389/fmicb.2012.00369

Tan Gana, N. H., Onuki, T., Victoriano, A. F., and Okamoto, T. (2012). MicroRNAs in HIV-1 infection: an integration of viral and cellular interaction at the genomic level. *Front. Microbiol.* 3:306. doi: 10.3389/fmicb.2012.00306

Yokoyama, M., Oka, T., Kojima, H., Nagano, T., Okabe, T., Katayama, K., et al. (2012). Structural basis for specific recognition of substrates by sapovirus protease. *Front. Microbiol.* 3:312. doi: 10.3389/fmicb.2012.00312

Zhu, J., O'Dell, S., Ofek, G., Pancera, M., Wu, X., Zhang, B., et al. (2012). Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Front. Microbiol.* 3:315. doi: 10.3389/fmicb.2012.00315

Received: 18 February 2013; accepted: 18 February 2013; published online: 07 March 2013.

*Citation: Sato H, Yokoyama M and Toh H (2013) Genomics and computational science for virus research. Front. Microbiol. **4**:42. doi: 10.3389/fmicb.2013.00042*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 11 September 2012
doi: 10.3389/fmicb.2012.00329

Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data

Niko Beerenwinkel[1,2]*, Huldrych F. Günthard[3], Volker Roth[4] and Karin J. Metzner[3]

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] Swiss Institute of Bioinformatics, Basel, Switzerland
[3] Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[4] Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland

Edited by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan

Reviewed by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan
Fabio Luciani, University of New South Wales, Australia

***Correspondence:**
Niko Beerenwinkel, Department of Biosystems Science and Engineering, ETH Zurich, WRO-1058 8.40, Mattenstrasse 26, 4058 Basel, Switzerland.
e-mail: niko.beerenwinkel@bsse.ethz.ch

Many viruses, including the clinically relevant RNA viruses HIV (human immunodeficiency virus) and HCV (hepatitis C virus), exist in large populations and display high genetic heterogeneity within and between infected hosts. Assessing intra-patient viral genetic diversity is essential for understanding the evolutionary dynamics of viruses, for designing effective vaccines, and for the success of antiviral therapy. Next-generation sequencing (NGS) technologies allow the rapid and cost-effective acquisition of thousands to millions of short DNA sequences from a single sample. However, this approach entails several challenges in experimental design and computational data analysis. Here, we review the entire process of inferring viral diversity from sample collection to computing measures of genetic diversity. We discuss sample preparation, including reverse transcription and amplification, and the effect of experimental conditions on diversity estimates due to *in vitro* base substitutions, insertions, deletions, and recombination. The use of different NGS platforms and their sequencing error profiles are compared in the context of various applications of diversity estimation, ranging from the detection of single nucleotide variants (SNVs) to the reconstruction of whole-genome haplotypes. We describe the statistical and computational challenges arising from these technical artifacts, and we review existing approaches, including available software, for their solution. Finally, we discuss open problems, and highlight successful biomedical applications and potential future clinical use of NGS to estimate viral diversity.

Keywords: next-generation sequencing, viral diversity, viral quasispecies, statistics, bioinformatics, haplotype inference, error correction, quasispecies assembly

INTRODUCTION

Many viruses, in particular RNA or single-stranded DNA viruses, exhibit extreme evolutionary dynamics. They have very high mutation rates, up to six orders of magnitude higher than in humans, short generation times, and large population sizes (Duffy et al., 2008). Under these conditions, genetic variants are produced constantly, and in each infected host, the virus population displays a high degree of genetic diversity. Rapidly evolving viruses are not only ideal systems for studying evolutionary mechanisms (Drummond et al., 2003), but many of them are significant pathogens of vital medical interest, including HIV, HCV, and Influenza (WHO, 2012).

Because of their diversity, intra-host virus populations are often referred to as mutant clouds, swarms, or viral quasispecies. The latter terms were originally introduced in the context of self-replicating macromolecules (Eigen, 1971; Eigen and Schuster, 1977) and have a precise mathematical meaning. A quasispecies is the equilibrium distribution of mutants in a mathematical model that accounts for mutation and selection (Eigen et al., 1988, 1989). In the framework of classical population genetics, it can be regarded as a coupled mutation-selection balance (Wilke, 2005). The main prediction of the quasispecies model is that selection acts on the population as a whole and hence the population dynamics cannot be understood from the fittest strain alone (Van Nimwegen et al., 1999; Wilke et al., 2001). The quasispecies model has later been applied to RNA viruses (Nowak, 1992; Domingo and Holland, 1997), hence the term viral quasispecies. The impact of the quasispecies model is not only due to its mathematical feasibility, but also its conceptual focus on the population as the target of natural selection (Burch and Chao, 2000).

The diversity of virus populations has repeatedly been shown to provide a selective advantage. For example, decreasing the mutation rate of poliovirus artificially, while maintaining its replication rate, resulted in reduced genomic diversity and in failure to adapt to adverse growth conditions (Vignuzzi et al., 2006). Similarly, pre-existing minority drug-resistant variants of HIV-1 have been shown to facilitate rapid viral adaptation leading to failure of antiretroviral therapy (Metzner et al., 2009; Li et al., 2011). In general, viral diversity is advantageous when the virus faces different selection pressures that need to be overcome by evolutionary escape (Iwasa et al., 2003, 2004). Changing selection pressures are common in the life of viruses, for example, after infecting a new host with a different immune response (Pybus and Rambaut, 2009), when infecting different cell types, while being exposed to different chemical agents, or due to changing multiplicity of infection (Ojosnegros et al., 2010). Understanding

and modeling the escape dynamics of these processes is of direct relevance for clinical and public health decisions.

With the introduction of next-generation sequencing (NGS) technologies, the experimental analysis of viral genetic diversity has changed dramatically. Rather than using labor-intensive limiting dilution and individual cloning of viruses followed by traditional Sanger sequencing, NGS now allows for sampling the virus population in a highly parallel fashion in a single experiment. However, the novel high-throughput approach has several pitfalls associated with both the experimental protocol and the statistical analysis of the data. We address both aspects in this review and discuss several successful applications of NGS to viral diversity studies, including drug resistance, immune escape, and epidemiology.

SAMPLE PREPARATION

The usefulness of NGS for viral diversity estimation depends crucially on the quality of the sample and on the procedure to prepare the sample. NGS sequence reads mirror the accumulation of errors, some of them preventable others unavoidable. To minimize the error rate, each step requires careful handling, starting with biological sample retrieval and storage up to the last steps of the NGS procedure itself (**Figure 1**).

Viral genomes are usually protected by the viral capsid and some of them additionally by an envelope, for instance, HIV and HCV. However, retrieval and storage conditions of biological specimens are especially important when studying RNA viruses due to the fragility of RNA (Holodniy et al., 1995; Jose et al., 2005), because degraded RNA will jeopardize all further steps of the analysis. Before starting the extraction of viral genomes, the viral load of the specimen should be considered. The final number of genome copies sequenced provides the basis for assessing viral diversity from the sequence reads (Metzner et al., 2003; Casbon et al., 2011). Low amounts might require a concentrating step, for instance, ultracentrifugation of plasma.

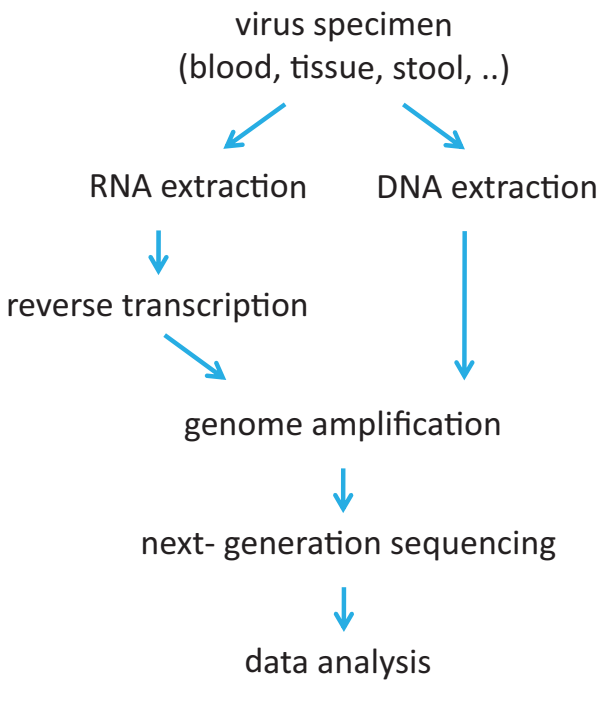


FIGURE 1 | Flow chart of sample processing for next-generation sequencing (NGS) of virus samples.

The choice of protocols used for genome extraction and elimination of contaminating RNA and DNA from other sources like host cells depends on the intended downstream procedures. Numerous kits are offered to extract viral DNA or RNA whose pros and cons will not be discussed here. A more critical point is the enrichment of viral genomes in the context of sample complexity. Three scenarios can be envisioned. (1) The virus is known and an amplicon approach is chosen for NGS. Here, the specificity of the primers might allow for amplifying the viral genome without any upstream enrichment. Nevertheless, it is often beneficial to eliminate contaminating DNA or RNA by DNase or RNase treatment. For instance, investigating HIV RNA genomes requires the elimination of proviral DNA genomes (Fischer et al., 2002). (2) The virus is known, but a random approach is chosen for NGS. Due to the high heterogeneity of some viruses, it might be disadvantageous to use virus-specific primers for amplification due to potential primer bias or even complete failure of amplification (Metzner et al., 2003). In contrast, any random approach, including amplification using degenerated or random primers as well as non-specific adaptor ligation and subsequent amplification using adaptor-specific primers, cannot differentiate between the viral genome and any other nucleic acid (Reyes and Kim, 1991; Chang et al., 1992). Thus, the elimination of contaminating nucleic acids is mandatory when a high coverage of viral genomes is required, as for studying diversity, since the viral genomes represent only a low-abundant fraction in almost all biological specimens (Daly et al., 2011). DNase and RNase treatment, filtration, density gradient centrifugation, and their combinations are commonly used procedures. Enrichment strategies based on hybridization capture might also be suitable (Turner et al., 2009; Althaus et al., 2012) and, potentially, freeze thaw nuclease digestion protocols may also be beneficial to minimize contaminating RNA or DNA (Fischer et al., 2002) (3). The virus in unknown, therefore, random approaches have to be applied. The enrichment of viral genomes is an even greater challenge in this set-up. In this review, we focus on estimating viral diversity from NGS data, a second step after virus discovery (Lipkin, 2010).

After viral genome extraction, an amplification procedure has to be performed, because the current NGS technologies require a high input DNA amount and the viral genome amount is several orders of magnitude lower. Furthermore, RNA genomes have to be reverse transcribed prior to PCR. Every amplification process introduces errors. Reverse transcriptases (RTs) are error-prone enzymes, because of the lack of any proof-reading activity (Preston et al., 1988; Roberts et al., 1988). Some RTs are less error-prone than others, but, in general, RT errors are unavoidable and very difficult to distinguish from real mutations since they are introduced in the first step of amplification. Another important but often ignored problem with reverse transcription is that short, incomplete cDNA fragments can act as primers in subsequent

PCRs and lead to *in vitro* recombination. This phenomenon has been considered only for RT-PCRs amplifying several kilobases (kb) long fragments (Fang et al., 1998). We have recently shown that this effect also occurs very frequently when amplifying short cDNA fragments of a size of only 0.6 kb and can be minimized by using an RNaseH-negative RT (Di Giallonardo et al., submitted).

Four main types of errors can occur during PCR and are relevant for NGS data: (i) biased amplification due to primer mismatches, (ii) *in vitro* recombination due to premature termination of strand elongation and subsequent false hybridization of short DNA fragments acting as primers or, less frequently, due to template switching, (iii) nucleotide misincorporation due to the inaccuracy of DNA polymerases, and (iv) resampling due to, for instance, too low amounts of input DNA copies (Eckert and Kunkel, 1991; Liu et al., 1996; Kanagawa, 2003). Several precautions can be taken to minimize these errors. Primer mismatches can be diminished by choosing primer binding sites in conserved regions of the viral genome or by using degenerated primers. Chimera formation can be reduced by several improvements of PCR conditions such as increasing the elongation time, decreasing the number of cycles, and deleting the final extension step (Meyerhans et al., 1990; Judo et al., 1998). Nucleotide misincorporation can be lowered by using high-fidelity DNA polymerases, and resampling can be reduced, for instance, by optimizing the input copy number. Even when applying all these precautions, it is currently not possible to completely avoid these PCR errors. Furthermore, the discrimination between artificial and real viral variants can be very difficult if not impossible. One possibility is to perform several independent PCRs assuming that most of the errors occur randomly with regard to the sequence position and the timing of the error, i.e., in which PCR cycle the error occurs, resulting in different variants of different frequencies in the replicates. A recently described method uses primer identifiers (IDs) to uniquely label each cDNA molecule (Jabara et al., 2011). This is an elegant procedure to reduce or even eliminate PCR errors, although errors induced during the reverse transcription cannot be addressed in this manner. In addition, the method is only applicable to amplicon-based approaches and a high number of sequence reads are required to obtain a sufficient number of consensus sequences, each of which has to be derived from at least three reads with the same primer ID. Thus, all unique or twice occurring reads, which represent the majority of sequence reads, cannot be considered in the analysis.

Overall, sample preparation is a critical issue in the process of NGS. If unrecognized, errors during sample preparation can lead to an artificially increased diversity of the investigated virus population. To avoid such misinterpretation, the pitfalls of sample preparation need to be identified and properly addressed.

NEXT-GENERATION SEQUENCING

In the last decade, many NGS technologies have been developed and several are commercially available today or about to become available in the near future (Mardis, 2008b; Metzker, 2010). Due to its massively parallel approach, NGS allows for generating much larger volumes of sequencing data in a cost-effective manner as compared to conventional sequencing methods. The increase in throughput has been so far-reaching that NGS is considered revolutionary, because it facilitates many new sequencing applications that had been out of reach (Mardis, 2008a; Schuster, 2008). One of these novel applications is the inference of viral genetic diversity from a single deep-coverage NGS experiment.

All NGS technologies involve the steps of template preparation, sequencing, and imaging, followed by data analysis, but they differ in the realization of each step. 454/Roche pyrosequencing has been the first NGS method commercially available and until today it is the most commonly used technology for the analysis of viruses (Margulies et al., 2005). For pyrosequencing, DNA is isolated, amplified and/or fragmented, adaptor-annealed, and amplified on beads in a micro-droplet emulsion PCR. DNA and beads have to be used in a ratio allowing the hybridization of only one DNA molecule to one bead, i.e., the majority of beads do not contain any DNA molecule. Thus, on each DNA-hybridized bead, a single template gives rise to several thousand copies. These beads are separated from the empty beads and loaded into 1.6 million wells of a picotiter plate, one bead per well, and enzymes for pyrophosphate sequencing are added. Sequencing by synthesis proceeds by adding the four bases in a cyclic order. In each cycle, the light emission associated with base incorporation is detected and remaining chemicals are washed out. The intensity of the light signal is approximately proportional to the number of nucleotides that have been incorporated. All generated signals are recorded as a series of peaks, called a flowgram, from which DNA bases are eventually called (Margulies et al., 2005).

The Illumina Genome Analyzer and HiSeq systems are currently dominating the NGS market (Bentley et al., 2008). Rather than emulsion PCR, Illumina relies on solid-phase amplification, which consists of initial priming and extending of single-stranded templates, followed by bridge amplification of each immobilized template with adjacent primers. In multiple cycles of annealing, extension, and denaturation, around 200 million molecular clusters are formed. For sequencing, all four nucleotides are added simultaneously. Each nucleotide is labeled with a different dye and they are modified to terminate DNA synthesis after incorporation. Color imaging is used to detect the incorporated nucleotide. In a cleavage step, the fluorescent dye is removed and termination is reversed by regenerating the 3′-OH group. Bases are called from the resulting four-color images.

We focus here on the 454/Roche and Illumina platforms, because the vast majority of reported virus sequencing applications have used these systems, but several other technologies can, and are likely to, be used as well, including ABI SOLiD, Ion Torrent, PacBio RS, and Polonator. The technical details in which platforms differ can have important consequences for their applicability to viral sequencing studies. Among other aspects, NGS platforms differ in throughput, runtime, costs, read lengths, and error patterns (Metzker, 2010). The currently most powerful 454/Roche sequencer GS FLX Titanium XL+ can produce up to 1 million reads per run of 700 bp average length, while Illumina's largest machine, HiSeq 2500, can generate up to 1.2 billion paired-end reads of 2 × 150 bp length. Both companies also offer smaller benchtop devices of their platforms that may be preferable in certain diagnostic and clinical settings. The Roche/454 Junior produces up to 100,000 reads of 400 bp average length in a single

10-h run, and the Illumina MiSeq generates up to 30 million paired-end reads of 2 × 150 bp length in 24 h. Thus, longer reads can be produced with the 454/Roche technology, but ultra-deep coverage is easier to obtain with Illumina (Loman et al., 2012).

In addition to the various errors that can occur during sample preparation, as discussed in "Sample Preparation", all NGS platforms introduce sequencing errors. With 454/Roche pyrosequencing, insertions and deletions (indels) are the most common type of errors. They occur predominantly in homopolymeric regions of the target sequence, where the linear relationship between signal intensity and number of incorporated nucleotides starts to fail. Remaining nucleotides after washing can give rise to insertions or carry forward errors, while deletion errors can result from incomplete extension (Margulies et al., 2005; Balzer et al., 2011). The error rate has been shown to increase with read length and to depend on several other biological and technical factors, including the organism and genomic region to be analyzed and the position on the picotiter plate with respect to the flow of chemicals and the position of the camera (Gilles et al., 2011).

Illumina reads are not as susceptible to indel errors in homopolymeric regions, but artificial indels outside these regions and substitutions have similar frequencies (Archer et al., 2012). The Illumina mismatch rate also increases with read length and it further depends on the sequence context and the substitution type (Dohm et al., 2008; Kircher et al., 2009; Nakamura et al., 2011). Illumina reads are generated in forward and reverse direction, and errors predominantly occur on one of the two strands (Chapman et al., 2011; Varela et al., 2011). All NGS platforms report quality scores, defined as $Q = -10 \log_{10} p$, where p is the error probability (Ewing and Green, 1998), together with the called bases, but the calibration of these scores is challenging (Brockman et al., 2008; Kircher et al., 2009) and there is no consensus on how to compare scores across platforms.

Besides errors, the distribution of reads along the genome is critical for diversity estimation, especially if phasing of genetic variants is the goal. However, uniform coverage is difficult to achieve and, in practice, the read coverage often varies by orders of magnitude. The reasons for this variation are poorly understood, but for Illumina, the GC content of the target sequence is an important factor (Dohm et al., 2008). Uniform coverage is feasible within short segments by using a single amplicon. However, increasing the number of amplicons to cover longer segments can impair this uniformity, and shot-gun approaches introduce even more variation. For 454/Roche, Illumina, and ABI SOLiD, correlation of coverage and errors is fairly weak among the three different NGS platforms (Harismendy et al., 2009). Thus, for viral diversity estimation, where uniform coverage and error correction are critical, complementary sequencing strategies involving more than one platform may be more efficient than increasing the coverage on a single platform.

The large amounts of viral sequencing data obtained by NGS place substantial demands on information technology and computational data analysis in terms of storage, quality control, mapping, error correction, single nucleotide variant (SNV) calling, haplotype reconstruction, diversity estimation, and data integration (Pop and Salzberg, 2008; Vrancken et al., 2010; Barzon et al., 2011; Beerenwinkel and Zagordi, 2011). Data analysis usually starts by removing reads of exceptionally low quality. The rationale for this initial filtering step is that low-quality reads contribute disproportionally to the overall error rate, i.e., most errors occur on a few reads (Huse et al., 2007). Filtering can be based on quality scores or on properties of the read or the target sequence known to affect error rates, as discussed above. Optimized filtering has been shown to reduce the error rate in detecting genomic variation up to 300-fold (Reumers et al., 2011).

After filtering, the next step is to align the remaining reads. In re-sequencing studies of known viruses, this is typically done by mapping reads individually to a reference sequence and then aggregating the pairwise alignments into a multiple sequence alignment (MSA). For read mapping, local alignment using dynamic programming may be applied (Wang et al., 2007; Zagordi et al., 2011), but for larger data sets, efficient short read mappers are required. Several efficient mapping algorithms based on indexing techniques are available. Some of them can handle gaps, account for quality scores, and have a paired ends option (Trapnell and Salzberg, 2009; Wikipedia, 2012). In coding regions, a major goal of the alignment step is to identify indels that cause frameshifts. These alterations are likely to be sequencing errors, which are frequently observed using the 454/Roche platform. Hence, they are usually removed, but this bears the risk of losing virus variants harboring real indels. For correcting indel errors, a high-quality alignment is necessary, but in mixed samples, the use of a reference sequence can be suboptimal if reads originating from some subpopulations align only poorly to the reference sequence. To address this concern, a MSA may be computed directly, for example, by using a progressive MSA strategy that takes into account the approximate location of reads on the genome (Saeed et al., 2009). Similarly, for the HIV *env* gene, a multi-step procedure has been proposed, in which reads are located efficiently on a reference sequence by k-mer matching and MSAs are built locally in windows of width 70 nucleotides along the genome. From all local MSAs, in-frame consensus sequences are generated and concatenated. Finally, the reads are re-aligned to the global consensus sequence and all indels causing frameshifts are removed. Using the consensus rather than a reference sequence was shown to improve the alignment quality, especially if their divergence is high (Archer et al., 2010).

LOCAL DIVERSITY ESTIMATION

From the aligned reads, one wants to reconstruct the original virus population in the sample, meaning the composition and relative frequencies of all individual viral genomes, also referred to as strains or haplotypes. Even after filtering and removal of frameshift-causing indels, many reads are still erroneous. Therefore, in mixed samples, error correction and haplotype inference are intrinsically tied to each other and, in fact, addressed jointly by most computational methods. This is in contrast to the simpler task of error correction in clonal samples, where implausible variants can easily be discarded using either k-mers, suffix trees/arrays, or MSA (Yang et al., 2012).

The haplotype inference problem occurs at different spatial scales depending on the length of the genomic region to be analyzed for diversity (**Figure 2**). When only a single genomic site

is considered, diversity estimation means detecting SNVs. Local haplotype inference refers to analyzing windows in the MSA that are covered entirely by reads. Finally, global haplotype inference, also called quasispecies assembly, involves a jigsaw puzzling step of assembling local fragments into multiple haplotype sequences that span the entire genomic region of interest.

SNV calling is based on the observed nucleotide counts at a single sequence position. The simplest statistical model for separating errors from true variations is to assume that, at each genomic site, the number of errors follows the same Poisson distribution and to call SNVs that occur more often than expected by chance for a given error rate (Wang et al., 2007). This approach has been extended to account for site-specific error rates (Macalalad et al., 2012). The power and accuracy of SNV calling can be increased substantially by a control experiment, in which the same genomic region is sequenced from a clonal sample under conditions as similar as possible to those for the mixed sample. The rationale for this comparative sequencing approach is that the control experiment allows for estimating the specific error patterns of the experiment and hence for improved separation of biological signal from technical noise. In this setting, SNV detection is based on comparing nucleotide counts between two experiments, for example, using Fisher's exact test (Koboldt et al., 2012). Assuming independent Poisson distributions, another test is based on the difference of the number of observed nucleotides (Altmann et al., 2011). Count data from NGS experiments have repeatedly been shown to display more variation across sites than is captured by a binomial distribution, and the beta-binomial distribution is a popular choice for such overdispersed data (Flaherty et al., 2012; Gerstung et al., 2012). Based on this model and accounting for the strand-bias of sequencing errors, a sensitivity of up to 1/10,000 has been achieved for SNV calling at a coverage of around 10^5 (Gerstung et al., 2012).

By dropping the assumption of independence among sites, SNV calling can be further improved. Considering the number of joint sequencing errors at two positions has been shown to significantly decrease the minimal frequency at which a variant is detectable (Macalalad et al., 2012). This phasing of two SNVs is possible only at a distance smaller than the maximal read length. For small distances, the SNV pair will be covered by many reads, but for larger distances the benefit of phasing will be undone by the loss of joint coverage. In fact, for deep coverage, pairs are more informative than single sites only if their distance is not larger than the average read length (Macalalad et al., 2012).

The idea of phasing SNVs is further extended by comparing entire reads within a sequence window they overlap. The size of the window is subject to the same trade-off as the distance between two SNVs discussed above: Small windows contain many reads but few SNVs for robust pairwise comparisons of reads, while large windows contain less reads but more segregating sites. Local haplotype inference is based on clustering reads within a given window (**Figure 3**). The rationale for clustering is that reads originating from the same haplotype should be more similar to each other than to reads from other haplotypes. This assumption is only valid if the error rate is low relative to the diversity of the population, and the ability to identify haplotype clusters increases with coverage (Eriksson et al., 2008).

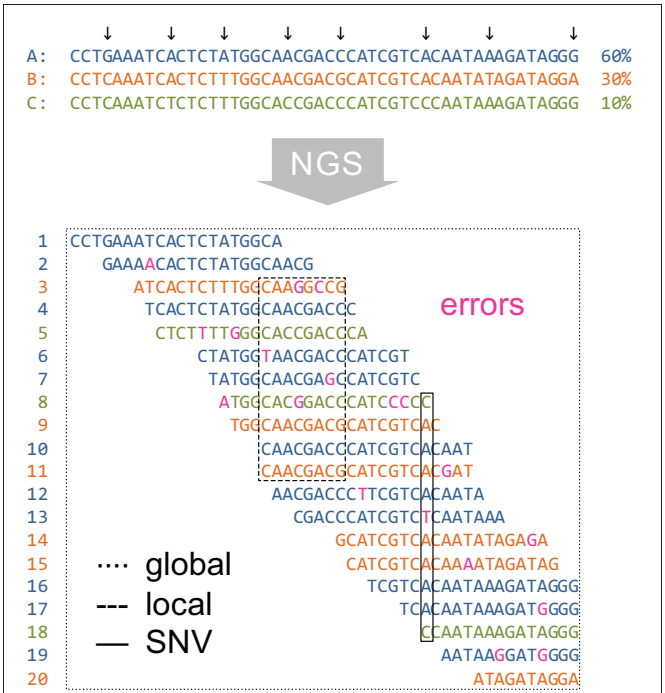


FIGURE 2 | Spatial scales of diversity estimation from NGS data. In this example, it is assumed that the true virus population (top of figure) consists of three haplotypes of relative frequencies 60% (A, blue), 30% (B, orange), and 10% (C, green). Segregating sites are indicated by arrows. Twenty short reads (labeled 1 through 20) are generated by NGS from the virus population subject to sequencing errors (indicated in magenta). Reads are displayed in a MSA and in the color of their corresponding parental haplotype. Diversity estimation can be approached at single sites (SNV detection, solid-line rectangle), in windows of the MSA (local haplotype inference, dashed-line rectangle), or over the entire genomic region (global haplotype reconstruction, dotted-line rectangle).

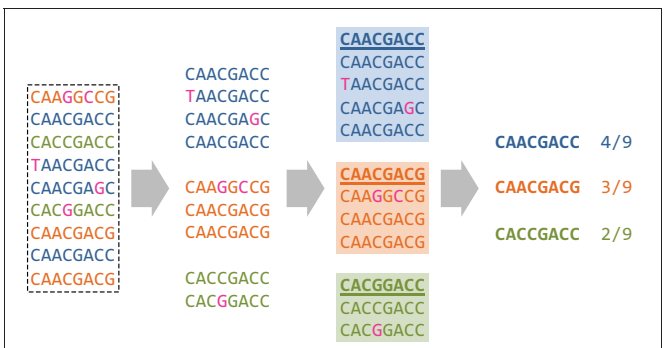


FIGURE 3 | Local read clustering. The local window of the MSA displayed in **Figure 2** is considered (dashed-line rectangle), with colors defined as in **Figure 2**. Reads that are more similar to each other than to other reads are grouped together which recovers the three original haplotypes A, B, and C of **Figure 2** as indicated by the three different colors. Each cluster center sequence is a predicted haplotype (bold, underlined) and the size of its corresponding cluster is an estimate of the frequency of the haplotype (here, 4/f/9, and 2/9, respectively).

Clustering was initially performed using the classical k-means algorithm (Jain and Dubes, 1981) and later formulated probabilistically and solved in a Bayesian fashion (Eriksson et al., 2008; Zagordi et al., 2010a). In particular, the latter approach allows for estimating the error rate and the number of clusters from the data—a notoriously difficult problem with any clustering method. The cluster centers are the predicted haplotypes and the cluster sizes are interpreted as the haplotype frequencies in the population. Error correction is based on a local read clustering solution by replacing all read bases with those of its cluster center (**Figure 3**). This method has been shown to reduce the per-base error rate after correction, to increase the sensitivity and specificity of local haplotype calling, and to improve the estimation of haplotype frequencies as compared to simple read counting or k-means clustering (Zagordi et al., 2010b). For the 454/Roche platform, a similar clustering approach called AmpliconNoise can be applied before base calling on the flowgrams (Quince et al., 2009, 2011). Here, the observed flowgrams are obtained from ideal flowgrams corresponding to read sequences subject to measurement noise. Whether clustering is based on sequences or on flowgrams, the distance measure between reads should reflect the pattern of experimental noise.

As an alternative to clustering, k-mer-based error correction, implemented in the program KEC, has been proposed for viral amplicon sequencing (Skums et al., 2012). This approach extends the EDAR error correction algorithm (Zhao et al., 2010) and initially does not require a read alignment. It consists of a number of heuristic steps with the goal of locating error regions in reads by considering rare k-mers and removing errors in these regions. In a final step, which eventually involves MSAs of the corrected reads, local haplotypes are reconstructed.

GLOBAL DIVERSITY ESTIMATION

The local methods discussed in the previous section focus on reconstructing haplotypes in a local window, the maximum size of which is effectively restricted to the average length of the reads. The global reconstruction problem, on the other hand, is defined as the genome-wide assembly of quasispecies, irrespective of machine-specific parameters like the average read length. The various approaches to solving this jigsaw puzzle described in the literature can be roughly divided into three groups: (1) graph-based methods that first aggregate the reads in a read graph and then search for a minimum set of paths through this graph, (2) probabilistic clustering models based on mixture models, and (3) *de novo* assembly methods which do not rely on the availability of a reference sequence.

Read graph-based global haplotype reconstruction consists in aggregating the reads in a read graph and subsequently identifying haplotypes as paths in this graph. The concept of a read graph has been independently introduced by Eriksson et al. (2008) and Westbrooks et al. (2008). The read graph contains the possibly pre-processed, for instance, locally error-corrected, reads as nodes. Directed edges connect two nodes when the reads agree on their non-empty overlap (**Figure 4**). The direction of the edge reflects the order of the starting positions on the reference sequence. The set of nodes is restricted to all irredundant reads, where a read is considered redundant if there is another read that overlaps completely and if both reads agree on this overlap. In a similar manner, the set of edges is restricted to include only those edges for which there would be no path between the corresponding nodes without this edge. The latter restriction is called transductive reduction in (Westbrooks et al., 2008), and it has been shown that this reduction can be computed efficiently. Finally, a source and a sink node are added to the graph, along with edges connecting all reads starting at the first position to the source and all reads ending at the last position to the sink (**Figure 4**).

Every path in the read graph connecting source and sink is a potential haplotype, and the problem of estimating the haplotypes present in a certain sample might be restated as finding a set of such source-sink paths that explains the reads well. Different formalizations of this problem lead to different optimization problems. One example is the search for the minimum set of paths that covers all reads implemented in ShoRAH (Eriksson et al., 2008; Zagordi et al., 2011). The same problem has been studied in a different way as a network flow problem (Westbrooks et al., 2008). A variant of the network flow formulation is the search for a set of haplotypes covering all reads with minimum costs (Westbrooks et al., 2008) and, in a slightly different

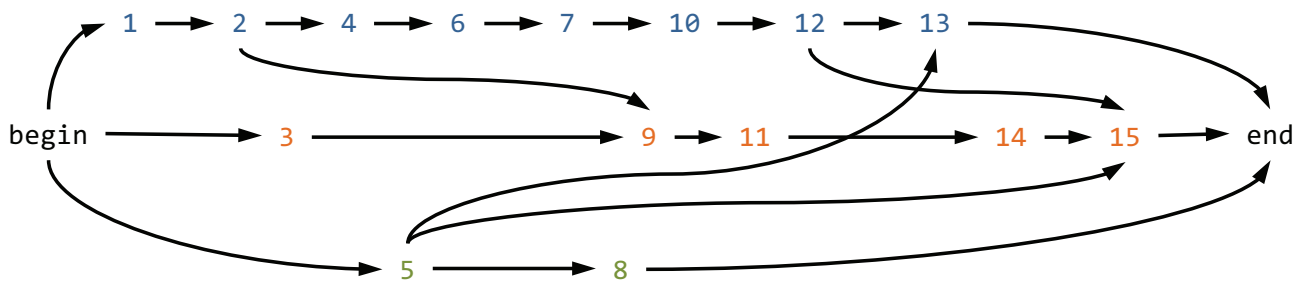


FIGURE 4 | Read graph-based global haplotype reconstruction. Shown is the read graph for the first 15 reads of the MSA shown in **Figure 2**. Each read is represented by its index and colored according to its parental haplotype (A, blue, first row; B, orange, second row; and C, green, third row). Reads are connected by a direct edge if they agree on their non-empty overlap. Each path from the begin node to the end node represents a potential global haplotype, but there are more paths in the graph than the original three haplotypes the reads have been derived from.

fashion relaxing the requirement of a complete read cover, implemented in ViSpA (Astrovskaya et al., 2011). The combinatorial reconstruction is followed by frequency estimation using an Expectation Maximization (EM) algorithm (Eriksson et al., 2008; Westbrooks et al., 2008; Astrovskaya et al., 2011).

In a related approach termed QuRe, the same read graph idea is used to find a set of consistent quasispecies explaining the reads (Prosperi et al., 2011; Prosperi and Salemi, 2012). It differs from the methods above in the optimization procedure for finding the quasispecies. This is formalized as minimizing the number of *in silico* recombinants instead of finding a path cover explaining the reads. However, both optimization strategies are similar in nature, since avoiding *in silico* recombinants can be regarded as avoiding redundant paths in the read graph. Another advantage of QuRe is that it explicitly addresses the blockwise structure of the reads due to amplicon-based sequencing in the statistical analysis (Prosperi et al., 2011; Prosperi and Salemi, 2012).

Haplotype assembly based on amplicon sequencing is also addressed by the BIOA software (Mancuso et al., 2011). Here, a read graph-based framework is proposed that includes balancing of haplotype frequencies between neighboring amplicons followed by quasispecies reconstruction using a maximum bandwidth approach or a greedy algorithm. In the assembly step, the parsimony criterion of explaining the observed reads with a minimal number of haplotypes is relaxed to finding a quasispecies of minimal entropy explaining the reads. This strategy was shown to outperform shotgun-based quasispecies assembly using ViSpA.

QColors is another method that relies on the read graph as the main source of information for assembling reads into haplotypes, but it uses in addition a conflict graph consisting of edges between reads that overlap but disagree on the overlap (Huang et al., 2011). The reconstruction problem is then to find a partition of the reads into a minimal number of non-conflicting subsets, which defines a vertex graph coloring problem, hence the name QColors. A potential problem with this approach might be the sensitivity of the conflict graph to sequencing errors and the uncertainty in placing alignment gaps, which are not explicitly dealt with.

Another method that uses the read graph approach is called Hapler (O'Neil and Emrich, 2012). This method is specifically designed for situations characterized by low haplotype diversity and low read coverage ($<25\times$), which, for instance, occur in the context of population-level *de novo* transcriptome assemblies or ecological studies. The minimum path cover problem is generalized and reformulated as a weighted bipartite graph matching problem, such that erroneous reads can be identified. Since, in general, the resulting path covers are again not unique, the analysis is equipped with a randomization step in which samples are drawn from the set of path covers, although this process seems to lack a clear probabilistic interpretation. Experiments under low-coverage conditions indicate that this method is successful in reconstructing local haplotypes over a region that is roughly determined by the average read length, which in our terminology would be classified as local reconstruction. Nevertheless, longer haplotype assemblies are possible with Hapler and specific care is taken in reconstructing consensus sequences with a minimal number of chimeric points.

A common property of all read graph-based approaches is that the haplotype reconstruction problem itself becomes deterministic in nature, while the unavoidable noise component present in observed reads is dealt with in a pre-processing error correction step—if at all.

Removing all the stochasticity in the observed reads by way of local error correction prior to global haplotype reconstruction has the limitation that corrections cannot be revised in the global context and miscorrections are propagated through subsequent steps. A probabilistic hierarchical model that circumvents this problem has been introduced (Jojic et al., 2008). The main idea is to model the generative stochastic process of read generation. Parameters and hidden variables in this method include the parental haplotype, the starting position, and the parameters related to the error transformation. Inference is carried out by maximizing the likelihood using the EM algorithm. A potential drawback of this approach is that the user has to fix the number of haplotypes to be reconstructed in advance, and no well-defined estimation process for this number is provided.

Probabilistic approaches are a second methodology for global haplotype reconstruction. PredictHaplo is one of these approaches which also automatically adjusts the number of haplotypes (Prabhakaran et al., 2010). In this model, a haplotype is represented as a set of position-specific probability tables over the four nucleotides, which can be augmented to include a fifth character representing alignment gaps (**Figure 5**). The underlying generative model assumes that reads are sampled from a mixture model, where each mixture component is interpreted as a haplotype, and the associated mixing proportion estimates the haplotype frequency. In order to avoid a priori specification of the number of mixture components, an infinite mixture model is employed (Ewens, 1972; Ferguson, 1973; Rasmussen, 2000), and for computational reasons, a truncated approximation of this stochastic process is used.

A further refinement of probabilistic haplotype reconstruction has been implemented in the program QuasiRecomb (Zagordi et al., 2012). Here, haplotypes are not reconstructed individually, but rather their distribution is estimated by a hidden Markov model. The model assumes that all haplotypes are generated from a small set of sequences by mutation and recombination. This model is taking into account that in some RNA viruses, such as HIV, recombination is very frequent and hence an important factor generating genetic diversity.

All approaches described so far make use of a known reference genome that serves as a fixed spatial coordinate system after read alignment. By contrast, *de novo* assembly methods are more general in nature since they do not require such reference genomes. Several assemblers specifically designed for certain NGS platforms like 454/Roche have been proposed in recent years (Finotello et al., 2012). The original goal of *de novo* assembly is reconstructing a single target genome sequence, rather than an ensemble of different genomes. Hence, the currently available genome assemblers are not designed to solve the whole-genome quasispecies assembly problem, but the different contigs they reconstruct may

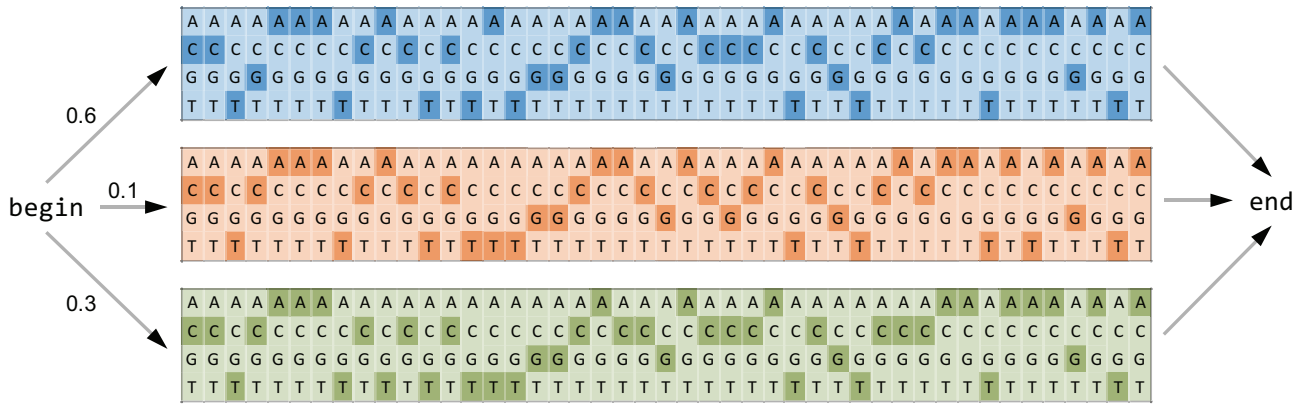


FIGURE 5 | Probabilistic global haplotype reconstruction using a generative mixture model. Each of the three haplotypes colored as in **Figure 2** (A, blue; B, orange; and C, green) is represented as a chain of probability tables over the four nucleotides, where darker shading of a base indicates higher probability. The probabilities of traversing from the begin node to one of the haplotypes serve as an estimate for the haplotype frequencies. Each read is regarded as an independent observation from this statistical model.

serve as a starting point for this jigsaw puzzle (Ramakrishnan et al., 2009).

Large-scale simulation studies show that all global reconstruction methods rely on the availability of relatively long reads. Coverage is also important when it comes to detecting low-abundant mutants, but even an arbitrarily high coverage cannot compensate for insufficient overlaps due to short reads. Given the typical diversity of virus populations, it appears that global haplotype reconstruction is currently only realistic for sequencing platforms producing long reads on the order of at least 300–500 bp. Accordingly, successful reconstructions have been reported predominantly for the 454/Roche sequencing platform.

Regarding the different computational approaches described above, it is generally difficult to conduct informative comparative simulation experiments, but two general trends have become evident. First, local read error correction has the tendency to under-correct the reads, which can lead to a large number of false positive global haplotypes, in particular, when combined with read graph approaches requiring a complete coverage of all reads. Quasispecies assembly methods that relax this coverage requirement (Astrovskaya et al., 2011; O'Neil and Emrich, 2012) or probabilistic approaches avoiding the read-graph construction (Jojic et al., 2008; Prabhakaran et al., 2010) are successful in decreasing the false positive rate. Second, the most problematic step in genome-wide reconstruction is the usually unavoidable (RT-)PCR pre-processing which can introduce significant artifacts. These artifacts might have a much stronger effect on the final quality of the haplotype reconstruction than the actual choice of the computational reconstruction method.

Computational methods for local and global haplotype reconstruction are summarized in **Table 1**. All of these tools have been developed in research environments and most are subject to continuous enhancements. Their usability and performance also depends on the quickly changing characteristics of the sequencing machines. In the future, comparative studies using simulated data, mixed control samples, or Sanger-sequenced gold standard samples are required to assess the performance of these tools under different conditions. In addition, software tools are available for NGS read data management and visualization. For example, Segminator II has been specifically designed to display sequence variability of temporally sampled virus populations (Archer et al., 2012).

Table 1 | Available software tools for viral quasispecies inference.

Program	Method	URL	References
QuRe	Read graph	https://sourceforge.net/projects/qure/	Prosperi and Salemi, 2012
ShoRAH	Read graph	http://www.cbg.ethz.ch/software/shorah	Zagordi et al., 2011
ViSpA	Read graph	http://alla.cs.gsu.edu/~software/VISPA/vispa.html	Astrovskaya et al., 2011
BIOA	Read graph	https://bitbucket.org/nmancuso/bioa/	Mancuso et al., 2011
Hapler	Read graph	http://nd.edu/~biocmp/hapler/	O'Neil and Emrich, 2012
AmpliconNoise	Probabilistic	http://code.google.com/p/ampliconnoise	Quince et al., 2011
PredictHaplo	Probabilistic	http://www.cs.unibas.ch/personen/roth_volker/HivHaploTyper	Prabhakaran et al., 2010
QuasiRecomb	Probabilistic	http://www.cbg.ethz.ch/software/quasirecomb	Zagordi et al., 2012

Table 2 | Applications of 454/Roche pyrosequencing and Illumina NGS technologies in clinical virology.

Virus	Study	NGS platform	NGS approach	Basis of analysis	References
CMV	Epidemiology	454/Roche	Amplicon-based	Reads	Gorzer et al., 2010
CMV	Epidemiology	454/Roche	Shotgun	Consensus sequence	Jung et al., 2011
EBV	Epidemiology	Illumina	Shotgun	SNV, consensus sequence	Liu et al., 2011
EBV	Epidemiology	Illumina	Shotgun (amplicons)	SNV	Kwok et al., 2012
HBV	Drug resistance	454/Roche	Amplicon-based	Reads, SNV	Solmone et al., 2009; Homs et al., 2011; Rodriguez-Frías et al., 2012
HBV	Drug resistance	454/Roche	Amplicon-based	SNV	Margeridon-Thermet et al., 2009; Ko et al., 2012; Sede et al., 2012
HBV	Drug resistance	Illumina	Shotgun	SNV	Nishijima et al., 2012
HCV	Drug resistance	454/Roche	Amplicon-based	Reads	Bolcic et al., 2012; Fonseca-Coronado et al., 2012
HCV	Drug resistance	Illumina	Shotgun (cDNA)	SNV	Hiraga et al., 2011
HCV	Drug resistance	454/Roche	Shotgun (amplicons)	SNV, consensus sequences	Lauck et al., 2012
HCV	Drug resistance	Illumina	Paired-end (amplicons)	SNV	Nasu et al., 2011
HCV	Drug resistance	454/Roche	Amplicon-based	SNV	Powdrill et al., 2011
HCV	Epidemiology	454/Roche	Amplicon-based	Reads	Escobar-Gutiérrez et al., 2012; Forbi et al., 2012
HCV	Epidemiology	Illumina	Shotgun (cDNA)	SNV, consensus sequences	Ninomiya et al., 2012
HIV	Drug resistance	454/Roche	Amplicon-based	SNV	Hoffmann et al., 2007; Wang et al., 2007; Mitsuya et al., 2008; Le et al., 2009; Simen et al., 2009; Varghese et al., 2009; Lataillade et al., 2010, 2012; Alteri et al., 2011; D'Aquila et al., 2011; Delobel et al., 2011; Gianella et al., 2011; Ji et al., 2011; Kozal et al., 2011; Moorthy et al., 2011; Stelzl et al., 2011; Fisher et al., 2012; Messiaen et al., 2012
HIV	Drug resistance	454/Roche	Amplicon-based	Reads, SNV	Hedskog et al., 2010; Ji et al., 2010; Mild et al., 2011; Mukherjee et al., 2011; Armenia et al., 2012
HIV	Epidemiology	454/Roche	Shotgun (amplicons)	Consensus sequence	Bruselles et al., 2009
HIV	Epidemiology	454/Roche	Amplicon-based	Consensus sequence	Eshleman et al., 2011
HIV	Epidemiology	454/Roche	Amplicon-based	Reads	Redd et al., 2012
HIV	Tropism	454/Roche	Amplicon-based	Reads	Archer et al., 2009; Rozera et al., 2009; Abbate et al., 2011; Swenson et al., 2010; Vandenbroucke et al., 2010; Baatz et al., 2011; Bunnik et al., 2011; Raymond et al., 2011; Saliou et al., 2011; Svicher et al., 2011; Swenson et al., 2011a,b; Vandekerckhove et al., 2011
Influenza A virus	Epidemiology	Illumina	Shotgun (amplicons)	SNV	Kuroda et al., 2010; Kampmann et al., 2011
Influenza A virus	Epidemiology	454/Roche	Shotgun (amplicons)	SNV	Bartolini et al., 2011
Influenza A virus	Epidemiology	454/Roche	Shotgun	Reads	Lorusso et al., 2011
norovirus	Epidemiology	454/Roche	Shotgun (amplicons)	SNV, haplotype recon-struction	Bull et al., 2012
rhinovirus	Epidemiology	Illumina	Shotgun (amplicons)	SNV, consensus sequences	Tapparel et al., 2011
rotavirus	Epidemiology	454/Roche	Shotgun (cDNA)	Consensus sequences	Jere et al., 2011
VZV	Epidemiology	454/Roche	Shotgun (amplicons)	Consensus sequences	Zell et al., 2012

BAL, bronchoalveolar lavage; CMV, cytomegalovirus; EBV, Epstein Barr virus; HBV, hepatitis B virus; HCV, hepatitis C virus; HIV, human immunodeficiency virus; SNV, single nucleotide variant; VZV, varicella zoster virus.

APPLICATIONS

NGS is widely applied to study viral diversity mainly in the context of drug resistance of clinically relevant viruses such as HIV, HCV, and HBV (**Table 2**). Most studies focus on pre-existing minority drug-resistant virus variants in treatment-naïve individuals and their impact on the success of antiviral therapy, epidemiological surveillance, and virus population dynamics during virological failure. The pathways of drug resistance development are of particular clinical importance, since they can lead to new drug design or new therapeutic strategies, for instance, avoiding cross resistance or rapid selection of resistant viruses (Beerenwinkel et al., 2003). Furthermore, epidemiological studies for a huge variety of human pathogenic viruses were performed using NGS technologies, including cytomegalovirus (CMV), Epstein Barr virus (EBV), HCV, influenza virus, norovirus, rhinovirus, rotavirus, and varicella zoster virus (VZV) (**Table 2**).

NGS is also increasingly used in more basic research areas, such as characterization of transmitted HIV (Fischer et al., 2010) and HCV (Wang et al., 2010; Bull et al., 2011), estimation of infection dates (Poon et al., 2011), evolution during the course of infection with HIV (Rozera et al., 2009; Poon et al., 2010; Wu et al., 2011), HCV (Bull et al., 2011), and rhinovirus (Cordey et al., 2010), and hypermutation patterns (Reuman et al., 2010; Knoepfel et al., 2011). Recently, NGS technologies have been applied to obtain the whole genome of HIV using a coverage allowing quasispecies analysis beyond the generation of consensus sequences to study, for instance, patterns of immune escape (Bimber et al., 2010; Willerth et al., 2010; Henn et al., 2012).

All these applications demonstrate the growing importance of NGS in studying viral diversity. With this technology, we will gain further insights into transmission traits, viral evolution, and its association with pathogenesis. World-wide viral diversity surveillance will be important for vaccine design and vaccination strategies. Currently, genetic diversity is mainly studied based on the detection and analyses of SNVs, rather than the reconstruction of linked mutations, due to the challenges in local and global haplotype reconstruction discussed above. It will be a huge step forward when haplotype reconstruction in heterogeneous viruses matures into a routine procedure based on standardized experimental protocols and validated, automatic data analysis pipelines.

OUTLOOK AND CONCLUSIONS

NGS opens up new roads to study viral diversity. It will tremendously increase our knowledge in virus evolution, fitness, selection pathways, and pathogenesis. Together with host genomics, viral diversity will allow insights into complex virus-host interactions. Full-length viral sequences may ultimately define truly conserved regions in viral genomes which might also be of relevance for vaccine and drug design. Clinically, the first application we can foresee is that in a single assay all drug targets relevant for antiviral treatment can be sequenced including information on minority drug-resistant variants. For all applications, sample procedures have to be chosen that minimize errors during sample preparation and sequencing. Several challenges in data analysis remain, especially in regard to alignments and global diversity estimation. In the future, some of these challenges might be diminished by upcoming third- and fourth-generation sequencing technologies, like single molecule or direct RNA sequencing.

Another not yet addressed future challenge will be making sense of the large amounts of genome data generated by NGS. For instance, clinical cut-offs need to be defined for minority drug-resistant virus variants, the clinical importance of new virus subtypes or even new viruses needs to be determined, and pathogenesis factors need to be confirmed in clinical settings. Thus, downstream analyses have to include large sets of well-documented patients, results from other experimental setups, etc. These are challenges as well as opportunities to answer important research questions which could not be addressed with conventional sequencing techniques.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation under grant number CR32I2_127017.

REFERENCES

Abbate, I., Vlassi, C., Rozera, G., Bruselles, A., Bartolini, B., Giombini, E., Corpolongo, A., D'Offizi, G., Narciso, P., Desideri, A., Ippolito, G., and Capobianchi, M. R. (2011). Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* 25, 611–617.

Alteri, C., Santoro, M. M., Abbate, I., Rozera, G., Bruselles, A., Bartolini, B., Gori, C., Forbici, F., Orchi, N., Tozzi, V., Palamara, G., Antinori, A., Narciso, P., Girardi, E., Svicher, V., Ceccherini-Silberstein, F., Capobianchi, M. R., and Perno, C. F. (2011). 'Sentinel' mutations in standard population sequencing can predict the presence of HIV-1 reverse transcriptase major mutations detectable only by ultra-deep pyrosequencing. *J. Antimicrob. Chemother.* 66, 2615–2623.

Althaus, C. F., Vongrad, V., Niederost, B., Joos, B., Di Giallonardo, F., Rieder, P., Pavlovic, J., Trkola, A., Gunthard, H. F., Metzner, K. J., and Fischer, M. (2012). Tailored enrichment strategy detects low abundant small noncoding RNAs in HIV-1 infected cells. *Retrovirology* 9, 27.

Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E. B., and Müller-Myhsok, B. (2011). vipR: variant identification in pooled DNA using R. *Bioinformatics* 27, i77–i84.

Archer, J., Baillie, G., Watson, S. J., Kellam, P., Rambaut, A., and Robertson, D. L. (2012). Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator, I. I. *BMC Bioinformatics* 13, 47.

Archer, J., Braverman, M. S., Taillon, B. E., Desany, B., James, I., Harrigan, P. R., Lewis, M., and Robertson, D. L. (2009). Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *AIDS* 23, 1209–1218.

Archer, J., Rambaut, A., Taillon, B. E., Harrigan, P. R., Lewis, M., and Robertson, D. L. (2010). The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time–an ultra-deep approach. *PLoS Comput. Biol.* 6:e1001022. doi: 10.1371/journal.pcbi.1001022

Armenia, D., Vandenbroucke, I., Fabeni, L., Van Marck, H., Cento, V., D'Arrigo, R., Van Wesenbeeck, L., Scopelliti, F., Micheli, V., Bruzzone, B., Lo Caputo, S., Aerssens, J., Rizzardini, G., Tozzi, V., Narciso, P., Antinori, A., Stuyver, L., Perno, C. F., and Ceccherini-Silberstein, F. (2012). Study of genotypic and phenotypic HIV-1 dynamics of integrase mutations during raltegravir treatment: a refined analysis by ultra-deep 454 pyrosequencing. *J. Infect. Dis.* 205, 557–567.

Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., and Zelikovsky, A. (2011). Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics* 12(Suppl. 6), S1.

Baatz, F., Struck, D., Lemaire, M., De Landtsheer, S., Servais, J. Y., Arendt, V., Schmit, J. C., and Perez Bercoff, D. (2011). Rescue of HIV-1 long-time archived X4 strains to escape maraviroc. *Antiviral Res.* 92, 488–492.

Balzer, S., Malde, K., and Jonassen, I. (2011). Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics* 27, i304–i309.

Bartolini, B., Chillemi, G., Abbate, I., Bruselles, A., Rozera, G., Castrignano, T., Paoletti, D., Picardi, E., Desideri, A., Pesole, G., and Capobianchi, M. R. (2011). Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing. *New Microbiol.* 34, 391–397.

Barzon, L., Lavezzo, E., Militello, V., Toppo, S., and Palù, G. (2011). Applications of next-generation sequencing technologies to diagnostic virology. *Int. J. Mol. Sci.* 12, 7861–7884.

Beerenwinkel, N., Lengauer, T., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., and Selbig, J. (2003). Methods for optimizing antiviral combination therapies. *Bioinformatics* 19(Suppl. 1), i16–i25.

Beerenwinkel, N., and Zagordi, O. (2011). Ultra-deep sequencing for the analysis of viral populations. *Curr. Opin. Virol.* 1, 413–418.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Cheetham, R. K., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Catenazzi, M. C. E., Chang, S., Cooley, R. N., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fajardo, K. V. F., Furey, W. S., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59.

Bimber, B. N., Dudley, D. M., Lauck, M., Becker, E. A., Chin, E. N., Lank, S. M., Grunenwald, H. L., Caruccio, N. C., Maffitt, M., Wilson, N. A., Reed, J. S., Sosman, J. M., Tarosso, L. F., Sanabani, S., Kallas, E. G., Hughes, A. L., and O'Connor, D. H. (2010). Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultra-deep pyrosequencing. *J. Virol.* 84, 12087–12092.

Bolcic, F., Sede, M., Moretti, F., Westergaard, G., Vazquez, M., Laufer, N., and Quarleri, J. (2012). Analysis of the PKR-eIF2alpha phosphorylation homology domain (PePHD) of hepatitis C virus genotype 1 in HIV-coinfected patients by ultra-deep pyrosequencing and its relationship to responses to pegylated interferon-ribavirin treatment. *Arch. Virol.* 157, 703–711.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* 18, 763–770.

Bruselles, A., Rozera, G., Bartolini, B., Prosperi, M., Del Nonno, F., Narciso, P., Capobianchi, M. R., and Abbate, I. (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.

Bull, R. A., Eden, J.-S., Luciani, F., McElroy, K., Rawlinson, W. D., and White, P. A. (2012). Contribution of intra- and interhost dynamics to norovirus evolution. *J. Virol.* 86, 3219–3229.

Bull, R. A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S. T., Chopra, A., Cameron, B., Maher, L., Dore, G. J., White, P. A., and Lloyd, A. R. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog.* 7:e1002243. doi: 10.1371/journal.ppat.1002243

Bunnik, E. M., Swenson, L. C., Edo-Matas, D., Huang, W., Dong, W., Frantzell, A., Petropoulos, C. J., Coakley, E., Schuitemaker, H., Harrigan, P. R., and van 't Wout, A. B. (2011). Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra-deep pyrosequencing. *PLoS Pathog.* 7:e1002106. doi: 10.1371/journal.ppat.1002106

Burch, C. L., and Chao, L. (2000). Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* 406, 625–628.

Casbon, J. A., Osborne, R. J., Brenner, S., and Lichtenstein, C. P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* 39, e81.

Chang, K. S., Vyas, R. C., Deaven, L. L., Trujillo, J. M., Stass, S. A., and Hittelman, W. N. (1992). PCR amplification of chromosome-specific DNA isolated from flow cytometry-sorted chromosomes. *Genomics* 12, 307–312.

Chapman, M. A., Lawrence, M. S., Keats, J. J., Cibulskis, K., Sougnez, C., Schinzel, A. C., Harview, C. L., Brunet, J.-P., Ahmann, G. J., Adli, M., Anderson, K. C., Ardlie, K. G., Auclair, D., Baker, A., Bergsagel, P. L., Bernstein, B. E., Drier, Y., Fonseca, R., Gabriel, S. B., Hofmeister, C. C., Jagannath, S., Jakubowiak, A. J., Krishnan, A., Levy, J., Liefeld, T., Lonial, S., Mahan, S., Mfuko, B., Monti, S., Perkins, L. M., Onofrio, R., Pugh, T. J., Rajkumar, S. V., Ramos, A. H., Siegel, D. S., Sivachenko, A., Stewart, A. K., Trudel, S., Vij, R., Voet, D., Winckler, W., Zimmerman, T., Carpten, J., Trent, J., Hahn, W. C., Garraway, L. A., Meyerson, M., Lander, E. S., Getz, G., and Golub, T. R. (2011). Initial genome sequencing and analysis of multiple myeloma. *Nature* 471, 467–472.

Cordey, S., Junier, T., Gerlach, D., Gobbini, F., Farinelli, L., Zdobnov, E. M., Winther, B., Tapparel, C., and Kaiser, L. (2010). Rhinovirus genome evolution during experimental human infection. *PLoS ONE* 5:e10588. doi: 10.1371/journal.pone.0010588

D'Aquila, R. T., Geretti, A. M., Horton, J. H., Rouse, E., Kheshti, A., Raffanti, S., Oie, K., Pappa, K., and Ross, L. L. (2011). Tenofovir (TDF)-selected or abacavir (ABC)-selected low-frequency HIV type 1 subpopulations during failure with persistent viremia as detected by ultradeep pyrosequencing. *AIDS Res. Hum. Retroviruses* 27, 201–209.

Daly, G. M., Bexfield, N., Heaney, J., Stubbs, S., Mayer, A. P., Palser, A., Kellam, P., Drou, N., Caccamo, M., Tiley, L., Alexander, G. J., Bernal, W., and Heeney, J. L. (2011). A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS ONE* 6:e28879. doi: 10.1371/journal.pone.0028879

Delobel, P., Saliou, A., Nicot, F., Dubois, M., Trancart, S., Tangre, P., Aboulker, J. P., Taburet, A. M., Molina, J. M., Massip, P., Marchou, B., and Izopet, J. (2011). Minor HIV-1 variants with the K103N resistance mutation during intermittent Efavirenz-containing antiretroviral therapy and virological failure. *PLoS ONE* 6:e21655. doi: 10.1371/journal.pone.0021655

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105.

Domingo, E., and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51, 151–178.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003). Measurably evolving populations. *Trends Ecol. Evol.* 18, 481–488.

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276.

Eckert, K. A., and Kunkel, T. A. (1991). DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.* 1, 17–24.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58, 465–523.

Eigen, M., McCaskill, J., and Schuster, P. (1988). Molecular quasi-species. *J. Phys. Chem.* 92, 6881–6891.

Eigen, M., McCaskill, J., and Schuster, P. (1989). The molecular quasi-species. *Adv. Chem. Phys.* 75, 149–263.

Eigen, M., and Schuster, P. (1977). The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften* 64, 541–565.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W., and Beerenwinkel, N. (2008). Viral population estimation using pyrosequencing. *PLoS Comput. Biol.* 4:e1000074. doi: 10.1371/journal.pcbi.1000074

Escobar-Gutiérrez, A., Vazquez-Pichardo, M., Cruz-Rivera, M., Rivera-Osorio, P., Carpio-Pedroza, J. C., Ruíz-Pacheco, J. A., Ruiz-Tovar, K., and Vaughan, G. (2012). Identification of hepatitis C virus transmission using a next-generation sequencing approach. *J. Clin. Microbiol.* 50, 1461–1463.

Eshleman, S. H., Hudelson, S. E., Redd, A. D., Wang, L., Debes, R., Chen, Y. Q., Martens, C. A., Ricklefs, S. M., Selig, E. J., Porcella, S. F., Munshaw, S., Ray, S. C., Piwowar-Manning, E., McCauley, M., Hosseinipour, M. C., Kumwenda, J., Hakim, J. G., Chariyalertsak, S., De Bruyn, G., Grinsztejn, B., Kumarasamy, N., Makhema, J., Mayer, K. H., Pilotto, J., Santos, B. R., Quinn, T. C., Cohen, M. S., and Hughes, J. P. (2011). Analysis of genetic linkage of HIV from couples enrolled in the HIV prevention trials network 052 trial. *J. Infect. Dis.* 204, 1918–1926.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3, 87–112.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194.

Fang, G., Zhu, G., Burger, H., Keithly, J. S., and Weiser, B. (1998). Minimizing DNA recombination during long RT-PCR. *J. Virol. Methods* 76, 139–148.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *Ann. Stat.* 1, 209–230.

Finotello, F., Lavezzo, E., Fontana, P., Peruzzo, D., Albiero, A., Barzon, L., Falda, M., Camillo, B. D., and Toppo, S. (2012). Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Brief Bioinform.* 13, 269–280.

Fischer, M., Wong, J. K., Russenberger, D., Joos, B., Opravil, M., Hirschel, B., Trkola, A., Kuster, H., Weber, R., and Gunthard, H. F. (2002). Residual cell-associated unspliced HIV-1 RNA in peripheral blood of patients on potent antiretroviral therapy represents intracellular transcripts. *Antivir. Ther.* 7, 91–103.

Fischer, W., Ganusov, V. V., Giorgi, E. E., Hraber, P. T., Keele, B. F., Leitner, T., Han, C. S., Gleasner, C. D., Green, L., Lo, C. C., Nag, A., Wallstrom, T. C., Wang, S., McMichael, A. J., Haynes, B. F., Hahn, B. H., Perelson, A. S., Borrow, P., Shaw, G. M., Bhattacharya, T., and Korber, B. T. (2010). Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5:e12303. doi: 10.1371/journal.pone.0012303

Fisher, R., Van Zyl, G. U., Travers, S. A., Pond, S. L. K., Engelbrech, S., Murrell, B., Scheffler, K., and Smith, D. (2012). Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen. *J. Virol.* 86, 6231–6237.

Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., Bell, J., Brown, S., Holodniy, M., Zhang, N., and Ji, H. P. (2012). Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.* 40, e2.

Fonseca-Coronado, S., Escobar-Gutierrez, A., Ruiz-Tovar, K., Cruz-Rivera, M. Y., Rivera-Osorio, P., Vazquez-Pichardo, M., Carpio-Pedroza, J. C., Ruiz-Pacheco, J. A., Cazares, F., and Vaughan, G. (2012). Specific detection of naturally occurring hepatitis C virus mutants with resistance to telaprevir and boceprevir (protease inhibitors) among treatment-naive infected individuals. *J. Clin. Microbiol.* 50, 281–287.

Forbi, J. C., Purdy, M. A., Campo, D. S., Vaughan, G., Dimitrova, Z. E., Ganova-Raeva, L. M., Xia, G. L., and Khudyakov, Y. E. (2012). Epidemic history of hepatitis C virus infection in two remote communities in Nigeria, West Africa. *J. Gen. Virol.* 93, 1410–1421.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumor cell populations. *Nat. Commun.* 3, 811.

Gianella, S., Delport, W., Pacold, M. E., Young, J. A., Choi, J. Y., Little, S. J., Richman, D. D., Pond, S. L. K., and Smith, D. M. (2011). Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. *J. Virol.* 85, 8359–8367.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., and Martin, J.-F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12, 245.

Gorzer, I., Guelly, C., Trajanoski, S., and Puchhammer-Stockl, E. (2010). Deep sequencing reveals highly complex dynamics of human cytomegalovirus genotypes in transplant patients over time. *J. Virol.* 84, 7195–7203.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 10, R32.

Hedskog, C., Mild, M., Jernberg, J., Sherwood, E., Bratt, G., Leitner, T., Lundeberg, J., Andersson, B., and Albert, J. (2010). Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS ONE* 5:e11345. doi: 10.1371/journal.pone.0011345

Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. A., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Gunthard, H. F., Brumme, Z. L., Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D., and Allen, T. M. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.* 8:e1002529. doi: 10.1371/journal.ppat.1002529

Hiraga, N., Imamura, M., Abe, H., Hayes, C. N., Kono, T., Onishi, M., Tsuge, M., Takahashi, S., Ochi, H., Iwao, E., Kamiya, N., Yamada, I., Tateno, C., Yoshizato, K., Matsui, H., Kanai, A., Inaba, T., Tanaka, S., and Chayama, K. (2011). Rapid emergence of telaprevir resistant hepatitis C virus strain from wild-type clone *in vivo*. *Hepatology* 54, 781–788.

Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M. Q., Tebas, P., and Bushman, F. D. (2007). DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res.* 35, e91.

Holodniy, M., Mole, L., Yen-Lieberman, B., Margolis, D., Starkey, C., Carroll, R., Spahlinger, T., Todd, J., and Jackson, J. B. (1995). Comparative stabilities of quantitative human immunodeficiency virus RNA in plasma from samples collected in VACUTAINER CPT, VACUTAINER PPT, and standard VACUTAINER tubes. *J. Clin. Microbiol.* 33, 1562–1566.

Homs, M., Buti, M., Quer, J., Jardi, R., Schaper, M., Tabernero, D., Ortega, I., Sanchez, A., Esteban, R., and Rodriguez-Frias, F. (2011). Ultra-deep pyrosequencing analysis of the hepatitis B virus preCore region and main catalytic motif of the viral polymerase in the same viral genome. *Nucleic Acids Res.* 39, 8457–8471.

Huang, A., Kantor, R., Delong, A., Schreier, L., and Istrail, S. (2011). "QColors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous

next generation sequencing reads," in *IEEE International Conference on Bioinformatics and Biomedicine Workshops*. Publisher is Institute of Electrical and Electronics Engineers (IEEE), 130–136.

Huse, S., Huber, J., Morrison, H., Sogin, M., and Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143.

Iwasa, Y., Michor, F., and Nowak, M. A. (2003). Evolutionary dynamics of escape from biomedical intervention. *Proc. Biol. Sci.* 270, 2573–2578.

Iwasa, Y., Michor, F., and Nowak, M. A. (2004). Evolutionary dynamics of invasion and escape. *J. Theor. Biol.* 226, 205–214.

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, I. D. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20166–20171.

Jain, A. K., and Dubes, R. C. (1981). *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall.

Jere, K. C., Mlera, L., Page, N. A., Van Dijk, A. A., and O'neill, H. G. (2011). Whole genome analysis of multiple rotavirus strains from a single stool specimen using sequence-independent amplification and 454(R) pyrosequencing reveals evidence of intergenotype genome segment recombination. *Infect. Genet. Evol.* 11, 2072–2082.

Ji, H., Li, Y., Graham, M., Liang, B. B., Pilon, R., Tyson, S., Peters, G., Tyler, S., Merks, H., Bertagnolio, S., Soto-Ramirez, L., Sandstrom, P., and Brooks, J. (2011). Next-generation sequencing of dried blood spot specimens: a novel approach to HIV drug-resistance surveillance. *Antivir. Ther.* 16, 871–878.

Ji, H., Masse, N., Tyler, S., Liang, B., Li, Y., Merks, H., Graham, M., Sandstrom, P., and Brooks, J. (2010). HIV drug resistance surveillance using pooled pyrosequencing. *PLoS ONE* 5:e9263. doi: 10.1371/journal.pone.0009263

Jojic, V., Hertz, T., and Jojic, N. (2008). "Population sequencing using short reads: HIV as a case study," in *Pacific Symposium on Biocomputing*, eds R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein (World Scientific), 114–125. ISBN 978-981-277-608-2.

Jose, M., Gajardo, R., and Jorquera, J. I. (2005). Stability of HCV, HIV-1 and HBV nucleic acids in plasma samples under long-term storage. *Biologicals* 33, 9–16.

Judo, M. S., Wedel, A. B., and Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res.* 26, 1819–1825.

Jung, G. S., Kim, Y. Y., Kim, J. I., Ji, G. Y., Jeon, J. S., Yoon, H. W., Lee, G. C., Ahn, J. H., Lee, K. M., and Lee, C. H. (2011). Full genome sequencing and analysis of human cytomegalovirus strain JHC isolated from a Korean patient. *Virus Res.* 156, 113–120.

Kampmann, M. L., Fordyce, S. L., Avila-Arcos, M. C., Rasmussen, M., Willerslev, E., Nielsen, L. P., and Gilbert, M. T. (2011). A simple method for the parallel deep sequencing of full influenza A genomes. *J. Virol. Methods* 178, 243–248.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* 96, 317–323.

Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol.* 10, R83.

Knoepfel, S. A., Di Giallonardo, F., Daumer, M., Thielen, A., and Metzner, K. J. (2011). In-depth analysis of G-to-A hypermutation rate in HIV-1 env DNA induced by endogenous APOBEC3 proteins using massively parallel sequencing. *J. Virol. Methods* 171, 329–338.

Ko, S.-Y., Oh, H.-B., Park, C.-W., Lee, H. C., and Lee, J.-E. (2012). Analysis of hepatitis B virus drug-resistant mutant haplotypes by ultra-deep pyrosequencing. *Clin. Microbiol. Infect*. doi: 10.1111/j.1469-0691.2012.03951.x. [Epub ahead of print].

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2, Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576.

Kozal, M. J., Chiarella, J., St. John, E. P., Moreno, E. A., Simen, B. B., Arnold, T. E., and Lataillade, M. (2011). Prevalence of low-level HIV-1 variants with reverse transcriptase mutation K65R and the effect of antiretroviral drug exposure on variant levels. *Antivir. Ther.* 16, 925–929.

Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., Sekizuka, T., Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y., Hata, S., Watanabe, M., and Sata, T. (2010). Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer. *PLoS ONE* 5:e10256. doi: 10.1371/journal.pone.0010256

Kwok, H., Tong, A. H. Y., Lin, C. H., Lok, S., Farrell, P. J., Kwong, D. L. W., and Chiang, A. K. S. (2012). Genomic sequencing and comparative analysis of Epstein-Barr virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PLoS ONE* 7:e36939. doi: 10.1371/journal.pone.0036939

Lataillade, M., Chiarella, J., Yang, R., Degrosky, M., Uy, J., Seekins, D., Simen, B., John, E. S., Moreno, E., and Kozal, M. (2012). Virologic failures on initial boosted-PI regimen infrequently possess low-level variants with major PI resistance mutations by ultra-deep sequencing. *PLoS ONE* 7:e30118. doi: 10.1371/journal.pone.0030118

Lataillade, M., Chiarella, J., Yang, R., Schnittman, S., Wirtz, V., Uy, J., Seekins, D., Krystal, M., Mancini, M., McGrath, D., Simen, B., Egholm, M., and Kozal, M. (2010). Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naive subjects in the CASTLE study. *PLoS ONE* 5:e10952. doi: 10.1371/journal.pone.0010952

Lauck, M., Alvarado-Mora, M. V., Becker, E. A., Bhattacharya, D., Striker, R., Hughes, A. L., Carrilho, F. J., O'connor, D. H., and Pinho, J. R. (2012). Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. *J. Virol.* 86, 3952–3960.

Le, T., Chiarella, J., Simen, B. B., Hanczaruk, B., Egholm, M., Landry, M. L., Dieckhaus, K., Rosen, M. I., and Kozal, M. J. (2009). Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS ONE* 4:e6079. doi: 10.1371/journal.pone.0006079

Li, J. Z., Paredes, R., Ribaudo, H. J., Svarovskaia, E. S., Metzner, K. J., Kozal, M. J., Hullsiek, K. H., Balduin, M., Jakobsen, M. R., Geretti, A. M., Thiebaut, R., Ostergaard, L., Masquelier, B., Johnson, J. A., Miller, M. D., and Kuritzkes, D. R. (2011). Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA* 305, 1327–1335.

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74, 363–377.

Liu, P., Fang, X., Feng, Z., Guo, Y.-M., Peng, R.-J., Liu, T., Huang, Z., Feng, Y., Sun, X., Xiong, Z., Guo, X., Pang, S.-S., Wang, B., Lv, X., Feng, F.-T., Li, D.-J., Chen, L.-Z., Feng, Q.-S., Huang, W.-L., Zeng, M.-S., Bei, J.-X., Zhang, Y., and Zeng, Y.-X. (2011). Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *J. Virol.* 85, 11291–11299.

Liu, S. L., Rodrigo, A. G., Shankarappa, R., Learn, G. H., Hsu, L., Davidov, O., Zhao, L. P., and Mullins, J. I. (1996). HIV quasispecies and resampling. *Science* 273, 415–416.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439.

Lorusso, A., Vincent, A. L., Harland, M. L., Alt, D., Bayles, D. O., Swenson, S. L., Gramer, M. R., Russell, C. A., Smith, D. J., Lager, K. M., and Lewis, N. S. (2011). Genetic and antigenic characterization of H1 influenza viruses from United States swine from 2008. *J. Gen. Virol.* 92, 919–930.

Macalalad, A. R., Zody, M. C., Charlebois, P., Lennon, N. J., Newman, R. M., Malboeuf, C. M., Ryan, E. M., Boutwell, C. L., Power, K. A., Brackney, D. E., Pesko, K. N., Levin, J. Z., Ebel, G. D., Allen, T. M., Birren, B. W., and Henn, M. R. (2012). Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput. Biol.* 8:e1002417. doi: 10.1371/journal.pcbi.1002417

Mancuso, N., Tork, B., Mandoiu, I. I., Skums, P., and Zelikovsky, A. (2011). "Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads," in *Proceedings of the 1st Workshop on Computational Advances in Molecular Epidemiology,* (IEEE), 94–101. ISBN: 978-1-4577-1612-6.

Mardis, E. R. (2008a). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141.

Mardis, E. R. (2008b). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402.

Margeridon-Thermet, S., Shulman, N. S., Ahmed, A., Shahriar, R., Liu, T.,

Wang, C., Holmes, S. P., Babrzadeh, F., Gharizadeh, B., Hanczaruk, B., Simen, B. B., Egholm, M., and Shafer, R. W. (2009). Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients. *J. Infect. Dis.* 199, 1275–1285.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

Messiaen, P., Verhofstede, C., Vandenbroucke, I., Dinakis, S., Van Eygen, V., Thys, K., Winters, B., Aerssens, J., Vogelaers, D., Stuyver, L. J., and Vandekerckhove, L. (2012). Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI-based regimen in treatment-naive patients. *Virology* 426, 7–11.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46.

Metzner, K. J., Bonhoeffer, S., Fischer, M., Karanicolas, R., Allers, K., Joos, B., Weber, R., Hirschel, B., Kostrikis, L. G., Günthard, H. F., and Study, T. S. H. C. (2003). Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. *J. Infect. Dis.* 188, 1433–1443.

Metzner, K. J., Giulieri, S. G., Knoepfel, S. A., Rauch, P., Burgisser, P., Yerly, S., Gunthard, H. F., and Cavassini, M. (2009). Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin. Infect. Dis.* 48, 239–247.

Meyerhans, A., Vartanian, J. P., and Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Res.* 18, 1687–1691.

Mild, M., Hedskog, C., Jernberg, J., and Albert, J. (2011). Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. *PLoS ONE* 6:e22741. doi: 10.1371/journal.pone.0022741

Mitsuya, Y., Varghese, V., Wang, C., Liu, T. F., Holmes, S. P., Jayakumar, P., Gharizadeh, B., Ronaghi, M., Klein, D., Fessel, W. J., and Shafer, R. W. (2008). Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J. Virol.* 82, 10747–10755.

Moorthy, A., Kuhn, L., Coovadia, A., Meyers, T., Strehlau, R., Sherman, G., Tsai, W. Y., Chen, Y. H., Abrams, E. J., and Persaud, D. (2011). Induction therapy with protease-inhibitors modifies the effect of nevirapine resistance on virologic response to nevirapine-based HAART in children. *Clin. Infect. Dis.* 52, 514–521.

Mukherjee, R., Jensen, S. T., Male, F., Bittinger, K., Hodinka, R. L., Miller, M. D., and Bushman, F. D. (2011). Switching between raltegravir resistance pathways analyzed by deep sequencing. *AIDS* 25, 1951–1959.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Ul-Amin, M. A., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39, e90.

Nasu, A., Marusawa, H., Ueda, Y., Nishijima, N., Takahashi, K., Osaki, Y., Yamashita, Y., Inokuma, T., Tamada, T., Fujiwara, T., Sato, F., Shimizu, K., and Chiba, T. (2011). Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. *PLoS ONE* 6:e24907. doi: 10.1371/journal.pone.0024907

Ninomiya, M., Ueno, Y., Funayama, R., Nagashima, T., Nishida, Y., Kondo, Y., Inoue, J., Kakazu, E., Kimura, O., Nakayama, K., and Shimosegawa, T. (2012). Use of illumina deep sequencing technology to differentiate hepatitis C virus variants. *J. Clin. Microbiol.* 50, 857–866.

Nishijima, N., Marusawa, H., Ueda, Y., Takahashi, K., Nasu, A., Osaki, Y., Kou, T., Yazumi, S., Fujiwara, T., Tsuchiya, S., Shimizu, K., Uemoto, S., and Chiba, T. (2012). Dynamics of hepatitis B virus quasispecies in association with nucleos(t)ide analogue treatment determined by ultra-deep sequencing. *PLoS ONE* 7:e35052. doi: 10.1371/journal.pone.0035052

Nowak, M. A. (1992). What is a quasispecies? *Trends Ecol. Evol.* 7, 118–121.

O'Neil, S. T., and Emrich, S. J. (2012). Haplotype and minimum-chimerism consesus determination using short sequence data. *BMC Genomics* 13, S4.

Ojosnegros, S., Beerenwinkel, N., Antal, T., Nowak, M. A., Escarmís, C., and Domingo, E. (2010). Competition-colonization dynamics in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2108–2112.

Poon, A. F., McGovern, R. A., Mo, T., Knapp, D. J., Brenner, B., Routy, J. P., Wainberg, M. A., and Harrigan, P. R. (2011). Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. *AIDS* 25, 2019–2026.

Poon, A. F., Swenson, L. C., Dong, W. W., Deng, W., Kosakovsky Pond, S. L., Brumme, Z. L., Mullins, J. I., Richman, D. D., Harrigan, P. R., and Frost, S. D. (2010). Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. *Mol. Biol. Evol.* 27, 819–832.

Pop, M., and Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet.* 24, 142–149.

Powdrill, M. H., Tchesnokov, E. P., Kozak, R. A., Russell, R. S., Martin, R., Svarovskaia, E. S., Mo, H., Kouyos, R. D., and Gotte, M. (2011). Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20509–20513.

Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., and Roth, V. (2010). "HIV haplotype inference using a constraint-based Dirichlet process mixture model," in *NIPS Workshop on Machine Learning in Computational Biology*.

Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science* 242, 1168–1171.

Prosperi, M. C. F., Prosperi, L., Bruselles, A., Abbate, I., Rozera, G., Vincenti, D., Solmone, M. C., Capobianchi, M. R., and Ulivi, G. (2011). Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics* 12, 5.

Prosperi, M. C. F., and Salemi, M. (2012). QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics* 28, 132–133.

Pybus, O. G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10, 540–550.

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F., and Sloan, W. T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6, 639–641.

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12, 38.

Ramakrishnan, M. A., Tu, Z. J., Singh, S., Chockalingam, A. K., Gramer, M. R., Wang, P., Goyal, S. M., Yang, M., Halvorson, D. A., and Sreevatsan, S. (2009). The feasibility of using high resolution genome sequencing of influenza A viruses to detect mixed infections and quasispecies. *PLoS ONE* 4:e7105. doi: 10.1371/journal.pone.0007105

Rasmussen, C. E. (2000). "The infinite gaussian mixture model," in *NIPS*, eds S. A. Solla, T. K. Leen, and K.-R. Müller (The MIT Press), 554–560.

Raymond, S., Saliou, A., Nicot, F., Delobel, P., Dubois, M., Cazabat, M., Sandres-Saune, K., Marchou, B., Massip, P., and Izopet, J. (2011). Frequency of CXCR4-using viruses in primary HIV-1 infections using ultra-deep pyrosequencing. *AIDS* 25, 1668–1670.

Redd, A. D., Mullis, C. E., Serwadda, D., Kong, X., Martens, C., Ricklefs, S. M., Tobian, A. A., Xiao, C., Grabowski, M. K., Nalugoda, F., Kigozi, G., Laeyendecker, O., Kagaayi, J., Sewankambo, N., Gray, R. H., Porcella, S. F., Wawer, M. J., and Quinn, T. C. (2012). The rates of HIV superinfection and primary HIV incidence in a general population in Rakai, Uganda. *J. Infect. Dis.* 206, 267–274.

Reuman, E. C., Margeridon-Thermet, S., Caudill, H. B., Liu, T., Borroto-Esoda, K., Svarovskaia, E. S., Holmes, S. P., and Shafer, R. W. (2010). A classification model for G-to-A hypermutation in hepatitis B virus ultra-deep pyrosequencing reads. *Bioinformatics* 26, 2929–2932.

Reumers, J., Rijk, P. D., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Loo, P. V., Bossche, M. V. D., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D., and Del-Favero,

J. (2011). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* 30, 61–68.

Reyes, G. R., and Kim, J. P. (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Mol. Cell. Probes* 5, 473–481.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science* 242, 1171–1173.

Rodriguez-Frías, F., Tabernero, D., Quer, J., Esteban, J. I., Ortega, I., Domingo, E., Cubero, M., Camós, S., Ferrer-Costa, C., Sánchez, A., Jardí, R., Schaper, M., Homs, M., Garcia-Cehic, D., Guardia, J., Esteban, R., and Buti, M. (2012). Ultra-deep pyrosequencing detects conserved genomic sites and quantifies linkage of drug-resistant amino acid changes in the hepatitis B virus genome. *PLoS ONE* 7:e37874. doi: 10.1371/journal.pone.0037874

Rozera, G., Abbate, I., Bruselles, A., Vlassi, C., D'offizi, G., Narciso, P., Chillemi, G., Prosperi, M., Ippolito, G., and Capobianchi, M. R. (2009). Massively parallel pyrosequencing highlights minority variants in the HIV-1 *env* quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* 6, 15.

Saeed, F., Khokhar, A., Zagordi, O., and Beerenwinkel, N. (2009). "Multiple sequence alignment system for pyrosequencing reads," in *BICoB 2009, LNBI 5462*, ed S. Rajasekaran (Berlin Heidelberg: Springer-Verlag), 362–375.

Saliou, A., Delobel, P., Dubois, M., Nicot, F., Raymond, S., Calvez, V., Masquelier, B., and Izopet, J. (2011). Concordance between two phenotypic assays and ultradeep pyrosequencing for determining HIV-1 tropism. *Antimicrob. Agents Chemother.* 55, 2831–2836.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* 5, 16–18.

Sede, M., Ojeda, D., Cassino, L., Westergaard, G., Vazquez, M., Benetti, S., Fay, F., Tanno, H., and Quarleri, J. (2012). Long-term monitoring drug resistance by ultra-deep pyrosequencing in a chronic hepatitis B virus (HBV)-infected patient exposed to several unsuccessful therapy schemes. *Antiviral Res*. 94, 184–187.

Simen, B. B., Simons, J. F., Hullsiek, K. H., Novak, R. M., MacArthur, R. D., Baxter, J. D., Huang, C., Lubeski, C., Turenchalk, G. S., Braverman, M. S., Desany, B., Rothberg, J. M., Egholm, M., and Kozal, M. J. (2009). Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. *J. Infect. Dis.* 199, 693–701.

Skums, P., Dimitrova, Z., Campo, D. S., Vaughan, G., Rossi, L., Forbi, J. C., Yokosawa, J., Zelikovsky, A., and Khudyakov, Y. (2012). Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics* 13(Suppl. 10), S6.

Solmone, M., Vincenti, D., Prosperi, M. C., Bruselles, A., Ippolito, G., and Capobianchi, M. R. (2009). Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen. *J. Virol.* 83, 1718–1726.

Stelzl, E., Proll, J., Bizon, B., Niklas, N., Danzer, M., Hackl, C., Stabentheiner, S., Gabriel, C., and Kessler, H. H. (2011). Human immunodeficiency virus type 1 drug resistance testing: evaluation of a new ultra-deep sequencing-based protocol and comparison with the TRUGENE HIV-1 genotyping kit. *J. Virol. Methods* 178, 94–97.

Svicher, V., Balestra, E., Cento, V., Sarmati, L., Dori, L., Vandenbroucke, I., D'Arrigo, R., Buonomini, A. R., Marck, H. V., Surdo, M., Saccomandi, P., Mostmans, W., Aerssens, J., Aquaro, S., Stuyver, L. J., Andreoni, M., Ceccherini-Silberstein, F., and Perno, C. F. (2011). HIV-1 dual/mixed tropic isolates show different genetic and phenotypic characteristics and response to maraviroc *in vitro*. *Antiviral Res.* 90, 42–53.

Swenson, L. C., Mo, T., Dong, W. W., Zhong, X., Woods, C. K., Jensen, M. A., Thielen, A., Chapman, D., Lewis, M., James, I., Heera, J., Valdez, H., and Harrigan, P. R. (2011a). Deep sequencing to infer HIV-1 co-receptor usage: application to three clinical trials of maraviroc in treatment-experienced patients. *J. Infect. Dis.* 203, 237–245.

Swenson, L. C., Mo, T., Dong, W. W. Y., Zhong, X., Woods, C. K., Thielen, A., Jensen, M. A., Knapp, D. J. H. F., Chapman, D., Portsmouth, S., Lewis, M., James, I., Heera, J., Valdez, H., and Harrigan, P. R. (2011b). Deep V3 sequencing for HIV type 1 tropism in treatment-naive patients: a reanalysis of the MERIT trial of maraviroc. *Clin. Infect. Dis.* 53, 732–742.

Swenson, L. C., Moores, A., Low, A. J., Thielen, A., Dong, W., Woods, C., Jensen, M. A., Wynhoven, B., Chan, D., Glascock, C., and Harrigan, P. R. (2010). Improved detection of CXCR4-using HIV by V3 genotyping: application of population-based and "deep" sequencing to plasma RNA and proviral DNA. *J. Acquir. Immune Defic. Syndr.* 54, 506–510.

Tapparel, C., Cordey, S., Junier, T., Farinelli, L., Van Belle, S., Soccal, P. M., Aubert, J. D., Zdobnov, E., and Kaiser, L. (2011). Rhinovirus genome variation during chronic upper and lower respiratory tract infections. *PLoS ONE* 6:e21163. doi: 10.1371/journal.pone.0021163

Trapnell, C., and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nat. Biotechnol.* 27, 455–457.

Turner, E. H., Ng, S. B., Nickerson, D. A., and Shendure, J. (2009). Methods for genomic partitioning. *Annu. Rev. Genomics Hum. Genet.* 10, 263–284.

Van Nimwegen, E., Crutchfield, J. P., and Huynen, M. (1999). Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9716–9720.

Vandekerckhove, L., Verhofstede, C., Demecheleer, E., De Wit, S., Florence, E., Fransen, K., Moutschen, M., Mostmans, W., Kabeya, K., Mackie, N., Plum, J., Vaira, D., Van Baelen, K., Vandenbroucke, I., Van Eygen, V., Van Marck, H., Vogelaers, D., Geretti, A. M., and Stuyver, L. J. (2011). Comparison of phenotypic and genotypic tropism determination in triple-class-experienced HIV patients eligible for maraviroc treatment. *J. Antimicrob. Chemother.* 66, 265–272.

Vandenbroucke, I., Van Marck, H., Mostmans, W., Van Eygen, V., Rondelez, E., Thys, K., Van Baelen, K., Fransen, K., Vaira, D., Kabeya, K., De Wit, S., Florence, E., Moutschen, M., Vandekerckhove, L., Verhofstede, C., and Stuyver, L. J. (2010). HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. *AIDS Res. Ther.* 7, 4.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M.-L., Teague, J., Bignell, G., Butler, A., Cho, J., Dalgliesh, G. L., Galappaththige, D., Greenman, C., Hardy, C., Jia, M., Latimer, C., Lau, K. W., Marshall, J., Mclaren, S., Menzies, A., Mudie, L., Stebbings, L., Largaespada, D. A., Wessels, L. F. A., Richard, S., Kahnoski, R. J., Anema, J., Tuveson, D. A., Perez-Mancera, P. A., Mustonen, V., Fischer, A., Adams, D. J., Rust, A., On, W. C., Subimerb, C., Dykema, K., Furge, K., Campbell, P. J., Teh, B. T., Stratton, M. R., and Futreal, P. A. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 469, 539–542.

Varghese, V., Shahriar, R., Rhee, S. Y., Liu, T., Simen, B. B., Egholm, M., Hanczaruk, B., Blake, L. A., Gharizadeh, B., Babrzadeh, F., Bachmann, M. H., Fessel, W. J., and Shafer, R. W. (2009). Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J. Acquir. Immune Defic. Syndr.* 52, 309–315.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., and Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439, 344–348.

Vrancken, B., Lequime, S., Theys, K., and Lemey, P. (2010). Covering all bases in HIV research: unveiling a hidden world of viral evolution. *AIDS Rev.* 12, 89–102.

Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., and Shafer, R. W. (2007). Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.* 17, 1195–1201.

Wang, G. P., Sherrill-Mix, S. A., Chang, K. M., Quince, C., and Bushman, F. D. (2010). Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *J. Virol.* 84, 6218–6228.

Westbrooks, K., Astrovskaya, I., Campo, D., Khudyakov, Y., Berman, P., and Zelikovsky, A. (2008). "HCV quasispecies assembly using network flows," in *ISBRA 2008, LNBI 4983*, eds I. Măndoiu, R. Sunderraman, and A. Zelikovsky (Berlin Heidelberg: Springer-Verlag), 159–170.

WHO. (2012). *World Health Organization* [Online]. Available online at: www.who.int [Accessed 1 May 2012].

Wikipedia (2012). List of sequence alignment software [Online]. Available: http://en.wikipedia.org/wiki/List_of_sequence_alignment_software # Short - Read _ Sequence _ alignment [Accessed 1 May 2012].

Wilke, C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evol. Biol.* 5, 44.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412, 331–333.

Willerth, S. M., Pedro, H. A., Pachter, L., Humeau, L. M., Arkin, A. P., and Schaffer, D. V. (2010). Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS ONE* 5:e13564. doi: 10.1371/journal.pone.0013564

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N. S., Louder, M., McKee, K., O'Dell, S., Perfetto, S., Schmidt, S. D., Shi, W., Wu, L., Yang, Y., Yang, Z. Y., Yang, Z., Zhang, Z., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Doria-Rose, N. A., Connors, M., Mullikin, J. C., Nabel, G. J., Roederer, M., Shapiro, L., Kwong, P. D., and Mascola, J. R. (2011). Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333, 1593–1602.

Yang, X., Chockalingam, S. P., and Aluru, S. (2012). A survey of error-correction methods for next-generation sequencing. *Brief Bioinform*. doi: 10.1093/bib/bbs015. [Epub ahead of print]

Zagordi, O., Bhattacharya, A., Eriksson, N., and Beerenwinkel, N. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12, 119.

Zagordi, O., Geyrhofer, L., Roth, V., and Beerenwinkel, N. (2010a). Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol.* 17, 417–428.

Zagordi, O., Klein, R., Däumer, M., and Beerenwinkel, N. (2010b). Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.* 38, 7400–7409.

Zagordi, O., Töpfer, A., Prabhakaran, S., Roth, V., Halperin, E., and Beerenwinkel, N. (2012). "Probabilistic inference of viral quasispecies subject to recombination," in *RECOMB 2012, LNBI 7262*, ed B. Chor (Berlin Heidelberg: Springer-Verlag), 342–354.

Zell, R., Taudien, S., Pfaff, F., Wutzler, P., Platzer, M., and Sauerbrei, A. (2012). Sequencing of 21 varicella-zoster virus genomes reveals two novel genotypes and evidence of recombination. *J. Virol.* 86, 1608–1622.

Zhao, X., Palmer, L. E., Bolanos, R., Mircean, C., Fasulo, D., and Wittenberg, G. M. (2010). EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol.* 17, 1549–1560.

Conflict of Interest Statement: Karin J. Metzner has received travel grants and honoraria from Gilead, Roche Diagnostics, GlaxoSmithKline, Bristol-Myers Squibb, Tibotec, and Abbott, and has received research grants from Abbott, Gilead, and Roche Diagnostics. Huldrych F. Günthard has been an adviser and/or consultant for the following companies: GlaxoSmithKline, Abbott, Novartis, Gilead, Boehringer Ingelheim, Roche, Tibotec and Bristol-Myers Squibb, and has received unrestricted research and educational grants from Roche, Abbott, Bristol-Myers Squibb, GlaxoSmithKline, Gilead, Tibotec and Merck Sharp & Dohme (all money went to institution). The other 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.

Received: 15 June 2012; paper pending published: 06 July 2012; accepted: 24 August 2012; published online: 11 September 2012.

*Citation: Beerenwinkel N, Günthard HF, Roth V and Metzner KJ (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbio. **3**:329. doi: 10.3389/fmicb.2012.00329*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 24 August 2012
doi: 10.3389/fmicb.2012.00306

MicroRNAs in HIV-1 infection: an integration of viral and cellular interaction at the genomic level

Neil H. Tan Gana, Tomohiro Onuki, Ann Florence B. Victoriano and Takashi Okamoto*

Department of Molecular and Cell Biology, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Akio Kanai, Keio University, Japan
Akihide Ryo, Yokohama City University, Japan

***Correspondence:**
Takashi Okamoto, Department of Molecular and Cell Biology, Nagoya City University Graduate School of Medical Sciences, 1-Kawasumi, Mizuho-cho, Mizuho-ku, Nagoya 467-8601, Japan.
e-mail: tokamoto@med.nagoya-cu.ac.jp

The microRNA pathways govern complex interactions of the host and virus at the transcripts level that regulate cellular responses, viral replication and viral pathogenesis. As a group of single-stranded short non-coding ribonucleotides (ncRNAs), the microRNAs complement their messenger RNA (mRNA) targets to effect post-transcriptional or translational gene silencing. Previous studies showed the ability of human immunodeficiency virus 1 (HIV-1) to encode microRNAs which modify cellular defence mechanisms thus creating an environment favorable for viral invasion and replication. In corollary, cellular microRNAs were linked to the alteration of HIV-1 infection at different stages of replication and latency. As evidences further establish the regulatory involvement of both cellular and viral microRNA in HIV-1-host interactions, there is a necessity to organize this information. This paper would present current and emerging knowledge on these multi-dimensional interactions that may facilitate the design of microRNAs as effective antiretroviral reagents.

Keywords: microRNA, HIV-1 mechanisms, transcription factors, targets

INTRODUCTION

The human immunodeficiency virus 1 (HIV-1) is the retroviral agent causing acquired immunodeficiency syndrome (AIDS), a disease leading to systemic failure of the immune system with life threatening consequences. The decades old magnanimous problem of HIV-1 infection has challenged researchers to address its control and eradication. One of the most recent strategies introduced is the use of small non-coding ribonucleotides (ncRNAs) which includes microRNAs (Arbuthnot, 2011). The microRNAs are ubiquitous ∼22–25 nt endogenously expressed ncRNAs targeting specific messenger RNA (mRNA) sequences, thus inducing its degradation or effecting translational inhibition. As proven vital regulatory components of viral infection and immunity (Huang et al., 2011), microRNAs can be directed to target viral and cellular transcripts to suppress infection. In fact, numerous studies have been proposed to integrate cellular microRNAs as nucleotide-based therapy for HIV (Boden et al., 2004; Lo et al., 2007; Aagaard et al., 2008). However, HIV-1 is a fastidious mutant consequently making cellular microRNAs prone to losing its viral transcript target efficacy as constant genome revisions occur in the course of viral evolution. Thus, simultaneous expression of microRNAs aimed to repress multiple HIV-1 targets may deter the effects of escape mutants. In another scenario, the cellular microRNAs can target host gene products that regulate cell defense responses. Once the issues of microRNA off-target effects, cell toxicity and delivery systems are addressed, the development of microRNAs as an anti-HIV-1 therapeutic strategy becomes more realistic (Boden et al., 2007; Liu et al., 2011). Now, the greater challenge is to determine the specific roles of the current inventory of 1921 human and three HIV-1 microRNAs (Kozomara and Griffiths-Jones, 2011) in HIV-1 infection. This difficult task of functional assignments correlated to microRNA-mRNA interactions has been made easier with genomics-based predictive tools in the recent years (Tan Gana et al., 2012). In addition, significant improvements on techniques for microRNA discovery and functional elucidation are likely to further expand these emerging interactive networks.

Whereas the current knowledge on cellular and viral microRNA functions involved in HIV-1 infection is still considerably few, consensus evidences suggest complex interactions (Chiang and Rice, 2011; Sanchez-Del Cojo et al., 2011). **Figure 1** implies that microRNA regulation is anchored on genomic information processing on four scenarios that may possibly explain the confounded nature of their effects in virus-infected host systems. First, HIV-1 infection alters host microRNA networks to initiate successful viral invasion and latency, thus, affecting global host microRNA regulome. Second, HIV-1 microRNAs are produced from both sense and antisense transcripts to target either its own viral transcript or host genes for immune compromise. Third, the host microRNA systems may consequentially target the HIV-1 genomic elements or its genes to innate immune responses. Fourth, the interplay of microRNA and target mRNA between host and HIV-1 can be organized into regulatory modules (*cis*- and *trans*- regulation) of essential biochemical pathways as critical determinants of host cell fate and survival. This framework would be the basis of our paper discussion covering an update on the current information on microRNA biogenesis and mechanisms involved in host-virus interactions. Also the paper would contain recently elucidated cellular and viral microRNA functions in HIV-1 infection from computational and experimental literature. Lastly, the integration of information would define future roles of microRNAs in HIV-1 control.

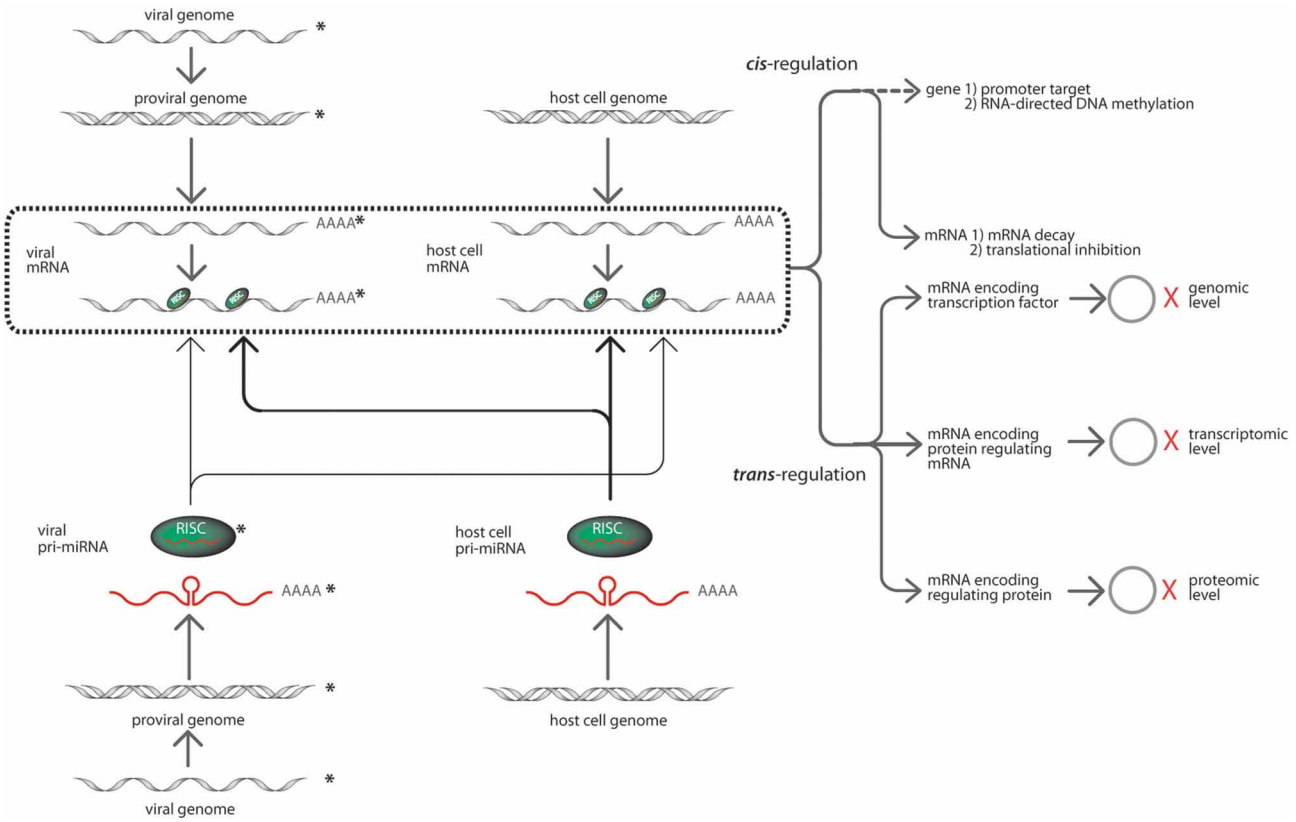


FIGURE 1 | The prospective targeting interactions of HIV-1 and cellular microRNAs (modified from Cullen, 2006). The interactions among HIV-1 and cellular microRNAs with their corresponding targets may occur in several modules. For instance, HIV-1 encoded microRNAs processed via the host RNAi machinery and incorporated into RISC (in green with the mature microRNA in red) are sourced from HIV-1 pro-viral strands (grey double helix) initially from precursor microRNAs (red lines with poly adenines). These HIV-1 microRNAs can target its viral transcripts or the cellular transcripts. The targeting interactions of microRNAs are shown in solid light arrow lines. In corollary, cellular microRNAs derived from precursor microRNAs (red lines with poly adenines) generated by the host cell genome (grey double helix). The host cellular microRNAs are encoded in the same manner and can target both viral and cellular transcripts where the targeting interactions are shown in solid bold arrow lines. The targeting of mRNA transcripts happens in a highly specific Watson and Crick base-pairing with either complete complementation or seed region complementarity. The box in bold broken lines consolidates all targeting events of the various microRNA-initiated regulatory activities within the systems biology of host-virus interaction. The type of microRNA silencing mechanisms may be grouped as a *cis*- and *trans*-regulation event. The *cis*-regulation event involves microRNA targeting of mRNAs initiating post-transcriptional regulatory responses via mRNA degradation and translational inhibition. Whereas *trans*-regulation is a tripartite regulatory event which include expression variation of microRNA target genes regulating various viral and cellular activities such as transcription factors, RNA regulatory proteins, interactive genes. The cascades of events cause changes in viral and cellular activities inducing transcriptional regulation, transcriptional variation and protein translational modifications as indicated by the hollow circles = protein products; X (in red) = regulation of expression. The HIV-1 components are distinguished from host cell components with asterisks beside the drawings.

MicroRNA BIOGENESIS PATHWAYS AND MECHANISMS

The genomic locations of the microRNA gene progenitors of the ∼100 nt, 5′ methyl-7G capped and 3′ poly-adenylated primary-microRNAs (pri-microRNA) transcribed by either RNA II or III polymerase determine the mode of microRNA biogenesis (**Figure 2**). The canonical pathway utilizes the microprocessor complex, an interaction between Drosha RNase III enzyme (Faller and Guo, 2008) and DiGeorge critical region gene 8 (DGCR8) (Faller et al., 2010), a ribonuclease binding protein (RBP) to cleave the pri-microRNA into 70 nt preliminary-microRNAs (pre-microRNAs). While, a non-canonical pathway is followed by mirtrons, a group of intron-derived pre-microRNAs utilizing spliceosomes (Okamura et al., 2007, 2008; Berezikov et al., 2010). Recently, an emerging mode of biogenic pathway has been proposed for a set of splicing-independent mirtrons called simtrons which neither utilize DGCR8, Dicer, Exportin-5, or Argonaute 2 (Ago2) in their biogenesis (Havens et al., 2012).

Then, the pre-microRNA associates with Ran/GTPase (Bohnsack et al., 2004; Okada et al., 2009) and exportins for cytoplasmic translocation from nucleus (Bohnsack et al., 2004). In the cytoplasm (**Figure 3**), catalytic hydrolysis of Ran/GTPase allows the dissociation of pre-microRNA and transporter proteins (Kim, 2004). Another enzyme called Dicer, splices the pre-microRNA into the mature ∼22 nt microRNA capable of mRNA duplexing (Carmell and Hannon, 2004; Cullen, 2004; Hammond, 2005; Harvey et al., 2008; Flores-Jasso et al., 2009). A complement strand from the mature double stranded microRNA is integrated into the RNA-induced silencing complex (RISC) which would be

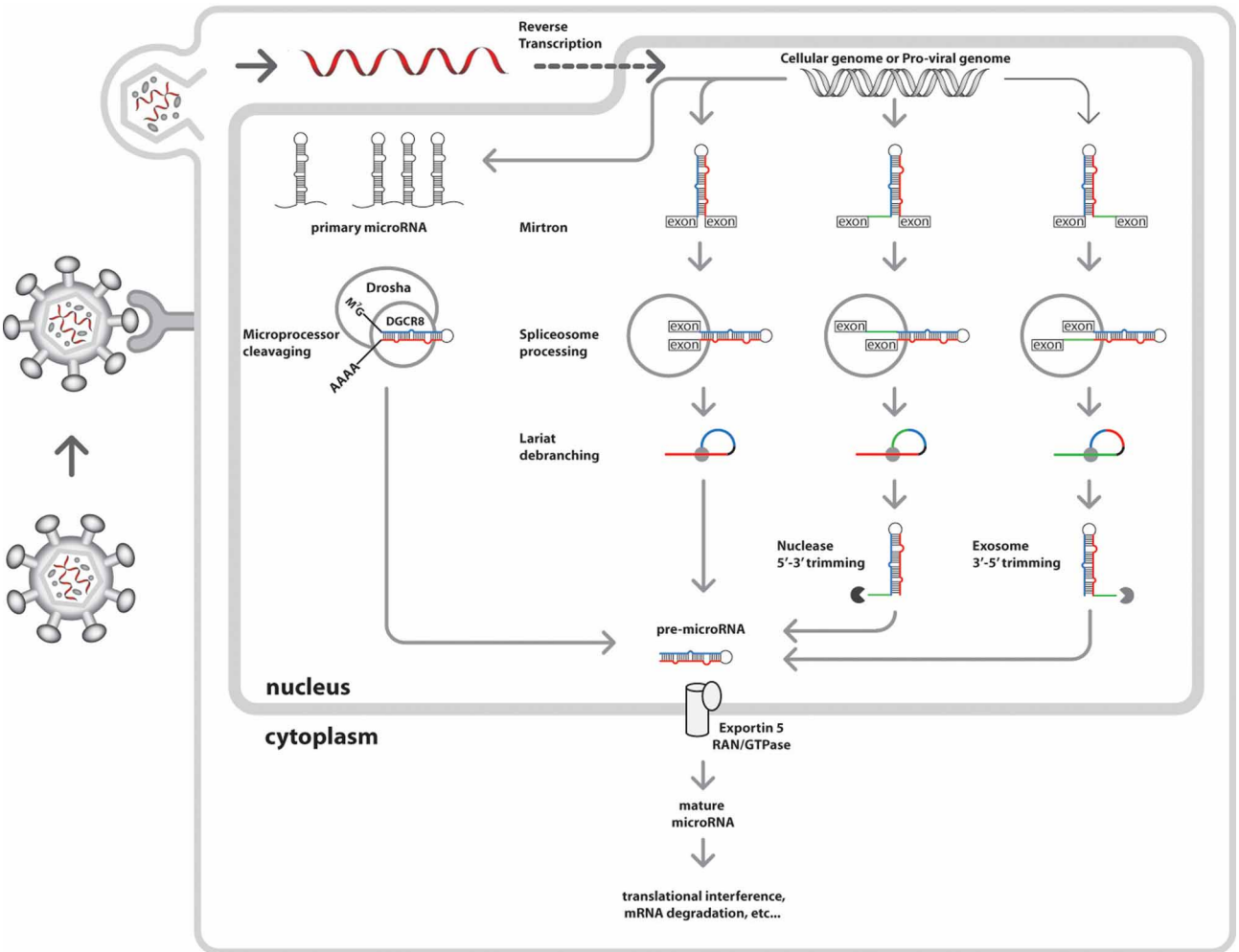


FIGURE 2 | Nuclear events of the integrated cellular and HIV-1 microRNA biogenic pathways. The nucleus of the host cell is the central site of both cellular and viral microRNA biogenesis. Initially, HIV-1 virion particles attach to host cells via CD4 receptors signaling viral attack. This would be followed by HIV-1 particle fusion with the cell membrane and uncoating to load its RNA genome into the cytoplasm. The viral replicase enzyme facilitates production of more RNA genome later to be shuttled into the nucleus for viral transcription and integration into cellular genome. HIV-1 microRNA biogenesis is synchronous to cellular microRNA production in the host cell nucleus where other microRNA biogenic enzymes are present. Independent nuclear events of miRNA biogenesis have several modes of pre-microRNA generation (viral and cellular), initially from primary microRNA transcribed by either RNA polymerase II or III. he canonical pathway is undertaken by intergenic microRNAs resulting from microprocessor cleavage (Drosha and DGCR8) of pri-microRNA transcripts into pre-microRNAs, An alternative pathway for intron-coded microRNAs called mirtrons produce pre-microRNAs via splicing by spliceosomes and debranching by lariat debranching enzyme (Ldbr). There are three possible variants of mirtron processing, namely regular, the 5′ tailed mirtrons (subject to nuclease processing) and 3′ tailed mirtons (subject to exosome processing) (Westholm and Lai, 2011; Westholm et al., 2012). The sections of these pre-microRNA variants are shown in different colors, which the future main mature microRNAs are in **blue**, the secondary mature microRNAs are in **red**, the loops in black, and the branches are in **green**. Once generated, the pre-miRNAs are ready cytoplasmic shuttling, where further processing into mature microRNAs are achieved. Lately, a new biogenic pathway has been proposed for a set of splicing-independent mirtrons called simtrons which independent from DGCR8, Dicer, Exportin-5, or Argonaute 2 (Ago2) (Havens et al., 2012) (*not shown*).

attached to the target transcript to elicit the regulatory processes (Kawamata and Tomari, 2010).

The gene regulatory effects caused by the microRNA and mRNA target interaction dictated by highly stringent base-complementation of the binding sites have been demonstrated extensively (Long et al., 2008; Brodersen and Voinnet, 2009; Ajay et al., 2010). Perfect complementation of the microRNA and mRNA causes endonucleolytic cleavage-induced gene silencing while non-perfect complementation initiate translational inhibition of proteins (Jackson and Standart, 2007; Seitz, 2009; Pan et al., 2010). Other factors shown to enhance these regulatory interactions include the presence of several binding sites within the target mRNA and CTG repeats extension (Hon and Zhang, 2007), deadenylation and decapping-mediated conformational changes within the microRNA-RISC complex (Lin et al., 2005; Eulalio et al., 2007), the A-U bias (Frank et al., 2010) and single nucleotide polymorphisms (SNPs) within the seed regions (Landi et al., 2011).

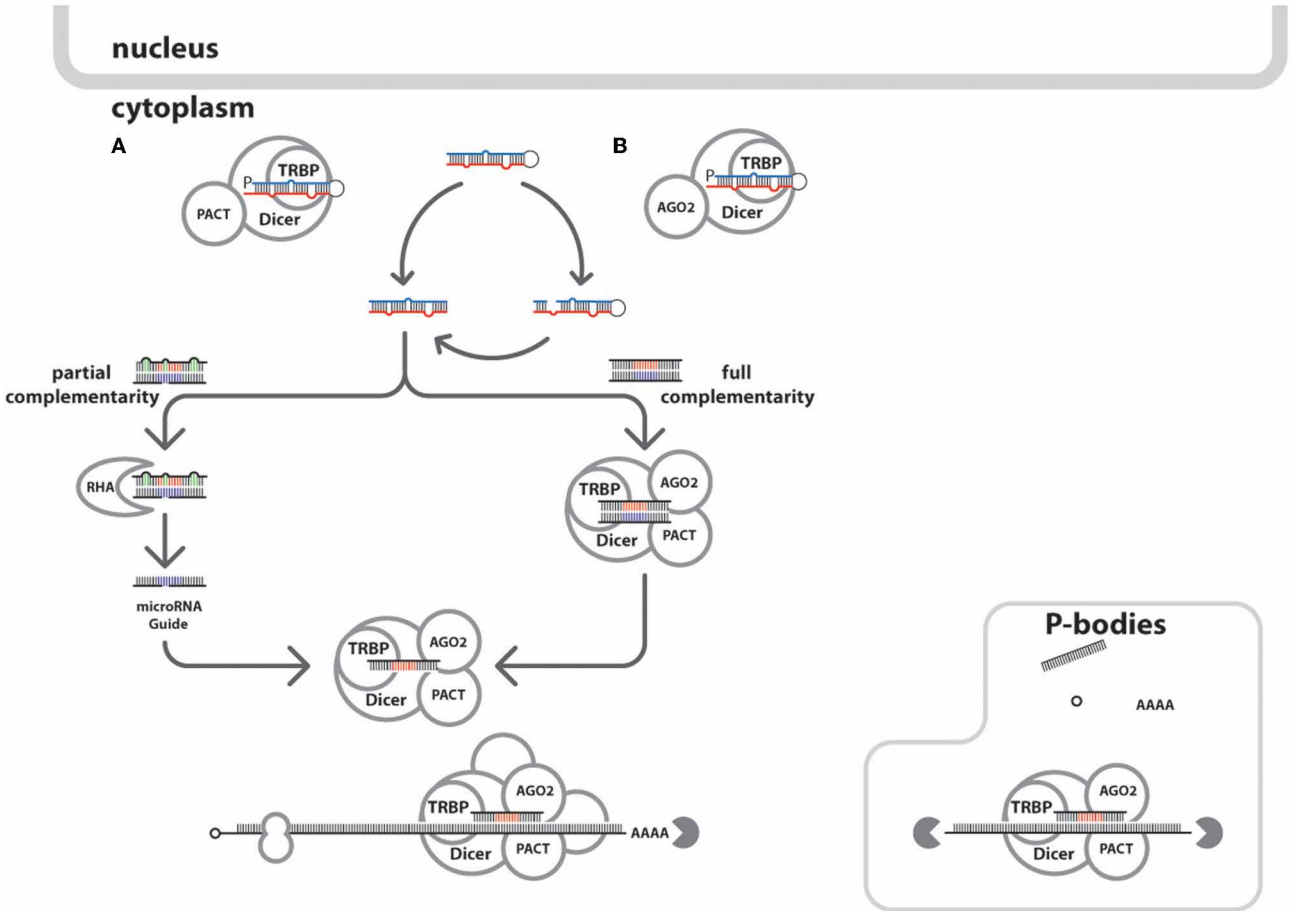


FIGURE 3 | Cytoplasmic events of the integrated cellular and HIV-1 microRNA biogenic pathways. Mature microRNAs are processed in the host cytoplasm after the pre-microRNAs are shuttled from the nucleus. There are two ways the processing can happen: **(A)** by the Dicer/TRBPorTARBP or TARBP/PACT leading to the generation of microRNA duplexes, or **(B)** via the Ago proteins-assisted generation of pre-cursor microRNA which in the end generate the microRNA duplexes via Dicer action. The circular arrows show the formation microRNA duplex produced after the circuitous cleavage reactions of either Dicer, Ago2 proteins. Also shown in the microRNAs duplexes are the representative seed sites in **red** and **blue** and **green** indicates mismatches and bulges. Depending on the degree of complementation of the guide strands in the microRNA:RISC assembly to its target mRNA, would either cause mRNA degradation or translational repression. Furthermore, microRNA duplexes undergoing asymmetric unwinding would be assembled into the RISC loading complex after the duplex unwinding by RNA helicase A (RHA, gray crescent), through a bypass mechanism. Once the microRNA guide is loaded into the RISC associated proteins (Dicer; TRBP; Ago2; PACT; GW182, not shown) forms the activated microRNA:RISC to seek and bind targets within mRNA transcript. The mRNA transcripts are shown with other components namely: methyl cap (in hollow small circle), ribosome (hollow figure eight), and poly-A tail (AAAA). In HIV-1, it was demonstrated that P-bodies modulate microRNA processing mechanisms (Nathans et al., 2009). The solid gray pac-man indicates mRNA degradation events.

Since, the viral genomic elements intersperse within the host genome during invasion, it is possible that these viral genomic fragments are processed into microRNAs by of the host microRNA machinery (**Figures 2**, **3**). As consequence, these viral sequences will follow several pathways of microRNA biogenesis similar to its host. As example are the observed non-significant differences among microRNA profiles in normalized Dicer and Drosha expressions in HIV-1 infected CD4+ cells for both *in vitro* and *in vivo* studies (Bignami et al., 2012). In contrast, it was confirmed that microRNA expression among monocytes could happen in the absence of the Dicer enzyme, thus, implying an alternative mode of viral microRNA production (Coley et al., 2010). Another possible consequence of repeated integration and recombination of the HIV-1 into the host genome is the generation of orthologous microRNAs. This is the case of hiv1-miR-N367 and hsa-miR-192 with identical seed sequences shown to down regulate similar functional targets in a dual fluorescent reporter study, thus, making them functional orthologs (You et al., 2012).

ALTERATIONS IN THE CELLULAR MicroRNA PATHWAYS DURING HIV-1 INFECTION

HIV-1 infection of host cells modifies the global RNA interference machinery which in effect changes the microRNA-regulated pathways via several bio-molecular interactions (Sanghvi and Steel, 2011b). Experiments confirmed the RNA silencing suppressor (RSS) activity of HIV-1 transactivator of transcription (Tat) (Bivalkar-Mehla et al., 2011). RSS is defined as a

molecule encoded by a virus which can counter the effect of host cell microRNA-mediated antiviral defense pathways, or natural immunity (Houzet and Jeang, 2011). The HIV-1 Tat protein via association with the trans-activation response (TAR) element at the terminal 5′end of HIV-1 transcripts, promotes viral transcription by recruiting and increasing the processivity of RNA polymerase II (Hayes et al., 2011). The molecular complex creates a stabilizing effect on transcriptional elongation elicited by a cyclin-dependent kinase (CDK9), another subunit of the positive transcription elongation factor b (P-TEFb) together with Cyclin T1 (CCNT1), which functions to phosphorylate the C-terminus of RNA polymerase II (Sanghvi and Steel, 2011a). Also the binding between Tat protein to the TAR element blocks the TAR element interaction with the Dicer protein thus influencing cellular silencing mechanism (Bennasser and Jeang, 2006; Qian et al., 2009). Further characterization of Tat protein exhibited two most essential prerequisites for RSS activity, namely by harboring dsRNA binding domain (RKKRRQRR) and GW/WG motif essential in sequestering Ago proteins thus preventing RISC formation (Qian et al., 2009; Bivalkar-Mehla et al., 2011; Houzet and Jeang, 2011). In another comparative gene expression profile analysis between HIV-1 infected and non-infected macrophages, it was showed that HIV-1-encoded Vpr (Viral Protein R) protein similarly suppresses Dicer function (Coley et al., 2010). **Table 1** includes the RNAi pathway related gene products that are targeted by HIV-1 microRNAs. With the inherent small size of

Table 1 | List of published HIV-1 microRNAs and their target HIV-1 and cellular gene products.

HIV-1 microRNA Name (A)	Mature sequence variants (B)	Gene product (mRNA) targets (C)	Function of mRNA targets (D)	References
hiv1-miR-TAR-5p	**4-UCUCUCUGGUUAGACCAGAUCUGA-27 (Ra)** **UGGGUCUCUGGUUAGACCAG (Bp)** **GGUCUCUGGUUAGACCA (Nb)** **GGGUCUCUCUGGUUAGACCA (Nb)** **CUCUGGCUAACUACUAGGGAACCC (Ns)** **UCUGGCUAACUACUAGGGAA (Ns)** **UCUGGCUAACUACUAGGGAACCCA (Ns)** **CUGGCUAACUACUAGGGAA (Ns)** **UGGCUAACUACUAGGGAA (Ns)** **UGGCUAACUACUAGGGAACCCAC (Ns)** **UGGCUAACUACUAGGGAACCCACU (Ns)** **UGGCUAACUACUAGGGAACCCACUG (Ns)** **GGCUAACUACUAGGGAACCCACUG (Ns)** **CUAACUACUAGGGAACCCACUGC (Ns)**	**LTR *d*** **TAR *d*** **RITS *d*** ERCC1(apo) ***d*** IER3 (apo) ***d***	Viral gene expression Regulation/anti-apoptosis Viral co-factor Cell apoptotic factor Cell apoptotic factor	Klase et al., 2007, 2009; Ouellet et al., 2008; Schopman et al., 2012
hiv1-miR-TAR-3p	**38-UCUCUGGCUAACUAGGGAACCCA-60 (Ra)** **CUAACUAGGGAACCCAC (Nb)** **GCUAACUAGGGAACCCAC (Nb)** **GCUAACUAGGGAACCCACUG (Nb)**	**TAR *d*** **RITS *d***	Viral gene expression Regulation/anti-apoptosis	Klase et al., 2007; Schopman et al., 2012
hiv1-miR-H1	**2-CCAGGG-AGGCGUGCCUGGGC-21 (Nb)** **CCAGGG-AGGCGUGgCaUGGGC (Mc)** **CCAGGG-AGGCGUGgCCUGGGC (Mc)** **CCAGGG-AGGCGUGgCCUGGGC (Mc)** **CCAuGGgAGGCGcGgCCUGGGC (Mc)** **CCAGGG-AGGCGUGgCCgGGGu (Mc)** **CCAGGGgAGGCGUGaCCUGGGC (Mc)**	AATF ***d*** **BCL2** ***d*** MYC ***d*** PAWR ***d***; **DICER** ***d*** hsa-miR-194 ***d***	Adaptive immunity Activated cell Proliferation regulator Pro-viral latency promoter MicroRNA processing Pre-microRNA Processing/microRNA binding	Kaul et al., 2009; Lamers et al., 2010
hiv1-miR-N367	**40-ACUGACCUUUGGAUGGUGCUUCAA-62 (Nb, Mc)**	**Nef *d***	Transcription factor and regulator	Omoto and Fujii, 2005

Notes: ***(A)*** *The official names of microRNAs as published in mirbase.org.* ***(B)*** *The sequences of mature HIV-1 microRNA variants as determined by several methods including: Bp, bioinformatic prediction; Ra, RNase protection assays; Nb, Northern blotting; Ns, next generation sequencing; Mc, molecular cloning. Please note that sequences are not aligned accordingly; lowercase letters indicate polymorphic sites when available. The numbers before and after the nucleotide sequences refer to the relative genomic position/s in the pre-microRNA sequences if made available by authors in literature.* ***(C)*** *The mRNA targets of the HIV-1 microRNAs immediately followed by italicized letters correspond to the type of regulation, where: u = up-regulation, d = down-regulation when described in literature. In addition, if targets are HIV-1 mRNA genes or mRNA transcripts they are typed in* ***boldface****, RNAi pathway-related gene products typed in* ***BLUE****; and literature-based standard HIV-1 linked cellular gene products are typed in* ***RED****.* ***(D)*** *The reported functional attributes of mRNA targets by HIV-1 microRNA among studies.*

microRNAs, bi-target co-regulation is a prospective occurrence when a microRNA seed sequence would complement both viral and cellular mRNAs simultaneously due to sequence similarities. Though, the mechanisms of these interactivities are not thoroughly explainable at the moment, computational analyses suggest the probabilities of their existence, in particular during viral infection (Veksler-Lublinsky et al., 2010).

ALTERATIONS IN CELLULAR MicroRNA EXPRESSION DURING HIV-1 INFECTION

HIV-1 infection induced changes in cellular microRNA expressions result from combinatorial molecular interactions among proteins, transcripts, and genomes. These manifest as circuitous microRNA attenuation of the different cellular host metabolic processes. At this point, the current knowledge of the exact mechanisms on how HIV-1 infection remains to be understood fully. The current data available are mostly derived from microarray data comparing non-infected and HIV-1 infected cell lines. Over expression analyses of microRNAs among various cell lines simulated with HIV-1 infection are usually used to validate these differences in an attempt to explain the possible regulatory mechanisms behind the miroRNA interactions. Examples include the (Houzet et al., 2008) investigation which reported 59 simultaneously down-regulated cellular microRNAs of HIV-1 infected individuals. Prior studies indicated that HIV-1 infection can down-regulate as much as 43% of the 312 microRNA gene arrays (Yeung et al., 2005). In addition, unique and variable global modifications in microRNA expressions were exhibited by different host cell types and lines in reaction to HIV-1 infection (Yeung et al., 2005; Bennasser et al., 2006; Noorbakhsh et al., 2010; Gupta et al., 2011). A most recent example is the notable differences among the 21 microRNA profiles between the elite long-term non-progressors where viral replication is continuously suppressed against multiple uninfected individuals from 377 microRNAs changes in HIV-1 infected CD4+ lymphocytes (Bignami et al., 2012). Determining the alterations in cellular microRNA expression patterns among various types of HIV-1 infected cell lines may account factors such as productive infection and constant exposure to HIV-1 that drive these changes. Also modifications in microRNA profiles may be attributed to temporal variability of immune responses during the course of HIV-1 infection. When fully elucidated, these patterned variations of microRNA expressions may reflect vital information in HIV-1 disease diagnostics and progression. Currently, microarray data has been the greatest source of these analyses of microRNA expression pattern changes, aided by complex algorithms to detect actual variation.

An in depth analyses of these global changes confirm the existence of clustered microRNA expression signatures in HIV-1 infected cells. For example, downregulation of polycistronic microRNA hsa-miR-17/92 consistently suppressed viral production as observed among various HIV-1 infected cells (Triboulet et al., 2007). While, hsa-miR-27b, hsa-miR-29b, hsa-miR-150, and hsa-miR-223 were identified as significantly down-regulated upon CD4(+) T cell activation (Chiang et al., 2012). In contrast, hsa-miR-28, hsa-miR-125b, hsa-miR-150, hsa-miR-223, and hsa-miR-382, which were enriched in resting CD4+ T cells against the activated CD4+ T cells (Huang et al., 2007). Moreover, the T-cell-specific microRNAs, namely hsa-miR-150, hsa-miR-191, hsa-miR-223, hsa-miR-16, and hsa-miR-146b, showed variable expression patterns at various stages of HIV-1 infection. Recently, gene expression profile analyses of cellular microRNAs in HIV-1 infected CD4+ T cells demonstrated the down-regulation of hsa-miR-21, hsa-miR-26a, hsa-miR-155, hsa-miR-29a, hsa-miR-29b, and hsa-miR-29c, contrary to the observed upregulation of hsa-miR-223 (Sun et al., 2012). The identification of microRNA families may hold significance in correlating the targets as orthologous modules as previously mentioned.

HIV-1 ENCODED MicroRNAs AND THEIR INTERACTIONS

The low number of verified HIV-1 encoded microRNAs (**Table 1**) in the miRBase (2012) confirm the difficulty of their identification thus making them among the least characterized of RNA virus-generated microRNAs (Grundhoff and Sullivan, 2011). This scarcity may be due to their inherent low number because of the small genome size or low levels of expression currently undetectable by conventional biochemical techniques thus may require enrichment processes to be detected (Althaus et al., 2012). Previously, Lin and Cullen (2007) estimated that retroviral microRNAs comprise only 0.5% of the total microRNAs detectable in HIV-1 infected cells. In addition, the limited access of viruses to nuclear microRNA processing machinery and the natural destabilization effects of microRNA processing may also limit their biogenesis (Grundhoff and Sullivan, 2011). Also, several reports confirmed the endonucleolytic effects of Dicer or Drosha against viral RNA genomes thus reducing viral mRNA production (Ouellet and Provost, 2010).

However, the advent of highly sensitive technologies like next generation sequencing and RNAse protection assays (RPA), as well as improved computational prediction may contribute to the discovery of new HIV-1 microRNA species. Recent pyrosequencing results estimated at least 40% or 125 of the candidates as putative HIV-1 microRNAs originating from the TAR, RRE and *nef* region, and major components of non-coding RNAs in HIV-1 infected cells (Yeung et al., 2009). The deep sequencing report of (Schopman et al., 2012) further supports these observations, as HIV-1 microRNAs are suggested to arise from structured regions of the genome which facilitate Drosha and Dicer mediated RNA processing. Hence, with the increased possibilities that many putative HIV-1 microRNAs identified by these breakthrough procedures, they require further investigations on their isolation and functional characterization.

In general, the target interactions of HIV-1 microRNAs with its mRNA seem to function as viral genome regulators (**Table 1**). However, current experiments open this into a subject of debate and further investigation. Although, functional studies suggest auxiliary functions of HIV-1 microRNAs which target host cellular transcripts mainly for immune evasion (Boss and Renne, 2011). These observations are explained further in succeeding discussions below.

HIV-1 TAR MicroRNA

The 50 nt HIV-1 TAR element within the 5′ region of the viral RNA serves as the progenitor of hiv1-miR-TAR via asymmetrical

processing of the transcript (Ouellet et al., 2008). Experiments confirmed hiv1-miR TAR to target host cell microRNA-related proteins, namely, Dicer, trans-activation responsive RNA binding protein (TARBP2, TRBP), and the RNA induced transcriptional silencing (RITS) complex. Recent functional studies showed that HIV-1 TAR microRNA down-regulates the DNA excision repair (ERCC1) and radiation-inducible immediate-early gene IEX-1 proteins (IER3) thus exerting its anti-apoptotic effect in infected cells (Klase et al., 2009).

Computer simulation studies established the HIV-1 TAR element as a potential microRNA rich region because of the following evidences (Narayanan et al., 2011): (a) the hairpin formation of the TAR element concurs with Dicer substrate specifications, allowing the complement fit to five distinct Dicer element, (b) TAR binds to important microRNA proteins, Dicer, and TARBP2, thus singling out its essential role in the microRNA-mediated gene regulatory processes. Moreover, the TARBP2 association with Dicer is necessary for efficient loading of microRNAs into the RISC, the consequential loss of TARBP2 function culminates in the loss of RNA silencing ability (Sanghvi and Steel, 2011a). The TARBP2 sequestration is known to restrict the availability of Dicer enzyme leading to modification of microRNA processing (Haase et al., 2005). Furthermore, TARBP2 and TAR element association suppresses interferon (IFN)-induced protein kinase R function (Gatignol et al., 2005)

Cloning studies of TAR-related microRNAs demonstrated a greater abundance of the 3′ mature sequence over the 5′ mature sequences involved in microRNA-derived silencing (Lamers et al., 2010). These observations collectively suggest that, in infected cells, hiv1-miR-TAR-3p is superior to the hiv1-miR-TAR-5p in suppressing gene expression, supporting speculations that there are preferential releases of these microRNA species. This also corroborates to the evident accumulation of the 3′ HIV-1 TAR RNAs *in vivo* (Ouellet et al., 2008).

HIV-1 H1 MicroRNA

The 81bp stem loop of HIV-1 transcript formed in the 3′-U3 (LTR) region known as the binding sites of the two nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB) is the origin of hiv1-miR-H1 (MI0006106). It was shown to degrade apoptosis antagonizing factor (AATF) which decreases cell viability and reduced expression of cellular factors, Bcl-2, c-myc, Par-4 as well as the microRNA Dicer protein (Kaul, 2007; Kaul et al., 2009). The report also indicated hiv1-miR-H1 interaction with hsa-mir-149, affecting the latter's target HIV-1 encoded Vpr protein (Kaul et al., 2009). It is assumed that hiv1-miR-H1 and hiv1-miR-TAR are antagonistic to one another as they have contrasted activities against apoptotic elements. Further functional studies demonstrated deletion-driven evolution patterns in hiv1-miR-H1 among various AIDS patients. In addition, causal association was suggested between the appearance of a less stable hiv1-miR-H1 variant and induction of AIDS-related lymphoma (Lamers et al., 2010).

HIV-1 NEF MicroRNA

Nef protein has been shown to downregulate cell surface CD4 and MHC class I molecules through the clathrin-mediated endocytic pathway (Lubben et al., 2007; Schaefer et al., 2008). It is also involved in cellular signal transduction pathway through interaction with non-receptor type Tyr kinase molecules such as, Fyn and Lyn. Since, Nef functions in favor of HIV-1 replication and it is relatively conserved among various HIV-1 variants, miRNA-mediated control of Nef could have a great effect on the viral life cycle and its pathogenesis (Arien and Verhasselt, 2008; Foster and Garcia, 2008; Malim and Emerman, 2008). Although, HIV-1 3′ LTR is partially overlapping with *nef* microRNA (hiv1-miR-N367), as proposed previously (Omoto and Fujii, 2005, 2006), it remained controversial due to its non-duplicability. These reports may support a hypothesis of hyper-evolution of HIV-1 genome as a consequence of peptide-based immunity and RNA interference mechanisms (Narayanan et al., 2011).

HIV-1 ANTISENSE MicroRNAs

Recent reports have indicated the ability of HIV-1 utilizing antisense transcripts in infected cells leading to discoveries of new viral microRNAs. In the RACE analyses of HIV-1 infected 293T and Jurkat cells, it was shown that cryptic transcription initiation sites in the 5′ border of the 3′ LTR and a new poly A signal within this LTR were present; also indicated was the possible role of the Tat protein as the modulator of transcription of this antisense RNA (Landry et al., 2007). In another study, an antisense peptide open reading frame (ORF) called "asp" coding for a hydrophobic protein was derived from Jurkat cells infected with HIV-1 although its origin, generation or the function is not yet clarified (Clerc et al., 2011). As the existence of antisense HIV-1 microRNAs remains to be proven, this concept opens a possibility where long antisense transcripts can complement with the sense transcripts within the viral genome. These sites in double stranded configuration can provide biogenic zones of Dicer-mediated microRNAs (Houzet and Jeang, 2011).

CELLULAR MicroRNAs INVOLVED IN HIV-1 INFECTION

HIV-1 infection triggers multi-modal cascades of host cell microRNA targeting interactions that either activate or inhibit viral invasion and replication as shown in **Figure 1**. These microRNA targeting scenarios are likely to occur on at least two fronts. First, the cellular microRNA might directly target the HIV-1 genome, either in sense or antisense orientation, to suppress the production of viral proteins. An outstanding example is hsa-miR-29 which targets the HIV-1 nef transcript (Hariharan et al., 2005; Ahluwalia et al., 2008). Since, the HIV-1 nef gene is located at the proviral DNA 3′ portion, cellular microRNA targeting of this region would have serious implications in the viral life cycle. The group of (Nathans et al., 2009) proved that ectopic expression of hsa-miR-29 can repress production of *nef* protein resulting to suppressed viral replication and infectivity. The study also reported that hsa-mir-29a/ HIV-1 interactions enhance viral mRNA associations with RISC and P-body structures, thus suggesting prospective roles of P-bodies to viral latency regulation. In another study, a set of microRNAs namely, hsa-mir-28, hsa-miR-125b, has-miR-150, hsa-miR-223, and hsa-miR-382 were shown to bind in the 3′ position of HIV-1 transcripts which triggers viral latency (Huang et al., 2007). Recently, Sun et al identified another set of cellular microRNA, namely hsa-miR-15a, hsa-miR-15b,

hsa-miR-16, hsa-miR-24, hsa-miR-29a, hsa-miR-29b, hsa-miR-150, and hsa-miR-223 that are directly targeting HIV-1 3′-UTR, and exhibiting weak but significant inhibitory effects on HIV-1 replication (Sun et al., 2012). In a review by Sun and Rossi, using the PITA software, 256 seed-match sites were identified to complement Nef-3′ LTR sequence (Sun and Rossi, 2011).

In the second scenario, the host cell as triggered by HIV-1 infection would initiate cellular microRNA production to attenuate cellular factors involved in antiviral responses against HIV-1. **Table 2** summarizes the list of cellular microRNAs with their validated host cellular protein targets and their corresponding cellular functions. As likely initial repercussions, these microRNAs may target the genes involved in immune responses for innate and adaptive immunity (Kulpa and Collins, 2011). This bipartite defense system initially triggers natural killer (NK) cell activities as elicited by partial detection of HIV-1 components. In a later reaction, adaptive immunity is induced through production of antigen-specific antibodies by B-cells and eliciting cell-mediated immunity through antigen-specific cytotoxic T lymphocytes, of which microRNAs were found to target various cellular receptors (Cobos-Jimenez et al., 2011). As examples are cellular microRNA interactions with chemokines (Zhou et al., 2010).

Transcriptional control is vital to the HIV- 1 proliferation, thus determining microRNA interactions among host transcription factors and regulators is a necessity (Victoriano and Okamoto, 2012). Among examples are reporter assays suggesting hsa-miR-223 bi-functional effects in HIV-1 replication are targets were varied in two different cell lines namely, Sp3 and LIF in NB4 cells, while RhoB and NF-1A in HEK293 cells (Sun et al., 2012). Next is the hsa-miR-29 family which also targets cellular proteins Mcl-1, DNMT 3A/B, Tcl1, p85, and CDC42, further establishes its diverse roles in HIV-1 latency (Sun et al., 2012; Witwer et al., 2012). Another is hsa-miR-198 targeting CCNT1 (Kaul et al., 2009), a key component in the Tat-mediated transcription of the virus, when suppressed can impair HIV-1 replication (Imai et al., 2009).

Table 2 | List of published cellular microRNAs and their target HIV-1 and cellular gene products.

Cellular microRNA Name (A)	Gene product (mRNA) targets (B)	Function (C)	References
hsa-let-7/g*	**DICER**	Pre-microRNA processing/microRNA binding	Faller and Guo, 2008; Faller et al., 2010; Desjardins et al., 2012; Swaminathan et al., 2012
	LIN28	Pre-microRNA processing regulation; repress maturation of hsa-let-7 family; blocks Drosha and Dicer processing of pri-/pre hsa-let-7 family via interaction with terminal loop; blocks Dicer processing of pre- hsa-miR-128	
	IL-10	Inflammatory response	
hsa-miR-17*/17-3p	**EP300**/CBP associated factor (PCAF) ***u***	Transcription factor and regulator/control of viral replication	Triboulet et al., 2007; Hayes et al., 2011
hsa-miR-17/17-5p	KAT8	HIV-1 Tat interactive protein	
	HA Tat co-factor	HIV-1 Tat interactive protein	
hsa-miR-92a-1*	HA Tat cofactor	HIV-1 Tat interactive protein	Sun et al., 2012
hsa-miR-125b-5p	**Nef-3′ UTR LTR**	Viral replication and promotion of viral latency in T-cells	Huang et al., 2007; Witwer et al., 2012
hsa-miR-125b-1*/125b-2*			
hsa-miR-125a-5p			
hsa-miR-125a-3p			
hsa-miR-128	SNAP25	Cellular receptor	Eletto et al., 2008
hsa-miR-146	**CCL8/MCP-2**	Innate immune response factor	Rom et al., 2010
hsa-miR-149	**Vpr**	Regulation of nuclear import of HIV-1 pre-integration complex; viral replication and cellular immune suppression	Kaul et al., 2009
hsa-miR-150/150*	**3′ end of HIV-1 RNA**	Viral replication and promotion of viral latency in T-cells	Huang et al., 2007; Witwer et al., 2012
	APOBEC3G/3F *d*	Cellular co-factor; relieves microRNA repression mechanisms	
	CCR5	HIV-1 receptor and natural ligand	
	CD4 *d*	HIV-1 receptor and natural ligand	
	CCNT1	Repression of HIV-1 tat co-factor for transcriptional *trans*-activation	
hsa-miR-155	Target not specified	Function not specified	Sun et al., 2012

(*Continued*)

Table 2 | Continued.

hsa-miR-198	**CCNT1** (P-TEFb) ***d***	Repression of HIV-1 Tat co-factor for transcriptional *trans*-activation	Sung and Rice, 2009
hsa-miR-20a	PCAF ***u*** KAT8 MCL1 DNMT3A/B TCL1A PIC3R1 **CDC42**	Co-factor of Tat *trans*-activation. Cellular transcription activator Cellular anti-apoptotic factor Cellular transcriptional regulator Interacts with IKB PI3 kinase subunit HIV-1 receptor and natural ligand	Hayes et al., 2011
hsa-miR-21	Target not specified	Function not specified	Sun et al., 2012
hsa-miR-27a*/27a hsa-miR-27b*/27b	**CCNT1**	Transcription factor and regulator; repression of HIV-1 Tat co-factor for transcriptional *trans*-activation	Chiang et al., 2012
hsa-miR-28-5p/28-3p	**3′ end of HIV-1 RNA** **CCR5** **CD4** ***d*** **APOBEC3G/3F** ***d***	Viral replication and promotion of viral latency in T-cells HIV-1 receptor and natural ligand HIV-1 receptor and natural ligand Cellular co-factor	Huang et al., 2007; Swaminathan et al., 2009
hsa-miR-29a/29a* hsa-miR-29-b1*/**29b**[1 *d*]/29-b2* hsa-miR-29c*/**29c**	**Nef protein coding mRNA** 3′-**UTR** (420) **RISC**, P bodies MCL-1 DNMT 3A/B TCL1a, p85a CDC42 **CCNT1**	Viral replication and latency Viral replication and latency Mature microRNA assembly/carrier Cellular anti-apoptotic factor Cellular transcriptional regulator Interacts with IKB PI3 kinase subunit HIV-1 receptor and natural ligand Transcription factor and regulator; repression of HIV-1 Tat co-factor for transcriptional *trans*-activation	Ahluwalia et al., 2008; Nathans et al., 2009; Chiang et al., 2012; Sun et al., 2012; Witwer et al., 2012
hsa-miR-217	SIRT1	Cellular stress response regulator	Zhang et al., 2012
hsa-miR-223*/**223**	**3′ end of HIV-1 RNA** **APOBEC3G/3F** ***d*** P3 *d* LIF *d* RobB *d* **CCNT1**	Viral replication and promotion of viral latency in T-cells Cellular co-factor Cellular co-factor Cellular co-factor Cellular co-factor Transcription factor and regulator	Huang et al., 2007; Swaminathan et al., 2009; Chiang et al., 2012; Sun et al., 2012 Chiang et al., 2012
hsa-miR-31/31*	Target not specified	Function not specified	Witwer et al., 2012
hsa-miR-34a	**CREBBP**	Transcription factor and regulator	Chiang et al., 2012
hsa-miR-382	**3′ end of HIV-1 RNA**	Viral replication and promotion of viral latency in T-cells	Huang et al., 2007

Notes: ***(A)*** *The official names of microRNAs as published in mirbase.org. The microRNAs in* ***boldface*** *are the dominant targeting species when reported in literature.* ***(B)*** *The mRNA targets of the HIV-1 microRNAs are immediately followed by italicized letters which correspond to the type of regulation, where: u = up-regulation, d = down-regulation when described in literature. In addition, if targets are HIV-1 mRNA genes or mRNA transcripts they are typed in* ***boldface****, RNAi pathway-related gene products typed in* ***BLUE****; and literature-based standard HIV-1 linked cellular gene products are typed in* ***RED****.* ***(C)*** *The reported functional attributes of mRNA targets by HIV-1 microRNA among studies.*

Cellular microRNAs linked to chromatin regulation show proof that microRNAs are critical elements of epigenetic control in HIV-1 infection (Obbard et al., 2009; Easley et al., 2010). Recent evidences support that chromatin modification may explain mechanisms of HIV-1 transcription and thus the maintenance of latency. Some microRNA species are considered involved in gene silencing by modulating methylation and deactylation of histone proteins (Triboulet et al., 2007).

The above mentioned functional gene product clusters are just few focal points of cellular microRNA interactivities related to HIV-1 infection. It is expected that as more interactions are validated, the complex nature of cellular microRNA regulation linked to HIV-1 infection and host response would be further characterized. However, the scope of cellular microRNA interactions may involve other non-listed prospective gene targets which may also influence HIV-1 infection.

BEYOND CROSSTALKS AMONG CELLULAR AND HIV-1 MicroRNA MACHINERIES

Preceding discussions on microRNA interactions in host-HIV-1 infection further confirm their inherent complexity. It perfectly illustrates the constant attenuation of gene regulatory networks to maintain homeostasis in the HIV-1 infected cells. However, as HIV-1 remains an incurable disease among humans, it is implied that it can successfully compromise host immune and defense reactions wherein microRNA regulation might play pivotal roles. Thus, future studies must focus on how to reprogram microRNAs to favorably initiate the cellular anti-HIV-1 defense response. To realize such goal, it becomes necessary to organize succeeding investigations as follows: First is to globally account cellular and viral microRNA interrelationships affecting biomolecular pathways in HIV-1 infection. This allows the possibility of unlocking the combination of molecular switches that would allow the host cell successfully defend itself against HIV-1. Second is to determine the simultaneous targets of viral and cellular microRNAs. These bi-targets may reveal signatures of gene families or microRNA clusters characterizing HIV-1 infection patterns. Third is to capture temporal changes among microRNA expression patterns during HIV-1 disease progression. In assessing the current amount of information on hand, there remains much work to be done in unlocking the ultimate roles of microRNAs in HIV-1 pathogencity.

ACKNOWLEDGMENTS

This work was partially supported by the Ministries of Education, Culture, Sports, Science and Technology, and Health, Labor, and Welfare of Japan. We sincerely express our gratitude to the invaluable support of Drs. Kaori Asamitsu, Satoshi Kanazawa, and Hiroaki Uranishi of the Department of Cell and Molecular Biology, Nagoya City University Graduate School of Medical Sciences. We also express our sincerest gratitude for Mr. Issey Takahashi of the Nagoya City University Graduate School of Design and Architecture for rendering the scientific illustrations.

REFERENCES

Aagaard, L. A., Zhang, J., Von Eije, K. J., Li, H., Saetrom, P., Amarzguioui, M., and Rossi, J. J. (2008). Engineering and optimization of the miR-106b cluster for ectopic expression of multiplexed anti-HIV RNAs. *Gene Ther.* 15, 1536–1549.

Ahluwalia, J. K., Khan, S. Z., Soni, K., Rawat, P., Gupta, A., Hariharan, M., Scaria, V., Lalwani, M., Pillai, B., Mitra, D., and Brahmachari, S. K. (2008). Human cellular microRNA hsa-miR-29a interferes with viral nef protein expression and HIV-1 replication. *Retrovirology* 5, 117.

Ajay, S. S., Athey, B. D., and Lee, I. (2010). Unified translation repression mechanism for microRNAs and upstream AUGs. *BMC Genomics* 11, 155.

Althaus, C. F., Vongrad, V., Niederost, B., Joos, B., Di Giallonardo, F., Rieder, P., Pavlovic, J., Trkola, A., Gunthard, H. F., Metzner, K. J., and Fischer, M. (2012). Tailored enrichment strategy detects low abundant small noncoding RNAs in HIV-1 infected cells. *Retrovirology* 9, 27.

Arbuthnot, P. (2011). MicroRNA-like antivirals. *Biochim. Biophys. Acta* 1809, 746–755.

Arien, K. K., and Verhasselt, B. (2008). HIV Nef: role in pathogenesis and viral fitness. *Curr. HIV Res.* 6, 200–208.

Bennasser, Y., and Jeang, K. T. (2006). HIV-1 Tat interaction with Dicer: requirement for RNA. *Retrovirology* 3, 95.

Bennasser, Y., Le, S. Y., Yeung, M. L., and Jeang, K. T. (2006). MicroRNAs in human immunodeficiency virus-1 infection. *Methods Mol. Biol.* 342, 241–253.

Berezikov, E., Liu, N., Flynt, A. S., Hodges, E., Rooks, M., Hannon, G. J., and Lai, E. C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nat. Genet.* 42, 6–9; Author reply 9–10.

Bignami, F., Pilotti, E., Bertoncelli, L., Ronzi, P., Gulli, M., Marmiroli, N., Magnani, G., Pinti, M., Lopalco, L., Mussini, C., Ruotolo, R., Galli, M., Cossarizza, A., and Casoli, C. (2012). Stable changes in CD4+ T lymphocyte miRNA expression after exposure to HIV-1. *Blood* 119, 6259–6267.

Bivalkar-Mehla, S., Vakharia, J., Mehla, R., Abreha, M., Kanwar, J. R., Tikoo, A., and Chauhan, A. (2011). Viral RNA silencing suppressors (RSS): novel strategy of viruses to ablate the host RNA interference (RNAi) defense system. *Virus Res.* 155, 1–9.

Boden, D., Pusch, O., and Ramratnam, B. (2007). Overcoming HIV-1 resistance to RNA interference. *Front. Biosci.* 12, 3104–3116.

Boden, D., Pusch, O., Silbermann, R., Lee, F., Tucker, L., and Ramratnam, B. (2004). Enhanced gene silencing of HIV-1 specific siRNA using microRNA designed hairpins. *Nucleic Acids Res.* 32, 1154–1158.

Bohnsack, M. T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* 10, 185–191.

Boss, I. W., and Renne, R. (2011). Viral miRNAs and immune evasion. *Biochim. Biophys. Acta* 1809, 708–714.

Brodersen, P., and Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. *Nat. Rev. Mol. Cell Biol.* 10, 141–148.

Carmell, M. A., and Hannon, G. J. (2004). RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol.* 11, 214–218.

Chiang, K., and Rice, A. P. (2011). Mini ways to stop a virus: microRNAs and HIV-1 replication. *Future Virol.* 6, 209–221.

Chiang, K., Sung, T.-L., and Rice, A. P. (2012). Regulation of Cyclin T1 and HIV-1 Replication by MicroRNAs in Resting CD4+ T Lymphocytes. *J. Virol.* 86, 3244–3252.

Clerc, I., Laverdure, S., Torresilla, C., Landry, S., Borel, S., Vargas, A., Arpin-Andre, C., Gay, B., Briant, L., Gross, A., Barbeau, B., and Mesnard, J. M. (2011). Polarized expression of the membrane ASP protein derived from HIV-1 antisense transcription in T cells. *Retrovirology* 8, 74.

Cobos-Jimenez, V., Booiman, T., Hamann, J., and Kootstra, N. A. (2011). Macrophages and HIV-1. *Curr. Opin. HIV AIDS* 6, 385–390.

Coley, W., Van Duyne, R., Carpio, L., Guendel, I., Kehn-Hall, K., Chevalier, S., Narayanan, A., Luu, T., Lee, N., Klase, Z., and Kashanchi, F. (2010). Absence of DICER in monocytes and its regulation by HIV-1. *J. Biol. Chem.* 285, 31930–31943.

Cullen, B. R. (2004). Derivation and function of small interfering RNAs and microRNAs. *Virus Res.* 102, 3–9.

Cullen, B. R. (2006). Viruses and microRNAs. *Nat. Genet.* 38(Suppl.), S25–S30.

Desjardins, A., Yang, A., Bouvette, J., Omichinski, J. G., and Legault, P. (2012). Importance of the NCp7-like domain in the recognition of pre-let-7g by the pluripotency factor Lin28. *Nucleic Acids Res.* 40, 1767–1777.

Easley, R., Van Duyne, R., Coley, W., Guendel, I., Dadgar, S., Kehn-Hall, K., and Kashanchi, F. (2010). Chromatin dynamics associated with HIV-1 Tat-activated transcription. *Biochim. Biophys. Acta* 1799, 275–285.

Eletto, D., Russo, G., Passiatore, G., Del Valle, L., Giordano, A., Khalili, K., Gualco, E., and Peruzzi, F. (2008). Inhibition of SNAP25 expression by HIV-1 Tat involves the activity

of mir-128a. *J. Cell. Physiol.* 216, 764–770.

Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S. F., Doerks, T., Dorner, S., Bork, P., Boutros, M., and Izaurralde, E. (2007). Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. *Genes Dev.* 21, 2558–2570.

Faller, M., and Guo, F. (2008). MicroRNA biogenesis: there's more than one way to skin a cat. *Biochim. Biophys. Acta* 1779, 663–667.

Faller, M., Toso, D., Matsunaga, M., Atanasov, I., Senturia, R., Chen, Y., Zhou, Z. H., and Guo, F. (2010). DGCR8 recognizes primary transcripts of microRNAs through highly cooperative binding and formation of higher-order structures. *RNA* 16, 1570–1583.

Flores-Jasso, C. F., Arenas-Huertero, C., Reyes, J. L., Contreras-Cubas, C., Covarrubias, A., and Vaca, L. (2009). First step in pre-miRNAs processing by human Dicer. *Acta Pharmacol. Sin.* 30, 1177–1185.

Foster, J. L., and Garcia, J. V. (2008). HIV-1 Nef: at the crossroads. *Retrovirology* 5, 84.

Frank, F., Sonenberg, N., and Nagar, B. (2010). Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature* 465, 818–822.

Gatignol, A., Laine, S., and Clerzius, G. (2005). Dual role of TRBP in HIV replication and RNA interference: viral diversion of a cellular pathway or evasion from antiviral immunity? *Retrovirology* 2, 65.

Grundhoff, A., and Sullivan, C. S. (2011). Virus-encoded microRNAs. *Virology* 411, 325–343.

Gupta, A., Nagilla, P., Le, H. S., Bunney, C., Zych, C., Thalamuthu, A., Bar-Joseph, Z., Mathavan, S., and Ayyavoo, V. (2011). Comparative expression profile of miRNA and mRNA in primary peripheral blood mononuclear cells infected with human immunodeficiency virus (HIV-1). *PLoS ONE* 6:e22730. doi: 10.1371/journal.pone.0022730

Haase, A. D., Jaskiewicz, L., Zhang, H., Laine, S., Sack, R., Gatignol, A., and Filipowicz, W. (2005). TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Rep.* 6, 961–967.

Hammond, S. M. (2005). Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett.* 579, 5822–5829.

Hariharan, M., Scaria, V., Pillai, B., and Brahmachari, S. K. (2005). Targets for human encoded microRNAs in HIV genes. *Biochem. Biophys. Res. Commun.* 337, 1214–1218.

Harvey, S. J., Jarad, G., Cunningham, J., Goldberg, S., Schermer, B., Harfe, B. D., McManus, M. T., Benzing, T., and Miner, J. H. (2008). Podocyte-specific deletion of dicer alters cytoskeletal dynamics and causes glomerular disease. *J. Am. Soc. Nephrol.* 19, 2150–2158.

Havens, M. A., Reich, A. A., Duelli, D. M., and Hastings, M. L. (2012). Biogenesis of mammalian microRNAs by a non-canonical processing pathway. *Nucleic Acids Res.* 40, 4626–4640.

Hayes, A. M., Qian, S., Yu, L., and Boris-Lawrie, K. (2011). Tat RNA silencing suppressor activity contributes to perturbation of lymphocyte miRNA by HIV-1. *Retrovirology* 8, 36.

Hon, L. S., and Zhang, Z. (2007). The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. *Genome Biol.* 8, R166.

Houzet, L., and Jeang, K. T. (2011). MicroRNAs and human retroviruses. *Biochim. Biophys. Acta* 1809, 686–693.

Houzet, L., Yeung, M. L., De Lame, V., Desai, D., Smith, S. M., and Jeang, K. T. (2008). MicroRNA profile changes in human immunodeficiency virus type 1 (HIV-1) seropositive individuals. *Retrovirology* 5, 118.

Huang, J., Wang, F., Argyris, E., Chen, K., Liang, Z., Tian, H., Huang, W., Squires, K., Verlinghieri, G., and Zhang, H. (2007). Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes. *Nat. Med.* 13, 1241–1247.

Huang, Y., Shen, X., Zou, Q., Wang, S., Tang, S., and Zhang, G. (2011). Biological functions of microRNAs: a review. *J. Physiol. Biochem.* 67, 129–139.

Imai, K., Asamitsu, K., Victoriano, A. F., Cueno, M. E., Fujinaga, K., and Okamoto, T. (2009). Cyclin T1 stabilizes expression levels of HIV-1 Tat in cells. *FEBS J.* 276, 7124–7133.

Jackson, R. J., and Standart, N. (2007). How do microRNAs regulate gene expression? *Sci. STKE* 2007, re1.

Kaul, D. (2007). Cellular AATF gene: armour against HIV-1. *Indian J. Biochem. Biophys.* 44, 276–278.

Kaul, D., Ahlawat, A., and Gupta, S. D. (2009). HIV-1 genome-encoded hiv1-mir-H1 impairs cellular responses to infection. *Mol. Cell. Biochem.* 323, 143–148.

Kawamata, T., and Tomari, Y. (2010). Making RISC. *Trends Biochem. Sci.* 35, 368–376.

Kim, V. N. (2004). MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends Cell Biol.* 14, 156–159.

Klase, Z., Kale, P., Winograd, R., Gupta, M. V., Heydarian, M., Berro, R., McCaffrey, T., and Kashanchi, F. (2007). HIV-1 TAR element is processed by Dicer to yield a viral micro-RNA involved in chromatin remodeling of the viral LTR. *BMC Mol. Biol.* 8:63. doi: 10.1186/1471-2199-8-63

Klase, Z., Winograd, R., Davis, J., Carpio, L., Hildreth, R., Heydarian, M., Fu, S., McCaffrey, T., Meiri, E., Ayash-Rashkovsky, M., Gilad, S., Bentwich, Z., and Kashanchi, F. (2009). HIV-1 TAR miRNA protects against apoptosis by altering cellular gene expression. *Retrovirology* 6, 18.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152–D157.

Kulpa, D. A., and Collins, K. L. (2011). The emerging role of HLA-C in HIV-1 infection. *Immunology* 134, 116–122.

Lamers, S. L., Fogel, G. B., and McGrath, M. S. (2010). HIV-miR-H1 evolvability during HIV pathogenesis. *BioSystems* 101, 88–96.

Landi, D., Barale, R., Gemignani, F., and Landi, S. (2011). Prediction of the biological effect of polymorphisms within microRNA binding sites. *Methods Mol. Biol.* 676, 197–210.

Landry, S., Halin, M., Lefort, S., Audet, B., Vaquero, C., Mesnard, J. M., and Barbeau, B. (2007). Detection, characterization and regulation of antisense transcripts in HIV-1. *Retrovirology* 4, 71.

Lin, J., and Cullen, B. R. (2007). Analysis of the interaction of primate retroviruses with the human RNA interference machinery. *J. Virol.* 81, 12218–12226.

Lin, S. L., Chang, D., and Ying, S. Y. (2005). Asymmetry of intronic pre-miRNA structures in functional RISC assembly. *Gene* 356, 32–38.

Liu, Y. P., Westerink, J. T., Ter Brake, O., and Berkhout, B. (2011). RNAi-inducing lentiviral vectors for anti-HIV-1 gene therapy. *Methods Mol. Biol.* 721, 293–311.

Lo, H. L., Chang, T., Yam, P., Marcovecchio, P. M., Li, S., Zaia, J. A., and Yee, J. K. (2007). Inhibition of HIV-1 replication with designed miRNAs expressed from RNA polymerase II promoters. *Gene Ther.* 14, 1503–1512.

Long, D., Chan, C. Y., and Ding, Y. (2008). Analysis of microRNA-target interactions by a target structure based hybridization model. *Pac. Symp. Biocomput.* 2008, 64–74.

Lubben, N. B., Sahlender, D. A., Motley, A. M., Lehner, P. J., Benaroch, P., and Robinson, M. S. (2007). HIV-1 Nef-induced down-regulation of MHC class I requires AP-1 and clathrin but not PACS-1 and is impeded by AP-2. *Mol. Biol. Cell* 18, 3351–3365.

Malim, M. H., and Emerman, M. (2008). HIV-1 accessory proteins–ensuring viral survival in a hostile environment. *Cell Host Microbe* 3, 388–398.

Narayanan, A., Kehn-Hall, K., Bailey, C., and Kashanchi, F. (2011). Analysis of the roles of HIV-derived microRNAs. *Expert Opin. Biol. Ther.* 11, 17–29.

Nathans, R., Chu, C. Y., Serquina, A. K., Lu, C. C., Cao, H., and Rana, T. M. (2009). Cellular microRNA and P bodies modulate host-HIV-1 interactions. *Mol. Cell* 34, 696–709.

Noorbakhsh, F., Ramachandran, R., Barsby, N., Ellestad, K. K., Leblanc, A., Dickie, P., Baker, G., Hollenberg, M. D., Cohen, E. A., and Power, C. (2010). MicroRNA profiling reveals new aspects of HIV neurodegeneration: caspase-6 regulates astrocyte survival. *FASEB J.* 24, 1799–1812.

Obbard, D. J., Gordon, K. H., Buck, A. H., and Jiggins, F. M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 99–115.

Okada, C., Yamashita, E., Lee, S. J., Shibata, S., Katahira, J., Nakagawa, A., Yoneda, Y., and Tsukihara, T. (2009). A high-resolution structure of the pre-microRNA nuclear export machinery. *Science* 326, 1275–1279.

Okamura, K., Chung, W. J., and Lai, E. C. (2008). The long and short of inverted repeat genes in animals: microRNAs, mirtrons and hairpin RNAs. *Cell Cycle* 7, 2840–2845.

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., and Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* 130, 89–100.

Omoto, S., and Fujii, Y. R. (2005). Regulation of human immunodeficiency virus 1 transcription by nef microRNA. *J. Gen. Virol.* 86, 751–755.

Omoto, S., and Fujii, Y. R. (2006). Cloning and detection of HIV-1-encoded microRNA. *Methods Mol. Biol.* 342, 255–265.

Ouellet, D. L., Plante, I., Landry, P., Barat, C., Janelle, M. E., Flamand, L., Tremblay, M. J., and Provost, P. (2008). Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element. *Nucleic Acids Res.* 36, 2353–2365.

Ouellet, D. L., and Provost, P. (2010). Current knowledge of MicroRNAs and noncoding RNAs in virus-infected cells. *Methods Mol. Biol.* 623, 35–65.

Pan, W., Xin, P., and Clawson, G. A. (2010). MicroRNAs align with accessible sites in target mRNAs. *J. Cell. Biochem.* 109, 509–518.

Qian, S., Zhong, X., Yu, L., Ding, B., De Haan, P., and Boris-Lawrie, K. (2009). HIV-1 Tat RNA silencing suppressor activity is conserved across kingdoms and counteracts translational repression of HIV-1. *Proc. Natl. Acad. Sci. U.S.A.* 106, 605–610.

Rom, S., Rom, I., Passiatore, G., Pacifici, M., Radhakrishnan, S., Del Valle, L., Pina-Oviedo, S., Khalili, K., Eletto, D., and Peruzzi, F. (2010). CCL8/MCP-2 is a target for mir-146a in HIV-1-infected human microglial cells. *FASEB J.* 24, 2292–2300.

Sanchez-Del Cojo, M., Lopez-Huertas, M. R., Mateos, E., Alcami, J., and Coiras, M. (2011). Mechanisms of RNA interference in the HIV-1-host cell interplay. *AIDS Rev.* 13, 149–160.

Sanghvi, V. R., and Steel, L. F. (2011a). The cellular TAR RNA binding protein, TRBP, promotes HIV-1 replication primarily by inhibiting the activation of double-stranded RNA-dependent kinase PKR. *J. Virol.* 85, 12614–12621.

Sanghvi, V. R., and Steel, L. F. (2011b). A re-examination of global suppression of RNA interference by HIV-1. *PLoS ONE* 6:e17246. doi: 10.1371/journal.pone.0017246

Schaefer, M. R., Wonderlich, E. R., Roeth, J. F., Leonard, J. A., and Collins, K. L. (2008). HIV-1 Nef targets MHC-I and CD4 for degradation via a final common beta-COP-dependent pathway in T cells. *PLoS Pathog.* 4:e1000131. doi: 10.1371/journal.ppat.1000131

Schopman, N. C., Willemsen, M., Liu, Y. P., Bradley, T., Van Kampen, A., Baas, F., Berkhout, B., and Haasnoot, J. (2012). Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs. *Nucleic Acids Res.* 40, 414–427.

Seitz, H. (2009). Redefining microRNA targets. *Curr. Biol.* 19, 870–873.

Sun, G., Li, H., Wu, X., Covarrubias, M., Scherer, L., Meinking, K., Luk, B., Chomchan, P., Alluin, J., Gombart, A. F., and Rossi, J. J. (2012). Interplay between HIV-1 infection and host microRNAs. *Nucleic Acids Res.* 40, 2181–2196.

Sun, G., and Rossi, J. J. (2011). MicroRNAs and their potential involvement in HIV infection. *Trends Pharmacol. Sci.* 32, 675–681.

Sung, T. L., and Rice, A. P. (2009). miR-198 inhibits HIV-1 gene expression and replication in monocytes and its mechanism of action appears to involve repression of cyclin T1. *PLoS Pathog.* 5:e1000263. doi: 10.1371/journal.ppat.1000263

Swaminathan, S., Suzuki, K., Seddiki, N., Kaplan, W., Cowley, M. J., Hood, C. L., Clancy, J. L., Murray, D. D., Méndez, C., Gelgor, L., Anderson, B., Roth, N., Cooper, D. A., and Kelleher, A. D. (2012). Differential regulation of the Let-7 family of microRNAs in CD4+ T cells alters IL-10 expression. *J. Immunol.* 188, 6238–6246.

Swaminathan, S., Zaunders, J., Wilkinson, J., Suzuki, K., and Kelleher, A. D. (2009). Does the presence of anti-HIV miRNAs in monocytes explain their resistance to HIV-1 infection? *Blood* 113, 5029–5030; Author reply 5030–5031.

Tan Gana, N. H., Victoriano, A. F., and Okamoto, T. (2012). Evaluation of online miRNA resources for biomedical applications. *Genes Cells* 17, 11–27.

Triboulet, R., Mari, B., Lin, Y. L., Chable-Bessia, C., Bennasser, Y., Lebrigand, K., Cardinaud, B., Maurin, T., Barbry, P., Baillat, V., Reynes, J., Corbeau, P., Jeang, K. T., and Benkirane, M. (2007). Suppression of microRNA-silencing pathway by HIV-1 during virus replication. *Science* 315, 1579–1582.

Veksler-Lublinsky, I., Shemer-Avni, Y., Kedem, K., and Ziv-Ukelson, M. (2010). Gene bi-targeting by viral and human miRNAs. *BMC Bioinformatics* 11, 249.

Victoriano, A. F., and Okamoto, T. (2012). Transcriptional control of HIV replication by multiple modulators and their implication for a novel antiviral therapy. *AIDS Res. Hum. Retroviruses* 28, 125–138.

Westholm, J. O., Ladewig, E., Okamura, K., Robine, N., and Lai, E. C. (2012). Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs. *RNA* 18, 177–192.

Westholm, J. O., and Lai, E. C. (2011). Mirtrons: microRNA biogenesis via splicing. *Biochimie* 93, 1897–1904.

Witwer, K. W., Watson, A. K., Blankson, J. N., and Clements, J. E. (2012). Relationships of PBMC microRNA expression, plasma viral load, and CD4+ T-cell count in HIV-1-infected elite suppressors and viremic patients. *Retrovirology* 9, 5.

Yeung, M. L., Bennasser, Y., Myers, T. G., Jiang, G., Benkirane, M., and Jeang, K. T. (2005). Changes in microRNA expression profiles in HIV-1-transfected human cells. *Retrovirology* 2, 81.

Yeung, M. L., Bennasser, Y., Watashi, K., Le, S. Y., Houzet, L., and Jeang, K. T. (2009). Pyrosequencing of small non-coding RNAs in HIV-1 infected cells: evidence for the processing of a viral-cellular double-stranded RNA hybrid. *Nucleic Acids Res.* 37, 6575–6586.

You, X., Zhang, Z., Fan, J., Cui, Z., and Zhang, X. E. (2012). Functionally orthologous viral and cellular microRNAs studied by a novel dual-fluorescent reporter system. *PLoS ONE* 7:e36157. doi: 10.1371/journal.pone.0036157

Zhang, H. S., Wu, T. C., Sang, W. W., and Ruan, Z. (2012). MiR-217 is involved in Tat-induced HIV-1 long terminal repeat (LTR) transactivation by down-regulation of SIRT1. *Biochim. Biophys. Acta* 1823, 1017–1023.

Zhou, Y., Wang, X., Liu, M., Hu, Q., Song, L., Ye, L., Zhou, D., and Ho, W. (2010). A critical function of toll-like receptor-3 in the induction of anti-human immunodeficiency virus activities in macrophages. *Immunology* 131, 40–49.

Conflict of Interest Statement: 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.

Received: 06 June 2012; paper pending published: 17 June 2012; accepted: 01 August 2012; published online: 24 August 2012.

Citation: Tan Gana NH, Onuki T, Victoriano AFB and Okamoto T (2012) MicroRNAs in HIV-1 infection: an integration of viral and cellular interaction at the genomic level. Front. Microbio. ***3****:306. doi: 10.3389/fmicb.2012.00306*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 August 2012
doi: 10.3389/fmicb.2012.00277

Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing

***Ponraj Prabakaran*[1,2]*, *Zhongyu Zhu*[1], *Weizao Chen*[1], *Rui Gong*[1], *Yang Feng*[1], *Emily Streaker*[1,2] and *Dimiter S. Dimitrov*[1]**

[1] *CCR Nanobiology Program, Protein Interactions Group, Frederick National Laboratory for Cancer Research, National Institutes of Health (NIH), Frederick, MD, USA*
[2] *Basic Research Program, Science Applications International Corporation-Frederick, Inc., NCI-Frederick, Frederick, MD, USA*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Takamasa Takeuchi, National Institute of Infectious Diseases, Japan
Shuzo Matsushita, Kumamoto University, Japan

***Correspondence:**
Ponraj Prabakaran, CCR Nanobiology Program, Protein Interactions Group, Frederick National Laboratory for Cancer Research, National Institutes of Health (NIH), Bldg 469, Rm 140, Frederick, MD 21702, USA.
e-mail: prabakaran.ponraj@nih.gov

Our understanding of how antibodies are generated and function could help develop effective vaccines and antibody-based therapeutics against viruses such as HIV-1, SARS coronavirus (SARS CoV), and Hendra and Nipah viruses (henipaviruses). Although broadly neutralizing antibodies (bnAbs) against the HIV-1 were observed in patients, elicitation of such bnAbs remains a major challenge when compared to other viral targets. We previously hypothesized that HIV-1 could have evolved a strategy to evade the immune system due to absent or very weak binding of germline antibodies to the conserved epitopes that may not be sufficient to initiate and/or maintain an effective immune response. To further explore our hypothesis, we used the 454 sequence analysis of a large naïve library of human IgM antibodies which had been used for selecting antibodies against SARS CoV receptor-binding domain (RBD), and soluble G proteins (sG) of henipaviruses. We found that the human IgM repertoires from the 454 sequencing have diverse germline usages, recombination patterns, junction diversity, and a lower extent of somatic mutation. In this study, we identified antibody maturation intermediates that are related to bnAbs against the HIV-1 and other viruses as observed in normal individuals, and compared their genetic diversity and somatic mutation level along with available structural and functional data. Further computational analysis will provide framework for understanding the underlying genetic and molecular determinants related to maturation pathways of antiviral bnAbs that could be useful for applying novel approaches to the design of effective vaccine immunogens and antibody-based therapeutics.

Keywords: HIV-1, vaccine, monoclonal antibody, IgM, immunogen, 454 sequencing

INTRODUCTION

Broadly neutralizing antibodies (bnAbs) against the HIV-1 are relatively rarely observed in patients; however, discovering HIV-1 vaccine candidates to elicit such bnAbs remains a challenge due to the extensive genetic sequence variability and complex immune evasion strategies of the HIV-1 (Burton, 2002; Johnson and Desrosiers, 2002; Haynes and Montefiori, 2006; Prabakaran et al., 2007). Among the different factors thwarting the induction of bnAbs, we previously found that all known HIV-1 bnAbs are highly divergent from germline antibodies; germline antibodies of bnAbs could not bind to the epitopes of respective mature antibodies, which led to a hypothesis that HIV-1 may have evolved to use the "holes" (absence of or weak binding to germline-lineaged bnAbs) in the human germline B cell receptor repertoire (Xiao et al., 2009). Consistent with our earlier hypothesis, we did not find any specific binders against the HIV-1 envelope glycoproteins (Envs) but only identified binders against the SARS CoV receptor-binding domain (RBD), and soluble Hendra virus G protein (sG) when combinatorial phage display libraries mimicking human antibody repertoire constructing from human IgM libraries had been used for panning experiments (Chen et al., 2012). These findings had indicated that the major problem could be related to a high level of somatic mutations required for bnAbs to accurately target the conserved structures on the HIV-1 Envs.

In this article, we have used high-throughput 454 sequencing of a large naïve library of human IgM antibodies to explore antibody repertoire landscape for finding germline usages, somatic mutations, intermediates, and phylogenetic relationships between the intermediates and corresponding antiviral-related bnAbs including the HIV-1, SARS CoV, and henipaviruses. This study helped to identify germline predecessors of bnAbs observed in normal individuals, and find maturation pathways of antiviral bnAbs. Indeed, most of the known HIV-1 bnAbs are highly divergent from their closest respective germlines as well as their intermediates as they undergo somatic mutations required for their neutralization function. The results corroborate that the HIV-1 may use a strategy to eliminate strong binding of germline antibodies due to the absence of closer anti-HIV antibody intermediates as an escape mechanism from adaptive immune responses, and finding of closer intermediates of bnAbs from rare individuals might help designing the effective vaccines against the HIV-1 and other viral diseases.

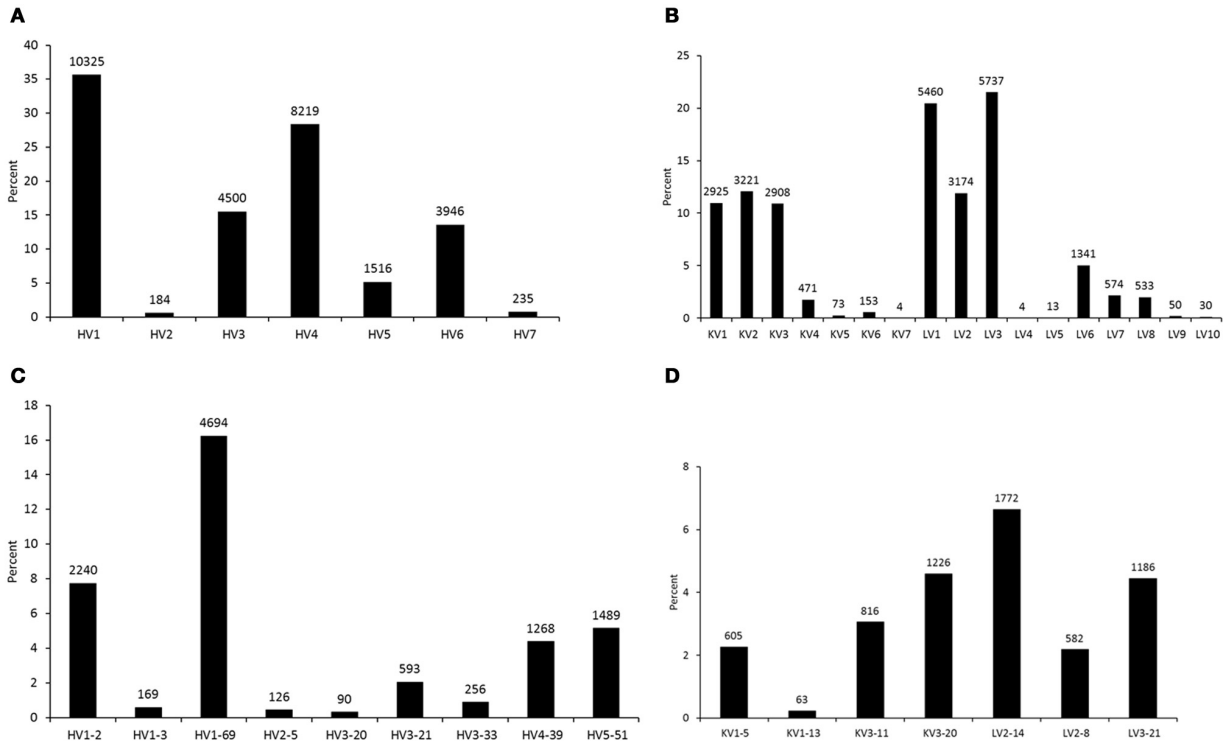


FIGURE 1 | Germline usage frequencies. The heavy (HV) and light (KV for κ and LV for λ) chains as observed in the 454 sequencing of a human naïve IgM library are shown in **(A)** and **(B)**, respectively. The frequencies of V-genes from heavy and light chains related to the antiviral bnAbs are shown in **(C)** and **(D)**, respectively. Data labels indicate the number of unique sequences at the top of data points in bar plots.

Table 1 | Antiviral bnAbs against the HIV-1, SARS CoV and Henipaviruses, and their related sequence, structure, and immunogenetics data.

bnAb	Viral Env target	PDB code	IGHV gene	IGKV/IGLV gene	Number of aa mutations (Identity/Similarity in percent)[a]		HCDR3 aa sequence (length)[b]
					V_H	V_L	
b12	gp120 CD4bs	2NY7	V1-3*01	KV3-20*01	20 (80/88)	21 (78/70)	ARVGPYSWDDSPQDNYYMDV (20)
2G12	gp120 glycans	1OP5	V3-21*01	KV1-5*03	31 (68/85)	15 (83/91)	ARKGSDRLSDNDPFDA (16)
X5	gp120 CD4i	2B4C	V1-69*01	KV3-20*01	18 (83/94)	8 (92/95)	ARDFGPDWEDGDSYDGSGRGFFDF (24)
VRC01	gp120 CD4bs	3NGB	V1-2*02	KV3-11*01	41 (58/74)	28 (68/75)	TRGKNCDYNWDFEH (14)
PG9	gp120 V1/V2	3U2S	V3-33*05	LV2-14*01	19 (81/87)	15 (85/90)	VREAGGPDYRNGYNYYDFYDGYYNYHYMDV (30)
CH01	gp120 V1/V2	3TCL	V3-20*01	KV3-20*01	28 (71/86)	16 (83/90)	ARGTDYTIDDAGIHYQGSGTFWYFDL (26)
PGT128	gp120 glycans	3TV3	V4-39*07	LV2-8*01	29 (65/81)	18 (77/87)	ARFGGEVLRYTDWPKPAWVDL (21)
2F5	gp41 MPER	1TJG	V2-5*10	KV1-13*02	14 (85/91)	14 (85/96)	AHRRGPTTLFGVPIARGPVNAMDV (24)
4E10	gp41 MPER	1TZG	V1-69*10	KV3-20*01	17 (83/95)	12 (88/93)	AREGTTGWGWLGKPIGAFAH (20)
m66	gp41 MPER	ND	V5-51*01	KV1-39*01	10 (90/96)	10 (90/94)	ARQNHYGSGSYFYRTAYYYAMDV (23)
m102	Henipa sG	ND	V1-69*10	KV3-20*01	6 (94/99)	9 (91/96)	ARGWGREQLAPHPSQYYYYYYGMDV (25)
m396	SARS RBD	2DD8	V1-69*05	LV3-21*03	5 (95/95)	2 (98/99)	ARDTVMGGMDV (11)

[a] Number of heavy chain (V_H) aa mutations were determined by IMGT/V-QUEST and confined to the V region only (excluding HCDR3 and Framework 4); Identity and similarity between aa sequences of bnAb and its germline counterpart were based on pairwise alignment using the Needleman-Wunsch algorithm.

[b] HCDR3, heavy chain complementarity determining region 3, lengths follow the CDR-IMGT definition.

bnAb, broadly neutralizing antibody; CD4bs, CD4 binding site; CD4i, CD4-induced; V1/V2, variable loops V1 and V2; MPER, membrane proximal epitope region; sG, soluble G glycoprotein; RBD, receptor binding domain; PDB, Protein Data Bank; ND, not determined; IGHV, IGKV and IGLV genes are V-REGIONS from V_H, V-KAPPA and V-LAMBDA domains respectively; aa, amino acids.

MATERIALS AND METHODS

PCR AMPLIFICATION AND HIGH-THROUGHPUT 454 SEQUENCING

To amplify IgM antibody sequences, cDNA was prepared from peripheral blood B cells of 10 healthy donors as received under the Research Donor Program of Frederick National Laboratory for Cancer Research, USA, which we previously used to construct a naïve human Fab phage display library for selecting antibodies against SARS CoV and henipaviruses. The complete set of primers used in the PCR amplification of IgM-derived heavy and light chains were described in detail elsewhere (Zhu and Dimitrov, 2009). For 454 sequencing, primer combinations used to amplify cDNA in separate reactions included the Roche A and B adaptor sequences along with target amplification sequence for heavy and light chain variable domains. The gene fragments were amplified in 20 cycles of PCR using the High Fidelity PCR Master from Roche. More detailed description of 454 sequencing can be found in our recent articles (Prabakaran et al., 2011, 2012). The standard Roche 454 GS Titanium shotgun library protocol was adapted as found in the Roche sequencing technical bulletin.

DATABASES AND TOOLS

For quality control of antibody sequences, we trimmed the 454 sequence data and retained only sequences of length more than 300 nucleotides (nt), covering the entire antibody variable domains consisting of the three complementarity determining regions (CDR) along with framework regions (FR). We used IMGT/HighV-QUEST (Alamyar et al., 2012), a high-throughput version for deep sequencing NGS data analysis resource for the immunogenetic analysis. The output results from the IMGT/HighV-QUEST analysis in CSV files were stored at PostgreSQL database, and Structured Query Language (SQL) was used to retrieve the data for the further analysis. Heatmap generation and statistical calculations involving distributions of antibody HCDR3 lengths and mutations were carried out using SAS JMP10® statistical software (SAS Institute, Cary, NC).

COMPUTATIONAL ANALYSES OF ANTIBODY SEQUENCES

Translated heavy and light chain variable sequences from the 454 sequencing that shared the IGHV genes of selected antiviral antibodies and associated immunogenetics data including the details of germlines, HCDR3 lengths, and mutations were retrieved from the database by using SQL. Sequence identities between the 454 sequence data and germlines were calculated based on the pairwise alignment using local BLAST as implemented in BioEdit v7.0.9 (Hall, 1999). Phylogenetic analysis was carried out using the Archaeopteryx software (Han and Zmasek, 2009).

RESULTS

GERMLINE GENE USAGES OF ANTIVIRAL bnAbs

To analyze germline origin of antiviral antibodies against the HIV-1, SARS CoV, and henipaviruses as expressed in the human IgM repertoire, we performed 454 sequencing of a non-immune library which was previously constructed from peripheral blood B cells of 10 healthy donors and used to select antibodies against SARS CoV and henipaviruses (Prabakaran et al., 2006; Zhu et al., 2006). A total of 113,139 sequences were obtained from which 91,528 sequences were found as unique with each had >300 nt in length. The total number of unique amino acid (aa) sequences for each V-gene subgroup in heavy and light chains that were found functionally productive as determined by IMGT/HighV-QUEST (Alamyar et al., 2012) are shown in **Figures 1A,B,** respectively. The read coverage or gene frequencies observed in the study suggested for biased germline usages and were comparable to the previous studies (Glanville et al., 2009; Prabakaran et al., 2012) but way far less than the theoretical diversity attainable by antibodies attributing to several factors such as library sampling, primer efficiency, and sequencing errors and limitations. Nevertheless, we selected known bnAbs against the viral targets including the HIV-1, SARS CoV, and henipaviruses (**Table 1**), and created sequence data sets related to those bnAbs from the 454 analysis as depicted in **Figures 1C,D** showing the germline usage frequencies of IGHV genes in the V_H domains, IGKV,

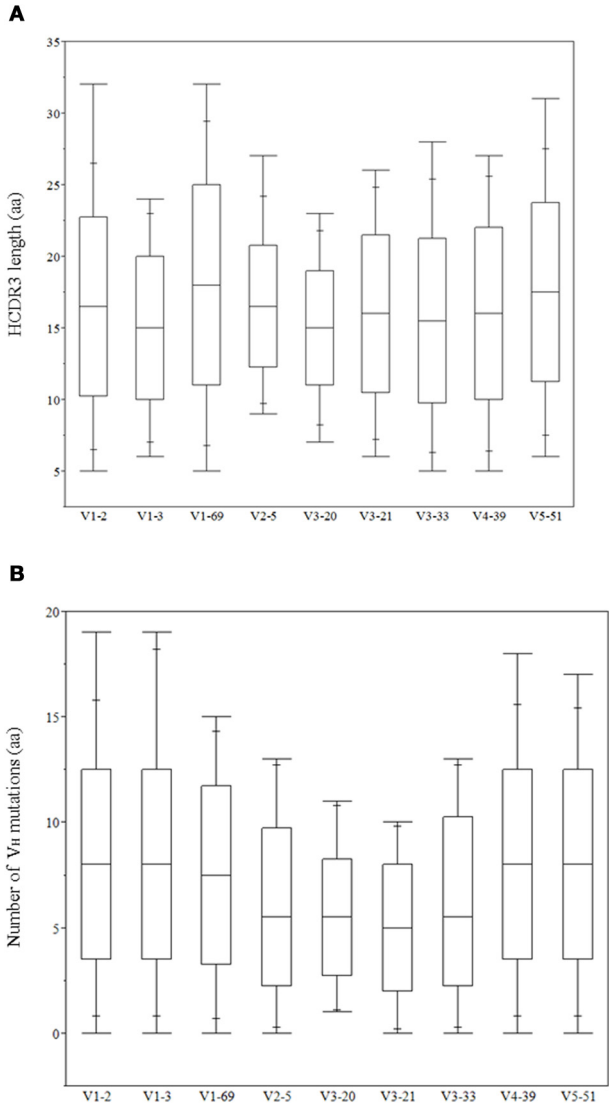


FIGURE 2 | Distribution of heavy chain complementarity determining region (HCDR3) lengths in (aa) (A) and V_H mutations (aa) (B) of IGHV genes related to the antiviral bnAbs.

and IGLV genes in the V_κ and V_λ domains, respectively. We found that while all antiviral-related germlines were expressed in human IgM repertoire, some preferential germline usages were noted, for example, HV1-69 gene in IGHV subgroups and KV3-20/LV2-14 genes in IGKV/IGLV subgroups were overrepresented (**Figures 1C,D**).

HCDR3 LENGTH DISTRIBUTIONS, SOMATIC V_H MUTATIONS AND UNIQUE VDJ FREQUENCIES

The role of heavy chains of antiviral antibodies in antigen recognition is found to be associated with longer HCDR3s and extensive V_H mutations (**Table 1**). Most of the bnAbs have longer HCDR3s with aa lengths ranging from 20 to 30, except for 2G12, VRC01 and m396. All of the V_H genes of anti-HIV-1 antibodies have a high degree of somatic mutations when compared to non-HIV-1 antiviral bnAbs. We analyzed HCDR3 length distributions and V_H mutations preexisting in germline-lineaged precursor antiviral antibodies from the IGHV genes of IgM repertoires from which bnAbs were generated. The box plots display the distributions of HCDR3 lengths and V_H mutations, **Figures 2A,B,** respectively, which indicates a high level HCDR3 length diversity and lesser extent of somatic mutations compared to bnAbs (**Table 1**).

To assess the VDJ repertoire usage among different antiviral related IGHV genes, we computed the frequencies of VDJ recombination patterns as observed in the V_H genes expressed in human IgM repertoire involving those IGHV genes of antiviral antibodies. The heatmap is shown in the **Figure 3** depicting the

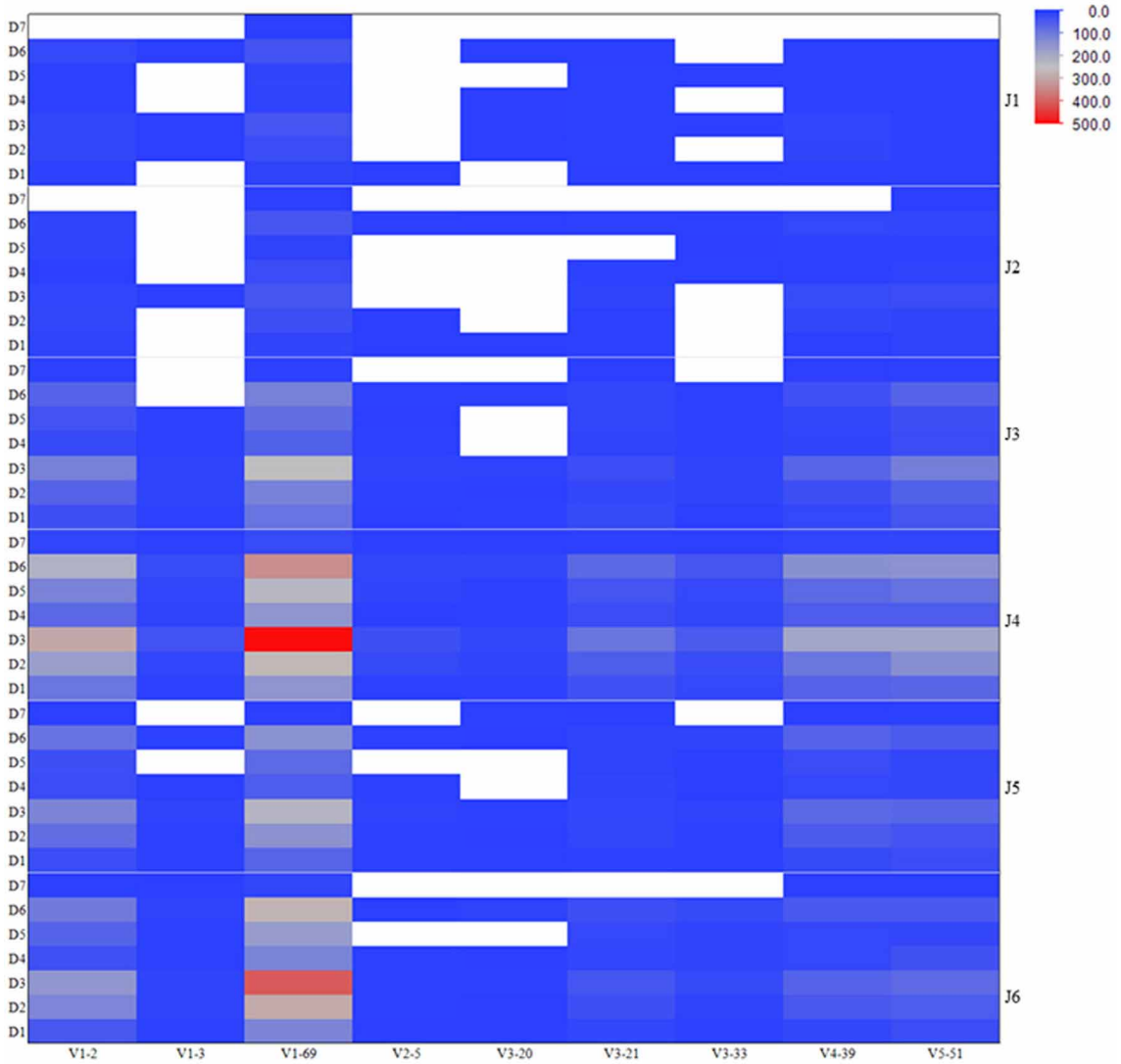


FIGURE 3 | Frequencies of VDJ recombination types as observed in the human IgM repertoire involving IGHV genes related to the antiviral bnAbs. The heatmap is colored according to the total number of unique VDJ patterns existing in the corresponding IGHV genes used in association with different IGHD and IGHJ genes, and is shown on a blue-to-gray-to-red scale. The white-colored space represents the missed or absent VDJ recombination types in the repertoire.

most (red) and least (blue) abundant VDJ types existing in the germline-lineaged repertoire for the corresponding IGHV genes used in association with different IGHD and IGHJ genes. The IGHV genes V1-69 and V1-2 were frequently found to recombine with IGHJ genes J4 and J6, and IGHD genes D3 and D6.

IDENTIFICATION OF INTERMEDIATE ANTIVIRAL bnAbs AND GERMLINE-LINAGE ANALYSIS

The intermediate antibodies corresponding to bnAbs against the HIV-1, SARS CoV, and henipaviruses were found by analyzing the human IgM repertoire, and such intermediates with the closest similarities to the matured antiviral bnAbs were selected for germline-linage analysis by using phylogenetic method. IGHV germline gene alleles of bnAbs were obtained from the IMGT database. The mid-point phylogenetic neighbor-joining tree showing the evolutionary relationships of different antiviral antibodies with their corresponding germlines and intermediates is given in **Figure 4**. We observed that some of the anti-HIV-1 antibodies (2G12, CH01, and VRC01) were found at distal nodes in the phylogenetic tree indicating high divergence from their corresponding germline and intermediate counterparts. In contrast, bnAbs against SARS CoV, and henipaviruses, m396 and m102, were found closer to their intermediates.

ANALYSIS OF INTERMEDIATES OF ANTI-HIV-1 bnAb b12 AND MAPPING OF SOMATIC V_H MUTATIONS TO THE COMPLEX STRUCTURE

We found 169 unique IGHV sequences from the V1-3 gene family as intermediates of bnAb b12 by using the 454 sequence analysis of a human IgM library. Phylogenetic analysis of those intermediates revealed two major groups, one group consisting of germline related antibodies and the other having potential intermediates closer to the bnAb b12. We then constructed a phylogenetic sub-tree selecting only the potential intermediates and the V1-3*01 germline along with bnAb 12. The tree was rooted at the known germline V1-3*01 of bnAb b12, and phylogram showed evolutionary relationship among the different intermediates (**Figure 5A**). One of the intermediates, G3JY1, had the maximum of 72% sequence identity (82% sequence similarity) at aa level to the bnAb b12 (**Figure 5B**). However, the HCDR3 length of that intermediate was found to be 17 aa long, which is 3 aa shorter than that of b12 antibody. To find the closest HCDR3 to that of b12, we scanned 28,925 unique HCDR3 sequences from the entire IgM 454 sequence data. We identified a HCDR3 with the same length (20 aa) and 50% sequence identity to that of b12 (**Figure 5C**), which was found to be the most similar to the HCDR3 of b12 but the IGHV gene associated with that HCDR3 was found to be V4-b. We used the HIV-1 gp120-b12 complex structure and mapped the V_H somatic mutations, which showed the overlapping of three mutated residues of b12 (N36 from HCDR1, Y59 from HCDR2, and W111.1 from HCDR3) that contribute to the most of binding interactions with the gp120 as previously observed (Zhou et al., 2007) (**Figure 5D**).

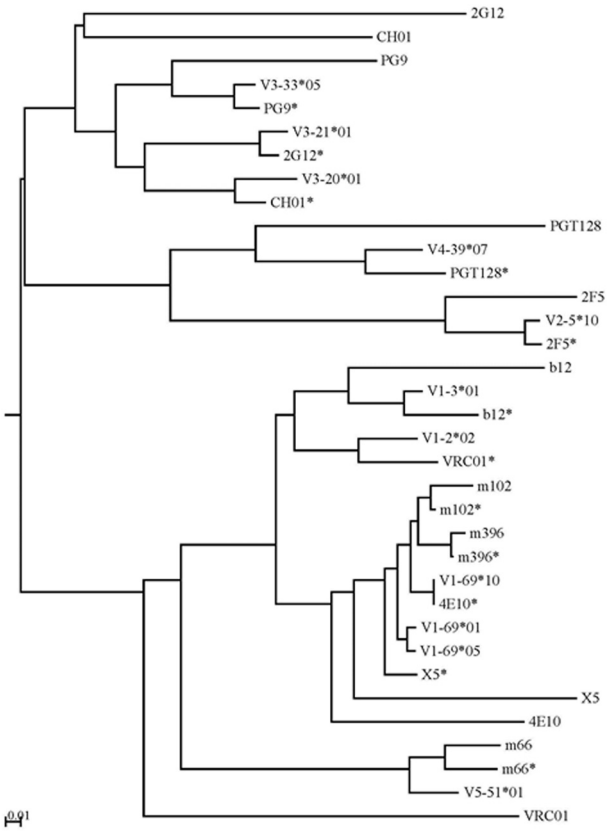


FIGURE 4 | The mid-point phylogenetic neighbor-joining tree shows the evolutionary relationships between different IGHVs of (bnAbs) with their corresponding germlines and intermediates. The IGHV germline gene alleles follow the IMGT nomenclature and the closest intermediates of bnAbs as found from the human IgM repertoire were designated with asterisks along with names of bnAbs. Some of the anti-HIV-1 antibodies (2G12, CH01, and VRC01) were found at distal nodes in the phylogenetic tree indicating high divergence from their corresponding germline and intermediate counterparts.

DISCUSSION

In this study, we have described the 454 sequence analysis of a large naïve library of human IgM antibodies, and carried out immunogenetic analysis to study the origin, diversity, and maturation of selected known bnAbs against the HIV-1, SARS CoV RBD, and henipaviruses sG proteins. We have found intermediates of antiviral related bnAbs, of which most of those against the HIV-1 were highly diverged from their mature forms of bnAbs as compared to other viral targets, SARS CoV, and henipaviruses.

Although antibodies are generated through various mechanisms involving VDJ recombination, junctional modification, and hypermutations, the V-genes sculpt the most of the antigen-combining sites, CDR1 and CDR2, and support frameworks for the CDR3. We found that antiviral antibodies targeting different Env binding regions of the HIV-1 and other viruses utilized different germline V-genes as the origins (**Table 1**). We noted that, among antiviral-related bnAbs, the V1-69 gene usage was dominated in the heavy chains while V3-20 and V2-14 genes of kappa and lambda were used with the highest frequencies in the light chains of human IgM repertoire (**Figure 1**). Accordingly, four of the V_H genes of bnAbs (4E10, X5, m102, and m396) originated

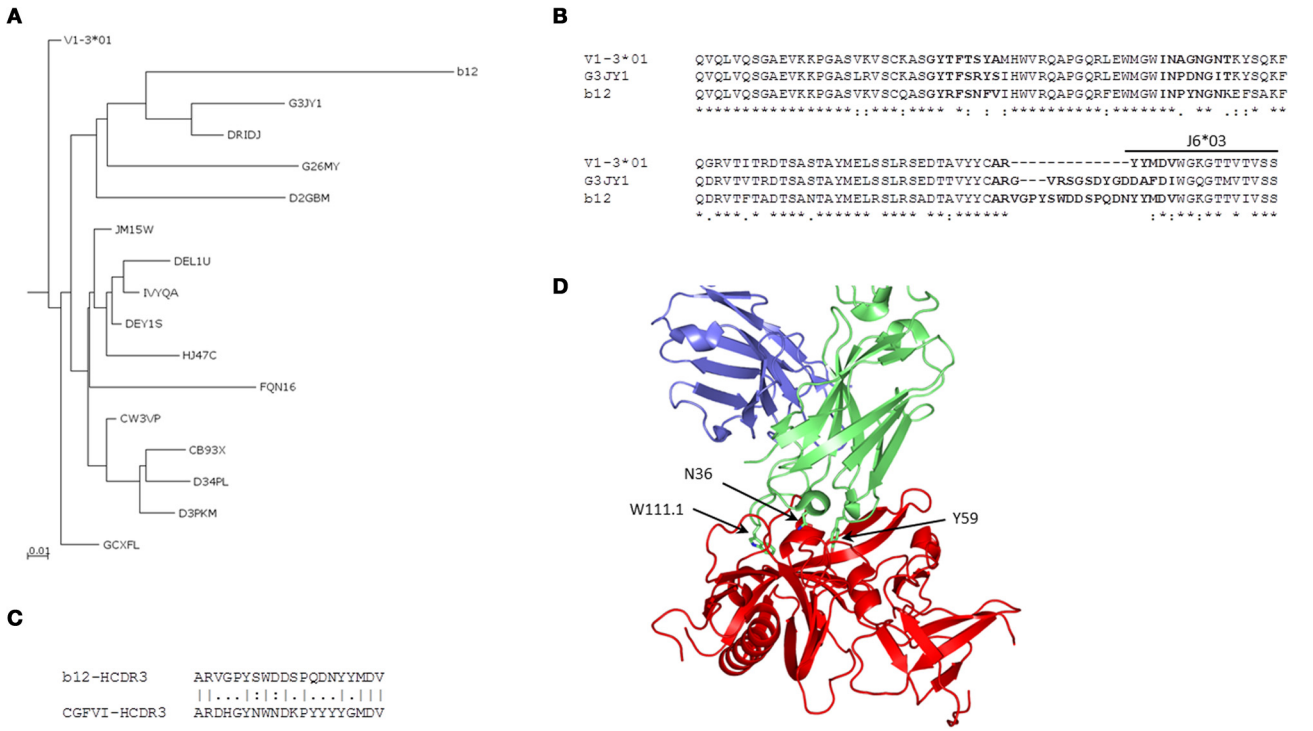


FIGURE 5 | Phylogenetic and somatic mutation analyses of anti-HIV bnAb b12. (A) A phylogenetic sub-tree constructed from a total of 169 unique sequences from the HV1-3 subgroup, which were obtained from the 454 sequence analysis of a human IgM library, along with the heavy chain V1-3*01 germline and V_H sequence of b12 antibody showing the evolutionary relationship among different intermediates. Five-letter alphanumeric codes denote sequence labels as used in the pyrosequencing. **(B)** Multiple sequence alignment shows the differences in aa sequence between germline, intermediate, G3JY1, and b12 neutralizing antibodies. The locations of CDRs according to IMGT numbering are shown in bold face. **(C)** The most similar HCDR3 sequence to that of bnAb b12 from the IgM repertoire was found to originate from V4-b*01 gene and is shown in the pairwise alignment. **(D)** Mapping of three of the somatically mutated residues N36, Y59, and W111.1, as per IMGT numbering scheme, from each of the HCDRs are shown as sticks using the complex crystal structure of HIV-1 gp120-b12 (PDB code 2NY7).

from the V1-69, and three of them paired with the kappa V3-20 gene. One possible reason for dominance in the usage of those germline genes could be reflecting from the relatively higher frequencies of distributions observed in the expressed IgM repertoire (**Figures 1A,B**). The HV3 gene was used in the three of the HIV-1 bnAbs, 2G12, PG9, and CH01. The structural data for most of the bnAbs selected in this analysis were known and the heavy chains of these bnAbs were dominantly used. The increased number of V_H mutations and longer HCDR3s are characteristics for the HVI-1 bnAbs when compared to other antiviral bnAbs (Breden et al., 2011). We analyzed the distribution of HCDR3 lengths and extent of somatic V_H mutations in the human IgM repertoire to compare with that of antiviral-related bnAbs (**Figure 2**). The results showed that the longer HCDR3s and low level of somatic V_H mutations as compared to the HIV-1 bnAbs existed in the intermediates as found from the 454 sequencing. The somatic diversity through VDJ recombination involving antiviral-related V-genes in the IgM repertoire was found high; the most abundant VDJ combination consisted of the HV1-69 gene with certain D and J genes as depicted in gray and red (**Figure 3**), which might be the reason for the preferential usage of that HV1-69 in many other viral diseases (Sui et al., 2009).

Further, bnAbs against the SARS CoV and henipaviruses shared the heavy chain V-gene germline, HV1-69, with two of the HIV-1 bnAbs, 4E10, and X5. All of these four bnAbs were less divergent from their V-germlines and intermediates, when compared to other HIV-1 bnAbs, and formed a single cluster at a mid-point rooted phylogenetic tree (**Figure 4**). The gp41 membrane-proximal epitope region (MPER) binding site bnAbs, 2F5, and m66, were moderately divergent from their V-germlines and intermediates and formed distinct clusters. The V-gene of VRC01 bnAb was the most divergent from its respective germline as well as the closest intermediate, and was placed at a distal branch of HV1 subgroup of bnAbs. For the mid-point rooted phylogenetic analysis, we included the closest intermediates only; however, favored maturation pathways could involve other intermediates too. We created the germline-rooted phylogenetic tree as a use-case for the bnAb b12 (**Figure 5A**) and analyzed the maturation pathway along different V-gene intermediates from HV1-3 gene family. The closest b12 intermediate, designated as G3JY1, had three mutations each at HCDR1 and HCDR2 compared to the germline, and were found similar though not identical to that of mature b12 (**Figure 5B**). Interestingly, we also identified a HCDR3 with the same length (20 aa) and 50% sequence identity to that of b12 (**Figure 5C**),

which was found to be the most similar to the HCDR3 of b12 but the IGHV gene associated with that HCDR3 was found to be V4-b. This might suggest for the possible maturation mechanism of bnAbs which could be involving the VH replacement (Chen et al., 1995). These two mutated residues (N36 from HCDR1 and Y59 from HCDR2) from the V-gene and a Trp residue from the D-gene (W111.1 from HCDR3) contributed to the most of binding interactions with the gp120 (**Figure 5D**) (Zhou et al., 2007).

In summary, the 454 sequence analysis of a large naïve human antibody repertoire corresponding to the selected antiviral-related bnAbs revealed the germline V-gene usage, VDJ rearrangement, HCDR3 length diversity, and somatic mutations of potential intermediate antibodies of HIV-1 and other viruses such as SARS CoV and henipaviruses. Thus, B cell germline-lineage analysis using the 454 sequence data from different sources could help finding appropriate antibody intermediates, pathways, and mechanisms useful in the development of bnAbs and vaccines against the HIV-1 and other viral diseases.

ACKNOWLEDGMENTS

We thank the Laboratory of Molecular Technology of SAIC-Frederick Inc. for providing Roche 454 sequencing service. We are grateful to Eltaf Alamyar and to the IMGT® team for providing access to IMGT/HighV-QUEST. We thank Ms. Maria G. Singarayan for constructing the PostgreSQL database and JAVA applications and helping with SQL. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, and by Federal funds from the NIH, National Cancer Institute, under Contract No. NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US Government.

REFERENCES

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., and Lefranc, M.-P. (2012). IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Res.* 8, 1–15.

Breden, F., Lepik, C., Longo, N. S., Montero, M., Lipsky, P. E., and Scott, J. K. (2011). Comparison of antibody repertoires produced by HIV-1 infection, other chronic and acute infections, and systemic autoimmune disease. *PloS ONE* 6:e16857. doi: 10.1371/journal.pone.0016857

Burton, D. R. (2002). Antibodies, viruses and vaccines. *Nat. Rev. Immunol.* 2, 706–713.

Chen, C., Nagy, Z., Prak, E. L., and Weigert, M. (1995). Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing. *Immunity* 3, 747–755.

Chen, W., Streaker, E. D., Russ, D. E., Feng, Y., Prabakaran, P., and Dimitrov, D. S. (2012). Characterization of germline antibody libraries from human umbilical cord blood and selection of monoclonal antibodies to viral envelope glycoproteins: implications for mechanisms of immune evasion and design of vaccine immunogens. *Biochem. Biophys. Res. Commun.* 417, 1164–1169.

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M., Cox, D., Rajpal, A., and Pons, J. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20216–20221.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* 41, 95–98.

Han, M. V., and Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10, 356.

Haynes, B. F., and Montefiori, D. C. (2006). Aiming to induce broadly reactive neutralizing antibody responses with HIV-1 vaccine candidates. *Expert. Rev. Vaccines* 5, 579–595.

Johnson, W. E., and Desrosiers, R. C. (2002). Viral persistance: HIV's strategies of immune system evasion. *Annu. Rev. Med.* 53, 499–518.

Prabakaran, P., Chen, W., Singarayan, M. G., Stewart, C. C., Streaker, E., Feng, Y., and Dimitrov, D. S. (2012). Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64, 337–350.

Prabakaran, P., Dimitrov, A. S., Fouts, T. R., and Dimitrov, D. S. (2007). Structure and function of the HIV envelope glycoprotein as entry mediator, vaccine immunogen, and target for inhibitors. *Adv. Pharmacol.* 55, 33–97.

Prabakaran, P., Gan, J., Feng, Y., Zhu, Z., Choudhry, V., Xiao, X., Ji, X., and Dimitrov, D. S. (2006). Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. *J. Biol. Chem.* 281, 15829–15836.

Prabakaran, P., Streaker, E., Chen, W., and Dimitrov, D. S. (2011). 454 antibody sequencing – error characterization and correction. *BMC Res. Notes* 4, 1–7.

Sui, J., Hwang, W. C., Perez, S., Wei, G., Aird, D., Chen, L. M., Santelli, E., Stec, B., Cadwell, G., Ali, M., Wan, H., Murakami, A., Yammanuru, A., Han, T., Cox, N. J., Bankston, L. A., Donis, R. O., Liddington, R. C., and Marasco, W. A. (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nat. Struct. Mol. Biol.* 16, 265–273.

Xiao, X., Chen, W., Feng, Y., Zhu, Z., Prabakaran, P., Wang, Y., Zhang, M. Y., Longo, N. S., and Dimitrov, D. S. (2009). Germline-like predecessors of broadly neutralizing antibodies lack measurable binding to HIV-1 envelope glycoproteins: implications for evasion of immune responses and design of vaccine immunogens. *Biochem. Biophys. Res. Commun.* 390, 404–409.

Zhou, T., Xu, L., Dey, B., Hessell, A. J., Van Ryk, D., Xiang, S. H., Yang, X., Zhang, M. Y., Zwick, M. B., Arthos, J., Burton, D. R., Dimitrov, D. S., Sodroski, J., Wyatt, R., Nabel, G. J., and Kwong, P. D. (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* 445, 732–737.

Zhu, Z., Dimitrov, A. S., Bossart, K. N., Crameri, G., Bishop, K. A., Choudhry, V., Mungall, B. A., Feng, Y. R., Choudhary, A., Zhang, M. Y., Feng, Y., Wang, L. F., Xiao, X., Eaton, B. T., Broder, C. C., and Dimitrov, D. S. (2006). Potent neutralization of Hendra and Nipah viruses by human monoclonal antibodies. *J. Virol.* 80, 891–899.

Zhu, Z., and Dimitrov, D. S. (2009). Construction of a large naive human phage-displayed Fab library through one-step cloning. *Methods Mol. Biol.* 525, 129–142. xv.

Conflict of Interest Statement: 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.

Received: 30 May 2012; paper pending published: 22 June 2012; accepted: 17 July 2012; published online: 02 August 2012.

*Citation: Prabakaran P, Zhu Z, Chen W, Gong R, Feng Y, Streaker E and Dimitrov DS (2012) Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. Front. Microbio. **3**:277. doi: 10.3389/fmicb.2012.00277*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 11 September 2012
doi: 10.3389/fmicb.2012.00315

Somatic populations of PGT135–137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics

Jiang Zhu[1], Sijy O'Dell[1†], Gilad Ofek[1†], Marie Pancera[1†], Xueling Wu[1†], Baoshan Zhang[1†], Zhenhai Zhang[2†], NISC Comparative Sequencing Program[3], James C. Mullikin[3], Melissa Simek[4], Dennis R. Burton[5,6,7], Wayne C. Koff[4], Lawrence Shapiro[1,2], John R. Mascola[1] and Peter D. Kwong[1]*

[1] *Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA*
[2] *Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA*
[3] *NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA*
[4] *International AIDS Vaccine Initiative, New York, NY, USA*
[5] *Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA, USA*
[6] *International AIDS Vaccine Initiative Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, CA, USA*
[7] *Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Takamasa Takeuchi, National Institute of Infectious Diseases, Japan
Dimiter Dimitrov, National Institutes of Health, USA

***Correspondence:**
Peter D. Kwong, Vaccine Research Center, NIAID/NIH, 40 Convent Drive, Building 40, Room 4508, Bethesda, MD 20892, USA.
e-mail: pdkwong@nih.gov

[†]Sijy O'Dell, Gilad Ofek, Marie Pancera, Xueling Wu, Baoshan Zhang and Zhenhai Zhang have contributed equally to this work.

Select HIV-1-infected individuals develop sera capable of neutralizing diverse viral strains. The molecular basis of this neutralization is currently being deciphered by the isolation of HIV-1-neutralizing antibodies. In one infected donor, three neutralizing antibodies, PGT135–137, were identified by assessment of neutralization from individually sorted B cells and found to recognize an epitope containing an N-linked glycan at residue 332 on HIV-1 gp120. Here we use next-generation sequencing and bioinformatics methods to interrogate the B cell record of this donor to gain a more complete understanding of the humoral immune response. PGT135–137-gene family specific primers were used to amplify heavy-chain and light-chain variable-domain sequences. Pyrosequencing produced 141,298 heavy-chain sequences of IGHV4-39 origin and 87,229 light-chain sequences of IGKV3-15 origin. A number of heavy and light-chain sequences of ∼90% identity to PGT137, several to PGT136, and none of high identity to PGT135 were identified. After expansion of these sequences to include close phylogenetic relatives, a total of 202 heavy-chain sequences and 72 light-chain sequences were identified. These sequences were clustered into populations of 95% identity comprising 15 for heavy chain and 10 for light chain, and a select sequence from each population was synthesized and reconstituted with a PGT137-partner chain. Reconstituted antibodies showed varied neutralization phenotypes for HIV-1 clade A and D isolates. Sequence diversity of the antibody population represented by these tested sequences was notably higher than observed with a 454 pyrosequencing-control analysis on 10 antibodies of defined sequence, suggesting that this diversity results primarily from somatic maturation. Our results thus provide an example of how pathogens like HIV-1 are opposed by a varied humoral immune response, derived from intrinsic mechanisms of antibody development, and embodied by somatic populations of diverse antibodies.

Keywords: antibody bioinformatics, high-throughput sequencing, HIV-1, immunity, N-linked glycan

INTRODUCTION

Recent years have seen revolutions in both genomics and computational science (Lander et al., 2001; Venter et al., 2001; Chen et al., 2012). In both of these fields, capabilities are advancing exponentially (Kahn, 2011). The impact of this non-linear development on biology is pervasive and multifaceted. With respect to virus research, the influence has been profound and is the focus of this special issue of *Frontiers*. Medical interest in viruses is focused on pathogens and their infection, and the biological mirror of infection is the host immune response. Advances in genomics and computational science have the potential for an equally profound impact on our understanding of the immune response. Here we focus on the application of new genomic and computational techniques, particularly 454 pyrosequencing of B cell transcripts (Reddy et al., 2009; Reddy and Georgiou, 2011; Wu et al., 2011) and systems-level bioinformatics (Kitano, 2002), to understand the antibody response to infection.

The human immunodeficiency virus type I, HIV-1, is the etiological agent of a global pandemic, which has killed over 30 million people, and currently infects ∼1% of adults worldwide (UNAIDS, 2010). HIV-1 is a retrovirus and member of the lentivirus genus (Gonda et al., 1985; Sonigo et al., 1985). Global genetic diversity of HIV-1 is extraordinarily high (Starcich et al., 1986; Korber et al., 2001), and this is thought to result from the low fidelity of its genome replication (Preston et al., 1988) as well as the persistent nature of the infection: the diversity of HIV-1 virus within a single

individual after 6 years of infection is equivalent to the global diversity of H1N1 influenza observed annually (Korber et al., 2001). Infection by HIV-1 elicits many antibodies, but in general these are not capable of neutralization of diverse strains of HIV-1. However, after several years of infection, 10–25% of infected individuals develop broadly neutralizing antibodies (Li et al., 2007; Gray et al., 2009; Sather et al., 2009; Simek et al., 2009; Stamatatos et al., 2009; Doria-Rose et al., 2010; Gnanakaran et al., 2010). These antibodies provide little or no benefit to the infected host, as the evolution of the virus outpaces the immune response (Parren et al., 1999; Poignard et al., 1999; Wei et al., 2003). Nevertheless these antibodies, when tested in humanized mice or macaque models by passive antibody transfer, impart effective immunity to challenge with HIV-1 or simian/human chimeric immunodeficiency viruses (Mascola et al., 1999, 2000; Parren et al., 2001; Mascola, 2003; Veazey et al., 2003; Hessell et al., 2009a,b; Balazs et al., 2011), indicating the potential for their use as targets for re-elicitation by rationally designed vaccines (reviewed in Walker and Burton, 2010; Kwong et al., 2011). Thus, substantial interest has focused on understanding human antibodies that effectively neutralize diverse strains of HIV-1.

A number of techniques have recently been applied to identification of such antibodies. These methods – including antigen-specific B cell sorting (Scheid et al., 2009; Wu et al., 2010) and direct assessment of neutralization by antibodies secreted from individually sorted B cells (Walker et al., 2009, 2011), each coupled to single B cell sequencing techniques – have so far yielded dozens of broadly HIV-1-neutralizing antibodies. These antibodies represent an extraordinarily sparse sampling of the humoral immune response, which typically generates roughly a billion new B cells in a healthy individual each day. We therefore asked whether the revolutionary new capabilities of next-generation sequencing (Mardis, 2008a,b; Boyd et al., 2010; Hawkins et al., 2010) and computational science could expand this sampling to generate a more complete understanding of the humoral immune response. In principle, memory B cells contain a persistent record of the antibody response to infection. As memory B cells are readily attained from blood, they provide a convenient means to access the antibody record, with B cell transcripts in peripheral blood mononuclear cells (PBMCs) providing a genetic representation. Using three antibodies, PGT135–137 from Protocol G donor 39 (Walker et al., 2011) as an example, we used 454 pyrosequencing of PCR-amplified heavy- and light-chain transcripts to capture a more comprehensive genetic record. We used bioinformatics approaches to interrogate this record, to identify populations of neutralizing antibodies, and to characterize their ontogenies. We link these ontogenies to the natural mechanisms of B cell development to provide a view of how somatic populations of antibodies engender a diverse immunological response to infection.

MATERIALS AND METHODS

HUMAN SPECIMENS

The PBMCs of the HIV-1 infected donor 39 were obtained from the International AIDS Vaccine Initiative (IAVI) protocol G. The same sample was used to isolate broadly neutralizing antibodies PGT135–137 (Walker et al., 2011). Human peripheral blood samples were collected after obtaining informed consent and appropriate Institutional Review Board (IRB) approval.

SAMPLE PREPARATION FOR 454 PYROSEQUENCING

Ten previously described heavy-chain plasmids with known sequences (Wu et al., 2011) were selected to assess 454 pyrosequencing error. Ten plasmids (100 ng each) were combined in 35 μl water, and 1 μl of the ten-plasmid combination was used to template polymerase chain reactions (PCRs). The heavy and kappa chain PCR samples for 454 pyrosequencing from donor 39 were prepared as described (Wu et al., 2011) with minor modifications. Briefly, mRNA was extracted from 20 million PBMCs into 200 μl of elution buffer (Oligotex kit, Qiagen), then concentrated to 10–30 μl by centrifuging the buffer through a 30 kD micron filter (Millipore). The reverse transcription was performed in one or multiple 35 μl-reactions, each composed of 13 μl of mRNA, 3 μl of oligo$(dT)_{12–18}$ at 0.5 μg/μl (Invitrogen), 7 μl of 5× first strand buffer (Invitrogen), 3 μl of RNase Out (Invitrogen), 3 μl of 0.1 M DTT (Invitrogen), 3 μl of dNTP mix (each at 10 mM), and 3 μl of SuperScript II (Invitrogen). The reactions were incubated at 42°C for 2 h. The cDNAs from each reaction were combined, applied to the NucleoSpin Extract II kit (Clontech), and eluted in 20 μl of elution buffer. In this way, 1 μl of the cDNA comprised transcripts from 1 million PBMCs. The immunoglobulin gene family-specific PCR was set up in a total volume of 50 μl, using 1 μl of the heavy-chain plasmid mix or 5 μl of the cDNA as template (equivalent of transcripts from 5 million PBMCs). The DNA polymerase systems used was the Platinum *Taq* High-Fidelity (HiFi) DNA Polymerase System (Invitrogen). According to the instructions of the manufacturer, the reaction mix was composed of water, 5 μl of 10× buffer, and 1 μl of supplied MgSO4, 2 μl of dNTP mix (each at 10 mM), 1–2 μl of primers (Table S1 in Supplementary Material) at 25 μM, and 1 μl of Platinum Taq HiFi DNA polymerase. The primers each contained the appropriate adaptor sequences (XLR-A or XLR-B) for subsequent 454 pyrosequencing. The PCRs were initiated at 95°C for 30 s, followed by 25 cycles of 95°C for 30 s, 58°C for 30 s, and 72°C for 1 min, then incubated at 72°C for 10 min. The PCR products at the expected size (∼500 bp) were gel extracted and purified (Qiagen), followed by further phenol/chloroform purification.

454 PYROSEQUENCING AND LIBRARY PREPARATION

The 454 pyrosequencing was carried out as described previously (Wu et al., 2011). Briefly, PCR products were quantified using Qubit (Life Technologies, Carlsbad, CA, USA). Library concentrations were determined using the KAPA Biosystems qPCR system (Woburn, MA, USA) with 454 pyrosequencing standards provided in the KAPA system. Pyrosequencing of the PCR products was performed on a GS FLX sequencing instrument (Roche-454 Life Sciences, Bradford, CT, USA) using the manufacturer's suggested methods and reagents. Initial image collection was performed on the GS FLX instrument and subsequent signal processing, quality filtering, and generation of nucleotide sequence and quality scores were performed on an off-instrument linux cluster using 454 application software (version 2.5.3). The amplicon quality filtering parameters were adjusted based on the manufacturer's recommendations (Roche-454 Life Sciences Application Brief No.

001-2010). Quality scores were assigned to each nucleotide using methodologies incorporated into the 454 application software to convert flowgram intensity values to Phred-based quality scores and as described (Brockman et al., 2008). The quality of each run was assessed by analysis of internal control sequences included in the 454 pyrosequencing reagents. Reports were generated for each region of the PicoTiterPlate (PTP) for both the internal controls and the samples.

BIOINFORMATICS ANALYSIS OF 454 PYROSEQUENCING-DETERMINED ANTIBODY SEQUENCES

Our previously described bioinformatics pipeline (Wu et al., 2011) was refined and currently consists of five steps. Starting from a 454 pyrosequencing-determined antibodyome, each sequence read was (1) reformatted and labeled with a unique index number; (2) assigned to variable (V), diverse (D), and joining (J) gene families and alleles using an in-house implementation of IgBLAST[1], and sequences with E-value $> 10^{-3}$ for V gene assignment were rejected; (3) subjected to a template-based error-correction procedure, in which 454 pyrosequencing homopolymer errors in V, D, and J regions were detected based on the alignment to their respective germline sequences. Note that only insertion and deletion errors of less than three nucleotides were corrected. D and J gene were corrected only when their gene assignment was reliable, indicated by E-value $< 10^{-3}$; (4) compared with the a set of template antibody sequences at both nucleotide level and amino-acid level using a global alignment module in CLUSTALW2 (Larkin et al., 2007); (5) subjected to a multiple sequence alignment (MSA)-based scheme to determine the third complementarity-determining region (CDR H3 or L3), which was further compared with a set of template CDR H3 or L3 sequences at nucleotide level, and to determine the sequence boundary of variable domain. For a large population of highly similar sequences, a "divide-and-conquer" procedure could be used to derive a consensus sequence to represent the population and to reduce random sequencing errors. First, a clustering using BLASTClust (Altschul et al., 1997) with a 95% sequence identity cutoff is performed on the sequence population. Then, the largest cluster is divided into 10–50 sets, for each of which a consensus can be derived from MSA. A final consensus is obtained by averaging over the subset consensuses.

Intra-donor phylogenetic analysis use the same procedure as cross-donor phylogenetic analysis, which has been described in detail in previous study (Wu et al., 2011), except that the template antibodies are from the same donor (intra-donor) rather than added exogenously (cross-donor), and intra-donor phylogenetic analysis is equally applicable to heavy and light chains. Briefly, the computational procedure consists of an iterative analysis based on the neighbor-joining (NJ) method (Kuhner and Felsenstein, 1994) implemented in CLUSTALW2 (Larkin et al., 2007) and a final analysis based on the maximum-likelihood (ML) method with molecular clock implemented in DNAMLK[2] in the PHYLIP package v3.69[3]. In the NJ-based analysis, donor sequences of a particular germline origin were first randomly shuffled and divided into subsets of no more than 5,000 sequences. Then, PGT135–137 and respective germline sequence, IGHV4-39*07 for heavy chain and IGKV3-15*01 for light chain, were added to each subset. A NJ tree was constructed for each subset using the "Phylogenetic trees" option in CLUSTALW2 (Larkin et al., 2007). The donor sequences that clustered in the smallest branch that contains PGT135–137 were extracted from each NJ tree and combined into a new data set for the next round of analysis. The analysis was repeated until convergence, where all the donor sequences resided within a subtree containing PGT135–137 and no other sequences resided between this subtree and the root, and where further repeat of the analysis did not change the NJ tree. The ML-based analysis was used to confirm the intra-donor dendrogram derived from the NJ-based analysis. Starting from the data set obtained from the last iteration of NJ analysis, the MSA generated by CLUSTALW2 (Larkin et al., 2007) was provided as input to construct a phylogenetic tree using DNAMLK. Usually, any sequences outside the ML-defined subtree were discarded, but in this study we tested light chains identified by NJ method but immediately outside the rooted ML-defined PGT135–137 subtree. The displayed phylogenetic trees were generated using Dendroscope (Huson et al., 2007), ordered to ladderize right and rooted at the germline genes.

A description of the antibodyomics software (Antibodyomics1.0) utilized in this paper is being prepared for publication.

ANTIBODY EXPRESSION AND PURIFICATION

Antibody production followed previously described procedures (Wu et al., 2011). Briefly, sequences were selected using the respective bioinformatics procedure and checked for sequencing errors using an automatic error-correction procedure followed by manual inspection. The corrected antibody sequences were synthesized (GenScript USA Inc. and Blue Heron Biotech, LLC.) and cloned into the CMV/R expression vector (Barouch and Nabel, 2005) containing the constant regions of IgG1. All synthesized heavy chains were paired with PGT137 light-chain DNA, and synthesized light chains were paired with PGT137 heavy-chain DNA for transfection. Full-length IgGs were expressed from transient transfection of 293F cells and purified using a recombinant protein-A column (Pierce).

HIV-1 NEUTRALIZATION

Neutralization was measured using HIV-1 Env-pseudoviruses to infect TZM-bl cells as described (Li et al., 2005; Wu et al., 2009; Seaman et al., 2010). Neutralization curves were fit by non-linear regression using a five-parameter hill slope equation as described (Seaman et al., 2010). The 50% and 80% inhibitory concentrations (IC_{50} and IC_{80}) were reported as the antibody concentrations required to inhibit infection by 50% and 80% respectively.

RESULTS

Experiments involving both sequencing technologies and computational analyses are described. Because variable region transcripts of antibodies are over 300 nucleotides in length and because the high similarity between different antibody transcripts precludes assembly of full sequences from fragments, we used 454 pyrosequencing, which is currently one of the few next-generation sequencing technologies to provide reads of sufficient length

[1]http://www.ncbi.nlm.nih.gov/igblast/
[2]http://evolution.genetics.washington.edu/phylip/doc/dnamlk.html
[3]http://evolution.genetics.washington.edu/phylip.html

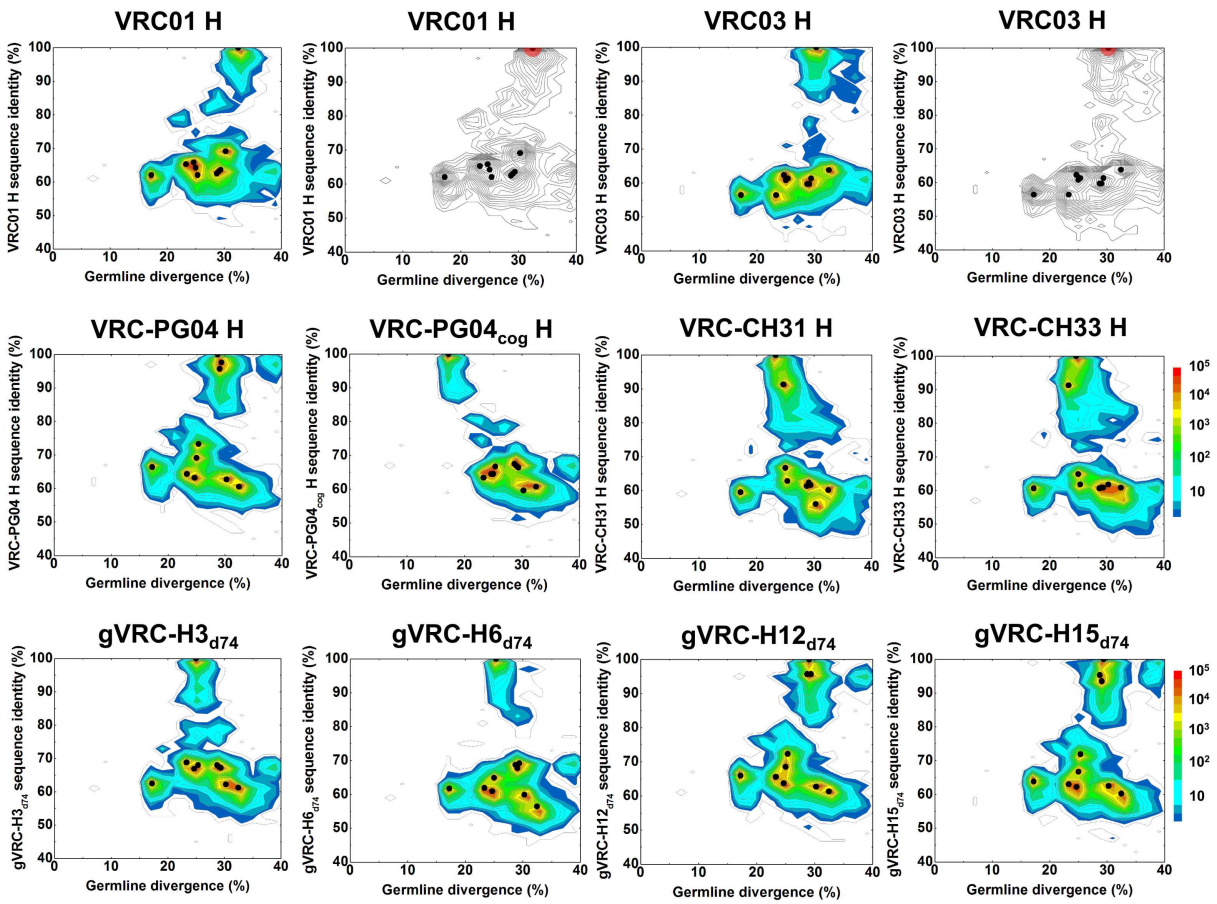


FIGURE 1 | Sequence variation as a consequence of 454 pyrosequencing for ten plasmid-control antibodies. To quantify sequencing error, ten antibodies, input as purified plasmid DNA, were subjected to 454 pyrosequencing. Tested plasmid antibodies included VRC01, VRC03, VRC-PG04, VRC-CH31, VRC-CH33, a codon-optimized version of inferred, reverted unmutated ancestor of VRC-PG04 (termed VRC-PG04$_{cog}$), gVRC-H3$_{d74}$, gVRC-H6$_{d74}$, gVRC-H12$_{d74}$, and gVRC-H15$_{d74}$. Heavy chain sequences are plotted as a function of sequence identity to the plasmid antibody (vertical axes) and of sequence divergence from their germline gene allele, IGHV1-2*02 (horizontal axes). The sequencing data used for divergence/identity analysis was processed by the standard bioinformatics pipeline without the error-correction step. Color coding indicates the number of sequences. For VRC01 and VRC03, additional contour plots displaying the estimated mutational error range (one root-mean-square deviation, 1.38% for VRC01 group and 1.26% for VRC03 group) have been shaded red around the input antibody.

(Reddy et al., 2009; Reddy and Georgiou, 2011; Wu et al., 2011). However, 454 pyrosequencing is known to suffer from high error rates (Prabakaran et al., 2011). We therefore begin by characterizing the accuracy of 454 pyrosequencing applied to a set of plasmid standards consisting of known HIV-neutralizing antibodies. We then describe 454 pyrosequencing of antibody heavy-chain transcripts from donor 39 (Walker et al., 2011), and analyze these data bioinformatically and functionally. We follow this with a similar analysis of donor 39 light-chain transcripts.

CHARACTERIZATION OF 454 PYROSEQUENCING ERRORS ON ANTIBODY TRANSCRIPTS

To investigate the extent of 454 pyrosequencing errors on the antibodyome analysis, we carried out a sequencing experiment on the heavy chains of 10 selected antibodies (Wu et al., 2011), including five from B cell sorting-based isolation, VRC01, VRC03, VRC-PG04, VRC-CH31, and VRC-CH33, one codon-optimized version of inferred reverted unmutated ancestor of VRC-PG04 (termed VRC-PG04$_{cog}$), and four identified from previous 454 pyrosequencing study, gVRC-H3$_{d74}$, gVRC-H6$_{d74}$, gVRC-H12$_{d74}$, and gVRC-H15$_{d74}$. The plasmid sequencing data was processed with the same bioinformatics pipeline used for donor sequencing data (Figure S1 in Supplementary Material). Sequence reads were subjected to an error-correction procedure, which was aimed to fix deletion and insertion errors that cause protein translation problems (see Materials and Methods). Results obtained with and without error correction were compared to examine the effect of error correction on observed sequence variation.

A divergence/identity analysis was first carried out on the 10 plasmid data set, obtained without (**Figure 1**) and with error correction (Figure S2 in Supplementary Material). Since divergence and identity were calculated at the nucleotide level, error correction appeared to have little effect on the sequence distribution. Ideally, if the 454 pyrosequencing did not produce any

Table 1 | Percent sequence-identity matrix of 10 plasmid antibody heavy-chain variable domains tested by 454 pyrosequencing.

	VRC01	VRC03	VRC-PG04	VRC-CH31	VRC-CH33	VRC-PG04$_{cog}$	gVRC-H3$_{d74}$	gVRC-H6$_{d74}$	gVRC-H12$_{d74}$	gVRC-H15$_{d74}$
VRC01	100.0	63.8	60.5	60.1	60.8	60.7	61.2	56.4	61.3	60.2
VRC03	69.1	100.0	62.7	56.0	61.8	59.6	62.2	59.9	62.7	62.5
VRC-PG04	62.5	59.7	100.0	61.3	60.6	67.5	68.0	68.7	95.7	95.3
VRC-CH31	65.3	56.4	64.3	100.0	91.3	63.4	68.8	61.9	65.6	63.0
VRC-CH33	65.8	62.3	63.2	91.3	100.0	64.5	66.9	60.9	63.7	62.2
VRC-PG04$_{cog}$	62.0	56.4	66.4	59.5	60.6	100.0	62.5	61.7	65.9	63.8
gVRC-H3$_{d74}$	64.2	60.8	69.1	66.7	64.9	64.5	100.0	64.9	68.5	66.7
gVRC-H6$_{d74}$	62.0	61.3	73.3	62.8	61.8	66.7	68.0	100.0	72.3	71.9
gVRC-H12$_{d74}$	63.1	59.7	95.7	62.3	60.8	66.9	67.5	67.7	100.0	93.5
gVRC-H15$_{d74}$	63.6	61.3	97.6	61.6	60.8	66.4	67.2	69.2	95.7	100.0

The heavy-chain variable domains of 10 antibodies were sequenced by 454 pyrosequencing, including VRC01, VRC03, VRC-PG04, VRC-CH31, VRC-CH33, a codon-optimized version of inferred reverted unmutated ancestor of VRC-PG04 (termed VRC-PG04$_{cog}$), and four neutralizing antibodies whose heavy-chain variable domains were identified from donor 74 in a previous study (Wu et al., 2011) – gVRC-H3$_{d74}$, gVRC-H6$_{d74}$, gVRC-H12$_{d74}$, and gVRC-H15$_{d74}$. The diagonal values, circled and all 100%, separate the upper- and lower-triangular portions of the matrix, which differ in how the sequence identity was calculated, specifically, which sequence was used as the "template" and which as the "query," as the differing sequences have different lengths.

errors, especially mutations, the distribution – irrespective of the antibody being used as template – would yield, on divergence/identity plots, 10 discrete points, each corresponding to one of the input sequences. In contrast, divergence/identity plots revealed broad islands centered around each of these 10 antibody sequences (**Figure 1**). The shape and area of each island provide a visual representation of the extent of the 454 pyrosequencing errors. As shown in **Table 1**, 5 of the 10 antibodies – those with an identity gap of 25% or greater to the next most closely related sequence – were easily distinguished from each other, while other more closely related variants, e.g., VRC-CH31 and VRC-CH33, overlapped (**Figure 1**). Based on identity considerations (**Table 1**) and the scope of each island in divergence/identity plots (**Figure 1**), a single cutoff of 75% was applied to group 454 pyrosequencing-determined sequences for VRC01, VRC03, VRC-PG04$_{cog}$, gVRC-H3$_{d74}$, and gVRC-H6$_{d74}$.

Each of these five 454 pyrosequencing-determined sequence groups was analyzed for mutations, insertions, and deletions relative to the input plasmid sequence, as well as total number of reads and their redundancy (**Table 2**). For four of the plasmids ∼50,000 reads were obtained; for gVRC-H6$_{d74}$, however, only about one fourth as many were obtained, which may relate to a lower efficiency of the primer used for gVRC-H6$_{d74}$. In terms of redundancy, for three of the plasmids between one fifth and one half of the reads were identical to the input plasmid, whereas for VRC01 and gVRC-H6$_{d74}$, only a small fraction ($<$1 and $<$10%) of the reads were identical to the input plasmid, a result of insertions in most of the sequences. Note that after error correction, 20–3254 more sequences became identical to the input antibodies (**Table 2**). Overall, for an antibody of typical length, ∼5-nucleotide mutations were observed between 454 pyrosequencing reads and corresponding input sequences. Error correction appeared to cause an increased count of mutation errors while decreasing insertion and deletion errors that produce stop codons and nonsense codons in protein translation. Currently used correction procedure was able to improve the identity of translated protein sequence to respective germline gene by an average of 14.1% (Figures S1C,D in Supplementary Material).

We then examined the accuracy of bioinformatically selected representative sequences for these five antibody groups. Note that all these sequences have been subjected to a template-based error-correction procedure in the pipeline processing. A "divide-and-conquer" procedure (See Materials and Methods) was used for sequence calculation. Remarkably, the representative sequence was 100% identical to the "true" sequence used as input for 454 pyrosequencing for VRC-PG04$_{cog}$, gVRC-H3$_{d74}$, and gVRC-H6$_{d74}$, while having one 1-nucleotide deletion and two 1-nucleotide insertions for VRC01 and VRC03, respectively. None had mutation errors. Such consensus-based sequence picking procedure may prove useful in the cases where a population of closely related sequences is observed on the divergence/identity plot, as indicated by a densely populated island.

454 PYROSEQUENCING OF DONOR 39 IGHV4 FAMILY AND BIOINFORMATICS ANALYSIS OF HEAVY CHAINS

We next performed 454 pyrosequencing of PGT135–137-related heavy-chain transcripts from donor 39 PBMCs. mRNA from ∼5 million PBMCs was used for reverse transcription to produce template cDNA, and PCR was used to amplify IgG and IgM heavy-chain sequences from the IGHV4 family using forward primers that overlapped the end of the V gene leader sequence and the start of the V region and reverse primers covering the start of the constant domain (Table S1 in Supplementary Material).

Next-generation pyrosequencing provided 918,298 reads, which were processed with a bioinformatics pipeline that involved assignment of germline origin genes, 454 pyrosequencing-error correction, and extraction of CDR H3 regions for lineage assignment. Overall about 85.3% of the raw reads spanned over 400 nucleotides, covering the entire variable domain. After computational assignment of V, D, and J gene components, 142,842 sequences were assigned to IGHV4-39 germline family, accounting for ∼16% of the expressed VH4 antibodyome. Each sequence

Table 2 | Statistical analysis of 454 pyrosequencing-induced errors for five plasmid antibodies.

Antibody	Length (nt)	N_{Seq}	N_{Iden}	Unnormalized			Normalized (per 100 nt)		
				RMS_{Mut} (nt)	RMS_{Ins} (nt)	RMS_{Del} (nt)	RMS_{Mut} (nt)	RMS_{Ins} (nt)	RMS_{Del} (nt)
NO ERROR CORRECTION									
VRC01	363	47542	289	5.0	2.2	1.7	1.38	0.61	0.47
VRC03	390	53734	12309	4.6	4.1	0.9	1.18	1.05	0.23
VRC-PG04$_{cog}$	369	43718	21281	5.0	1.7	0.5	1.36	0.46	0.14
gVRC-H3$_{d74}$	381	53147	19843	6.3	1.9	1.1	1.65	0.50	0.29
gVRC-H6$_{d74}$	399	13639	1013	6.4	2.8	1.6	1.60	0.70	0.40
WITH ERROR CORRECTION									
VRC01	363	47542	334	5.8	1.9	1.2	1.60	0.52	0.33
VRC03	390	53734	12948	4.7	4.0	0.9	1.21	1.03	0.23
VRC-PG04$_{cog}$	369	43718	22021	5.1	1.6	0.4	1.38	0.43	0.11
gVRC-H3$_{d74}$	381	53147	23097	6.5	1.7	0.9	1.71	0.45	0.24
gVRC-H6$_{d74}$	399	13639	1033	6.6	2.7	1.5	1.65	0.68	0.38

Columns include antibody name, nucleotide-sequence length of antibody heavy-chain variable domain, number of 454 pyrosequencing-determined heavy-chain variable-domain sequences for this antibody, number of sequences 100% identical to the sequenced antibody heavy-chain variable-domain, root-mean-square (RMS) fluctuation of 454 pyrosequencing-induced mutations, insertions, and deletions with respect to the input antibody sequence, and their values after normalization by a length of 100 nucleotides.

*As shown in the percent sequence identity matrix in **Table 1** and the divergence/identity plots in **Figure 1**, only five antibodies can be distinguished from others using a single sequence identity cutoff. After mapping the 454 pyrosequencing-determined heavy-chain variable-domain sequences onto 10 plasmid antibodies, a single cutoff of 75% was applied to extract sequences corresponding to VRC01, VRC03, VRC-PG04$_{cog}$, gVRC-H3$_{d74}$, and gVRC-H6$_{d74}$, respectively.*

RMS_{Mut}, RMS_{Ins}, and RMS_{Del} were calculated using the formula:

$$RMS_X = \sqrt{\sum_i^{N_{Seq}} \frac{(X_i - \overline{X})^2}{N_{Seq}}},$$

X denotes the type of sequencing error to be characterized, mutation (Mut), insertion (Ins), and deletion (Del), respectively; $\overline{X}$ denotes the averaged sequencing error; N_{Seq} denotes the total number of sequences within a given antibody group.

The RMS values were normalized using $RMS_{normalized} = RMS_{unnormalized}/length_{unnormalized} \times 100$ to take into account the difference in sequence length.

was subjected to an automatic error-correction scheme. For donor 39 heavy chains, the correction procedure improved the accuracy of protein translation, measured by protein sequence identity to inferred gemline gene, by an average of 20.4%. The results for pipeline processing of heavy-chain data set are listed in Figure S3 in Supplementary Material.

First, germline family analyses were performed using two standard methods – IMGT (Brochet et al., 2008) and IgBLAST (see text footnote 1; **Table 3**). These analyses assigned PGT135–137 gene origins to IGHV4-39 with two possible alleles (*03 or *07), to three potential D genes, and the J gene IGHJ5*02. An analysis of the third complementarity-determining region of the heavy chain (CDR H3) showed 80–90% sequence identity between PGT135–137, suggestive of a common lineage. The likely clonal origin of PGT135–137 indicates that they will all have the same V(D)J origin, with the different origin gene assignments by IMGT and IgBLAST likely due to their high divergence of ∼20% from ancestral gene.

Second, a divergence/identity analysis of 454 pyrosequencing-derived sequences assigned to IGHV4-39 origin was performed (**Figure 2**). The IGHV4-39-related sequences revealed a maximum divergence of 30.4% and an average divergence of 7.7% from germline. An island of sequences was observed at ∼90% identity to PGT137 with divergence of 20–25% from VH4-39, indicative of PGT137-related antibodies with similar evolutionary distance from the origin.

Third, intra-donor phylogenetic analysis (see Materials and Methods) was applied to identify the somatic variants of PGT135–137 from the donor 39 heavy-chain sequencing data. In this analysis, a set of clonally related template antibodies is used to interrogate sequences from the same donor using phylogenetic analysis. Phylogenetic analysis, using a tree rooted by the inferred germline gene IGHV4-39*07, produced a ML dendrogram with 202 heavy-chain variable-domain sequences identified by their co-segregation with PGT135–137 (**Figure 3**). Most of the intra-donor-identified sequences clustered with PGT137, and one sequence clustered with PGT136.

Fourth, CDR H3 variation was analyzed for the 202 PGT135–137-related heavy-chain variable-domain sequences. One hundred seven were found to have identical CDR H3 sequences, as the same as the nucleotide-sequence consensus. With a maximum of five mutations from the consensus, the average CDR H3 variation was 1.2, indicative of a rather conserved signature of PGT135–137 lineage.

PGT135–137 SOMATIC HEAVY-CHAIN POPULATIONS AND FUNCTIONAL CHARACTERIZATION

To gain insight into the functional diversity of the antibodies identified by 454 pyrosequencing and bioinformatics methods, a clustering procedure was used to analyze the 202 identified heavy chains and to select representative sequences for further characterization. We used BLASTClust (Altschul et al., 1997) clustering

Table 3 | Recombination origins of antibodies PGT135–137.

Antibody	V gene ($SeqID_{togerm}$)	D gene	J gene ($SeqID_{togerm}$)	$SeqID_{toPGT135}$
HEAVY CHAIN (CDR H3 LENGTH = 18)				
PGT135	IGHV4-39*07 (81.4%)	IGHD3-9*01	IGHJ5*02 (72.0%)	100.0%
PGT136	IGHV4-39*07 (82.1%)	IGHD2-8*02	IGHJ5*02 (78.0%)	83.0%
PGT137	IGHV4-39*03 (77.2%)	IGHD2-15*01	IGHJ5*02 (78.0%)	82.0%
LIGHT CHAIN (CDR-L3 LENGTH = 9)				
PGT135	IGKV3-15*01 (82.4%)	–	IGKJ1*01 (94.3%)	100.0%
PGT136	IGKV3-15*01 (86.4%)	–	IGKJ1*01 (97.1%)	84.4%
PGT137	IGKV3-15*01 (87.8%)	–	IGKJ1*01 (97.1%)	85.0%

V, D, and J gene components of PGT135–137 were determined by two servers: IgBLAST (http://www.ncbi.nlm.nih.gov/igblast/) and IMGT/V-Quest (http://www.imgt.org/IMGT_vquest/vquest).

$SeqID_{togerm}$ is the nucleotide-sequence identity with respect to the germline gene.

$SeqID_{toPGT135}$ is the nucleotide-sequence identity with respect to PGT135.

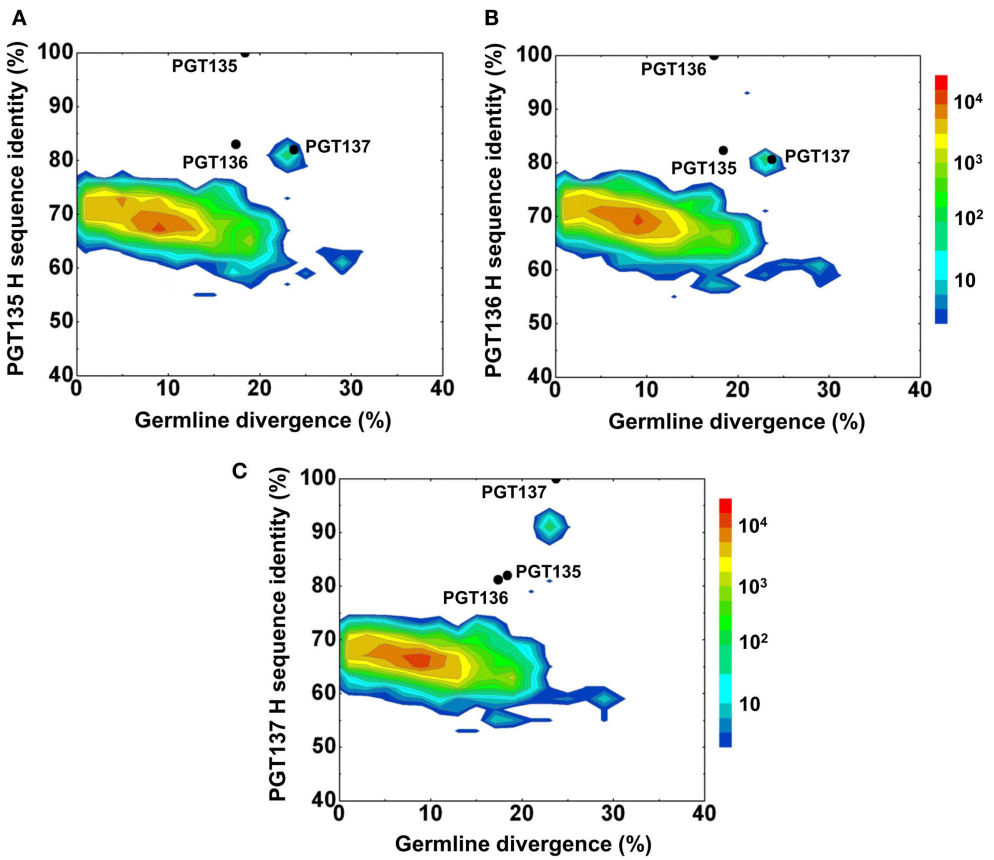


FIGURE 2 | Repertoire of donor 39 heavy-chain variable-domain sequences of IGHV4-39 origin determined by 454 pyrosequencing. After processing by a standard bioinformatics pipeline (see Materials and Methods), 1,41,298 full-length, heavy-chain variable-domain sequences from IGHV4-39 germline family were obtained. These are plotted as a function of sequence identity to the heavy-chain variable domain of PGT135 **(A)**, PGT136 **(B)**, and PGT137 **(C)** and of sequence divergence from inferred IGHV4-39 germline allele. Color coding indicates the number of sequences. The 10%-identity gap indicates that the sequences within the upper island in 2C are somatic variants of PGT137 and not caused by sequencing errors.

function and an identity cutoff of 95% to sample the natural variation. We chose this cutoff to be greater than the ∼1.6% "false" sequence variation induced by 454 pyrosequencing errors (**Table 2**). A total of 15 clusters emerged. In the BLASTClust output, the first sequence of each cluster was selected to "represent" the cluster (**Figure 4A**) and were synthesized and reconstituted with the PGT137 light chain for functional assessment of HIV-1 neutralization, which was carried out on two viruses sensitive

FIGURE 3 | Evolutionary similarity of PGT135–137 to donor 39 heavy-chain variable-domain sequences. Germline-rooted maximum-likelihood tree of PGT135–137 and 202 sequences identified by the iterative intra-donor phylogenetic analysis of donor 39 heavy-chain variable domain sequences determined by 454 pyrosequencing. The iterative intra-donor phylogenetic analysis was based on an implementation of neighbor-joining (NJ) method. Collapsed branches are indicated by *Collapse {N: M}*, in which *N* is the branch depth (number of intermediate nodes) and *M* is the number of sequences within the branch. All sequences are on the PGT137 branch except for 174091, which is somatically related to PGT136.

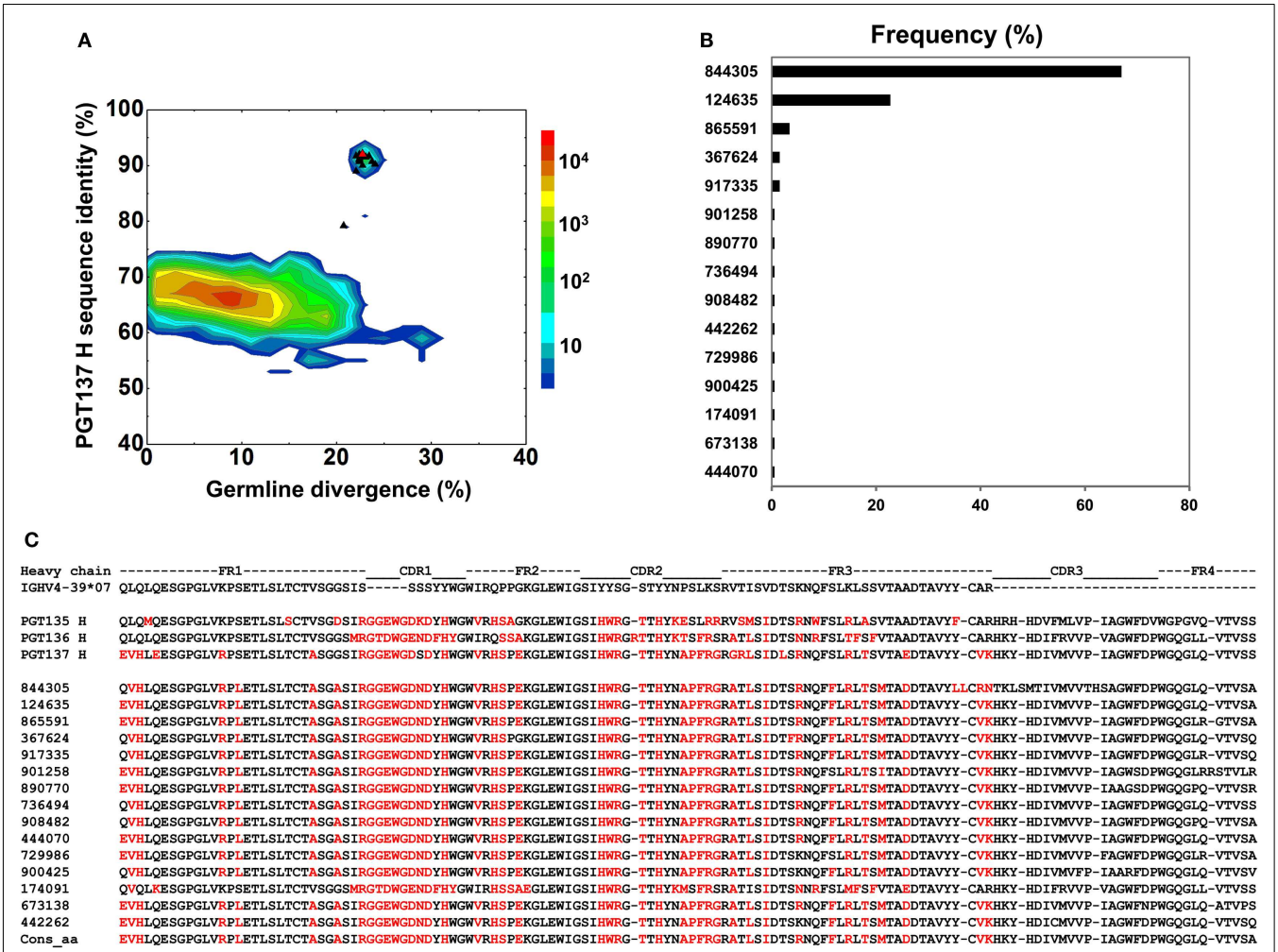


FIGURE 4 | Sequence selection for functional characterization of heavy chains from donor 39. (A) Divergence/identity analysis of 15 heavy-chain variable-domain sequences obtained from the clustering analysis of 202 sequences identified by intra-donor phylogenetic analysis. Sequences of IGHV4-39 origin are plotted as a function of sequence identity to PGT137 heavy chain and sequence divergence from inferred germline allele, with 15 selected sequences shown as black triangles and their amino-acid consensus as red triangle. **(B)** Percent population of 15 clusters obtained using a sequence identity cutoff of 95%. Each cluster is indicated by its representative sequence. "Frequency" refers to the total number of sequences observed for each cluster. **(C)** Protein sequences of 15 cluster representatives and their amino-acid consensus. Sequences are aligned to the inferred germline gene, IGHV4-39*07. Framework regions (FR) and complementarity-determining regions (CDRs) are based on Kabat nomenclature. Amino acids mutated from the germline gene are shown in red.

to PGT135–137 antibodies. Out of 15 tested heavy-chain variable domain sequences, when paired with PGT137 light chain, 11 reconstituted antibodies showed neutralization to different extents (**Table 4**).

The two largest clusters, with 136 and 46 sequences, respectively, accounted for ∼90% of the sequences (**Figure 4B**), while 10 of the 15 clusters contained only a single member. A consensus sequence (ConsAA), calculated from the alignment of 15 representative sequences (**Figure 4C**), was also synthesized. Notably, the reconstituted amino-acid consensus displayed neutralization almost on par with wild-type PGT137 (**Table 4**).

Despite their apparent clonality, the clustering procedure reveals 15 clusters. The topology of the dendrogram produced from phylogenetic analysis indicates that these 15 clusters represent populations of somatically related antibodies evolving along distinct branches by standard mechanisms of hypermutation (**Figure 3**). We analyzed these 15 somatic populations for prevalence of mutations, insertions, and deletions (Table S2 in Supplementary Material). Note that the representative sequence of cluster 1 (#844305) contained two insertions in the CDR H3 region which were not seen in other members of the cluster, suggesting that these insertions might be sequencing errors. Indeed, this heavy chain could not be expressed when reconstituted with PGT137 light chain. We also analyzed each of these populations by divergence/identity plot (**Figure 5**). Overall, sequences chosen to represent the 15 somatic populations showed diverse neutralization characteristics (Table S2 in Supplementary Material). Some antibodies, for example from clusters 2, 3, 14, and

Table 4 | Neutralization titers of 21 chimeric antibodies derived from 454 pyrosequencing of donor 39 against HIV-1 pseudoviruses from clade A and clade D.

Seq. index	Nab name	Neutralization IC_{50} titers (μg/ml)			Neutralization IC_{80} titers (μg/ml)		
		RW020.2	UG024.2	MuLV	RW020.2	UG024.2	MuLV
HEAVY-CHAIN VARIANT PAIRED WITH PGT135 LIGHT CHAIN							
124635	gVRC-H1$_{d39}$	0.005	0.021	>50	0.022	0.107	>50
865591	gVRC-H2$_{d39}$	0.004	0.017	>48	0.016	0.08	>48
367624	gVRC-H3$_{d39}$	0.009	0.253	>50	0.044	1.41	>50
917335	gVRC-H4$_{d39}$	0.003	0.243	>50	0.015	3.41	>50
736494	gVRC-H5$_{d39}$	0.003	0.458	>50	0.021	4.18	>50
442262	gVRC-H6$_{d39}$	0.409	18.4	>50	12.8	>50	>50
729986	gVRC-H7$_{d39}$	2.31	>50	>50	>50	>50	>50
900425	–	>50	>50	>50	>50	>50	>50
174091	gVRC-H8$_{d39}$	>50	0.02	>50	>50	0.072	>50
673138	gVRC-H9$_{d39}$	1.49	1.86	>50	23.6	11.5	>50
444070	gVRC-H10$_{d39}$	0.008	0.027	>50	0.03	0.138	>50
ConsAA	gVRC-H11$_{d39}$	0.003	0.016	>50	0.012	0.08	>50
LIGHT-CHAIN VARIANTS PAIRED WITH PGT137 HEAVY CHAIN							
107548	gVRC-L1$_{d39}$	0.0007	0.008	>50	0.006	0.036	>50
219622	–	>50	>50	>50	>50	>50	>50
210137	–	>50	>50	>50	>50	>50	>50
215528	–	>50	>50	>50	>50	>50	>50
425756	gVRC-L2$_{d39}$	<0.0006	0.007	>50	0.004	0.033	>50
121553	gVRC-L3$_{d39}$	0.04	0.423	>50	0.245	2.16	>50
303540	gVRC-L4$_{d39}$	0.075	1.11	>50	0.841	6.61	>50
378597	gVRC-L5$_{d39}$	0.03	0.375	>50	0.167	1.26	>50
521298	–	>50	>50	>50	>50	>50	>50
537707	gVRC-L6$_{d39}$	0.012	0.115	>50	0.057	0.436	>50
WILD-TYPE PGT135–137 ANTIBODIES AND VRC01 ANTIBODY AS CONTROLS							
PGT135	PGT135	>50	0.002	>50	>50	0.007	>50
PGT136	PGT136	>50	0.008	>50	>50	0.026	>50
PGT137	PGT137	0.001	0.007	>50	0.006	0.041	>50
VRC01	VRC01	0.157	0.207	>50	0.682	0.718	>50

Columns include sequence index (for heavy chains, the amino-acid consensus is denoted by "ConsAA"; for controls, the antibody name is used as sequence index), neutralizing antibody name based on the nomenclature used in previous studies (Wu et al., 2011; Zhu et al., under review), neutralization IC_{50} and IC_{80} titers for viruses RW020.2 (HIV-1 clade A), UG024.2 (HIV-1 clade D) and MuLV (murine leukemia virus). Two sets of chimeric antibodies, 22 in total, were expressed by pairing the 12 heavy chains derived from 454 pyrosequencing and the PGT137 light chain, and pairing the 10 light chains derived from 454 pyrosequencing and the PGT137 heavy chain.

The wild-type mAbs PGT135–137 and wild-type VRC01 were included as controls.

MuLV stands for murine leukemia virus, which was included as a negative control.

"–" denotes expressed but non-neutralizing sequence after reconstituted with the PGT137-partner chain.

15 (gVRC-H1$_{d39}$, gVRC-H2$_{d39}$, gVRC-H9$_{d39}$, and gVRC-H10$_{d39}$), neutralized clade A – RW020.2 and clade D – UG024.2 with roughly equal potency. Some antibodies, for example from clusters 4, 5, 8, and 10 (gVRC-H3-H6$_{d39}$), neutralized clade A-RW020.2 25-150-fold more potently than clade D. While the antibody from cluster 13 (gVRC-H8$_{d39}$) neutralized clade D – UG024.2 with at least 100-fold greater potency than clade A. These results provide an example for how somatically related antibodies can significantly differ in their neutralization specificities. This begins to provide insight into how populations of somatically related antibodies can engender neutralization breadth significantly different than any individual member.

454 PYROSEQUENCING OF DONOR 39 IGKV3 FAMILY AND BIOINFORMATICS ANALYSIS OF LIGHT CHAINS

We next performed 454 pyrosequencing of PGT135–137-related light-chain transcripts from donor 39 PBMCs. mRNA from ∼5 million PBMCs was used for reverse transcription to produce template cDNA, and PCR was used to amplify light-chain sequences from the IGKV3 family.

The 454 pyrosequencing provided 971,165 reads, which were then processed using a pipeline adapted for κ-chain analysis. For donor 39, about 83.3% of the raw reads were 400 nt or longer, effectively covering the light-chain variable domain. After V and J gene assignment, 91,951 sequences were determined to belong to

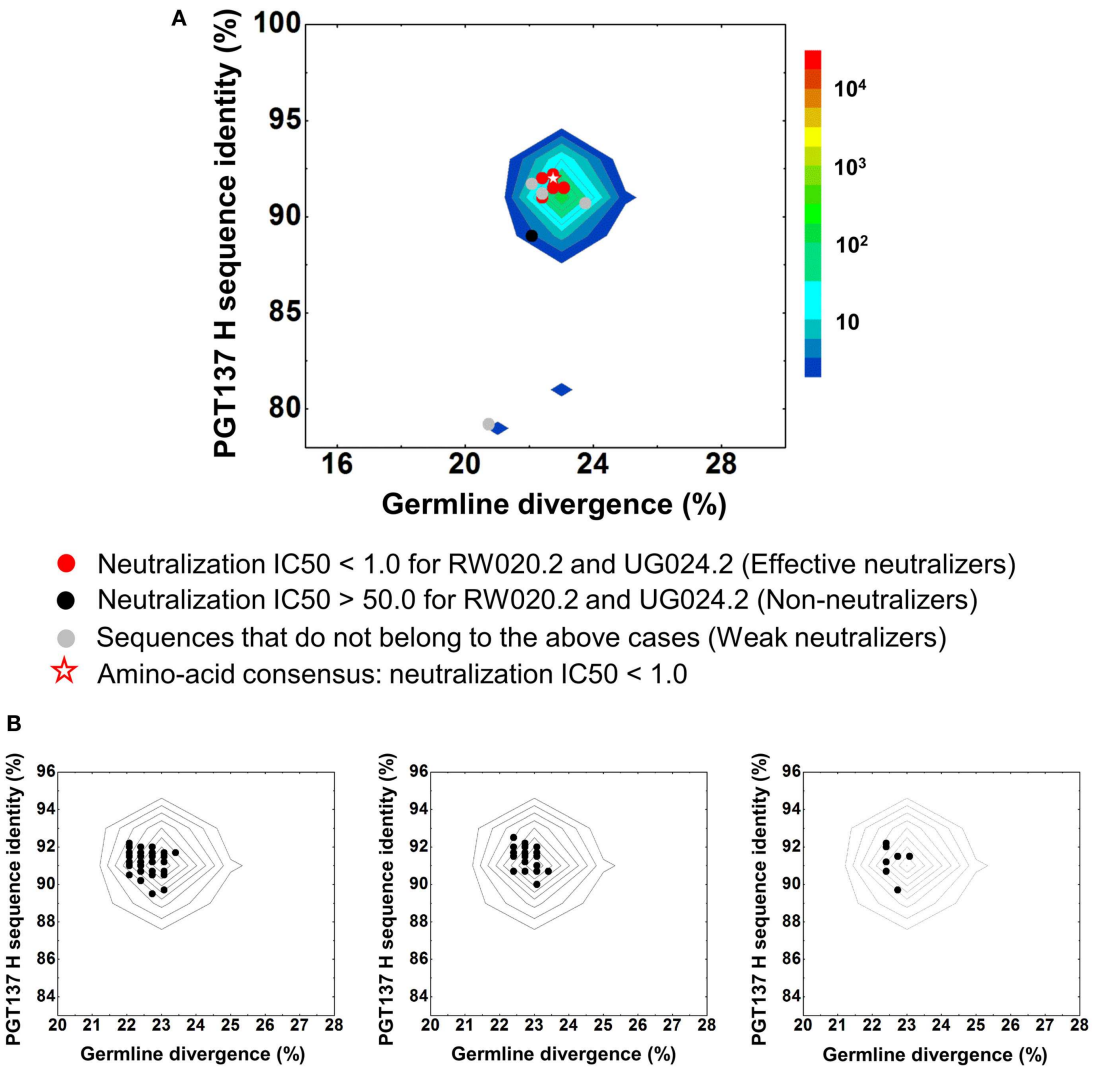


FIGURE 5 | Divergence/identity analysis of heavy-chain neutralization. (A) The expressed heavy-chain sequences color-coded based on the neutralization potency of reconstituted antibodies, with IC50 <1.0 for both viruses shown in red (effective neutralizers), IC50 >50.0 for both viruses in black (non-neutralizers), and other cases in gray (weak neutralizers). The amino-acid consensus, when reconstituted with PGT137 light chain, neutralized both viruses with an IC50 <1.0 and is shown as a red hollow star. **(B)** The three largest clusters are displayed on the enlarged divergence/identity plot, with 136, 46, and 7 members, respectively.

IGKV3-15 germline family, accounting for 10% of the light chain reads obtained. After error correction, the accuracy of protein translation measured by the protein sequence identity to inferred gemline gene was improved by an average of 16.5%. The results for pipeline processing of light-chain data set are listed in Figure S4 in Supplementary Material.

First, the recombination origins of PGT135–137 light chains were analyzed (**Table 3**). PGT135–137 light chains were assigned to the same germline V gene allele, IGKV3-15*01, recombined with the same J gene, IGKJ1*01, supporting the notion that the discrepancy in heavy-chain germline assignment was likely an artifact caused by their high divergence.

Second, the divergence/identity analysis of 454 pyrosequencing-derived sequences assigned to the IGKV3-15*01 origin was performed (**Figure 6**). The IGKV3-15*01-related sequences revealed a maximum divergence of 20.9% and an average divergence of 6.3% from germline. Distinct sequence islands were observed at ∼100% identity to PGT136 and 95% identity to PGT137 – both with divergence of 10–15% from IGKV3-15*01. No distinct sequence island was observed that was closely related to PGT135.

Third, to identify light-chain somatic variants, we performed intra-donor phylogenetic analysis that combined an iterative NJ procedure for the high-throughput screening of sequencing data, and a ML calculation to confirm the NJ analysis and to provide the final dendrogram (see Materials and Methods). Two methods were usually in agreement, e.g., for donor 39 heavy chains, but differed here. The NJ-based analysis yielded 72 sequences within the

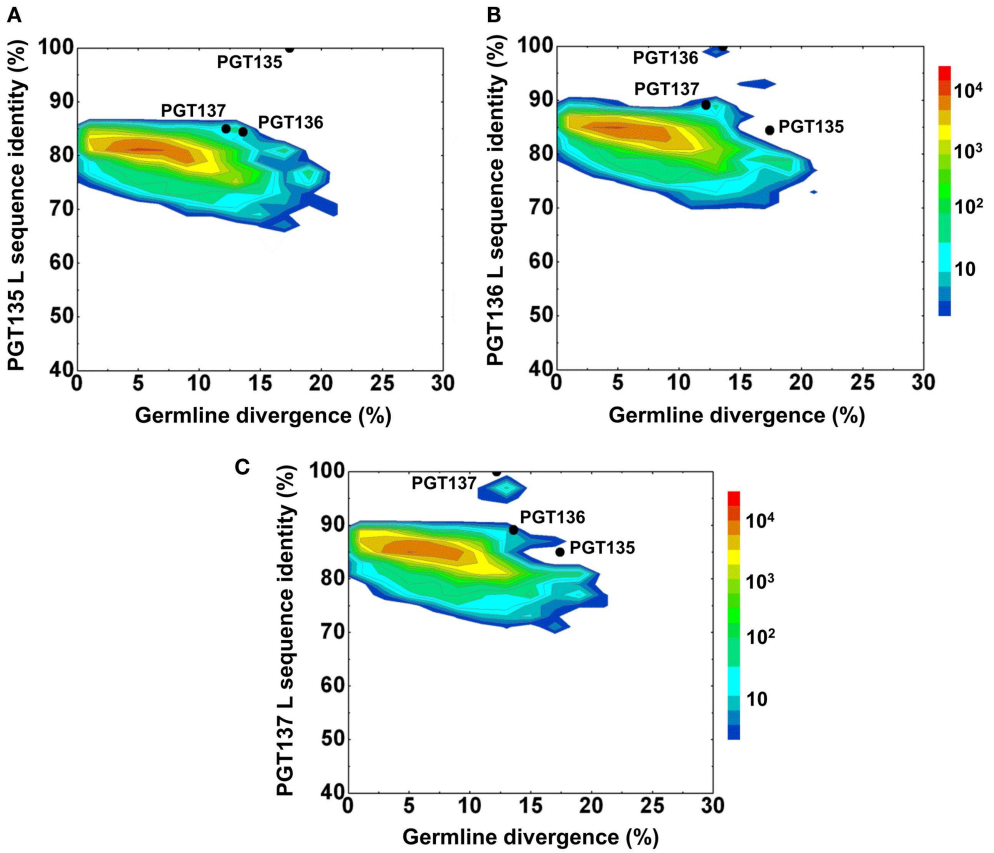


FIGURE 6 | Repertoire of donor 39 light-chain variable-domain sequences of IGKV3-15 origin determined by 454 pyrosequencing. After processed by a standard bioinformatics pipeline, 87,229 full-length, light-chain variable-domain sequences from IGKV3-15 germline family are plotted as a function of sequence identity to the light-chain variable-domain of PGT135 **(A)**, PGT136 **(B)**, and PGT137 **(C)** and of sequence divergence from inferred IGKV3-15 germline allele. Color coding indicates the number of sequences.

PGT135–137 subtree, whereas the subsequent ML-based analysis retained 57 of the 72 sequences within the PGT135–137 subtree (**Figure 7**), providing an example for functional characterization of similar but somatically unrelated sequences.

PGT135–137 SOMATIC LIGHT-CHAIN POPULATIONS AND FUNCTIONAL CHARACTERIZATION

By using the same 95% clustering procedure as for heavy chains, 14 light-chain clusters were identified from the phylogenetic tree. Representative sequences were selected, also as described for heavy chains, from the first 10 clusters for functional characterization (**Figure 8A**). We analyzed these 10 clusters for prevalence of mutations, insertions, and deletions (Table S3 in Supplementary Material). The largest cluster, lying within the population of PGT137-like sequences, contained 45 sequences or 63% of the subtree sequences (**Figure 8B**). All selected light-chain sequences possessed CDR L3s of the same length except for the sequences selected from the clusters 2 and 3 (**Figure 8C**). Out of 10 tested light-chain variable domain sequences, when reconstituted with the PGT137 heavy chain, six antibodies – representing six sequence clusters – showed neutralization of two HIV-1 strains from clade A and clade D. Notably, two of the light chains (gVRC-L1$_{d39}$ and gVRC-L2$_{d39}$) showed neutralization breadth slightly better than PGT135–137, and the light-chain variants neutralized clade A about 10-fold more effectively than the clade D (**Table 4**).

In contrast to the 454 pyrosequencing-identified heavy chains, the six neutralizing light-chain clusters were not localized to a single divergence/identity island (**Figure 9**). Indeed, neutralization was observed with clusters from at least three diverse locations on the divergence/identity plot. Nevertheless, the topology of the light-chain phylogenetic analysis indicates that these six clusters represent populations of somatically related antibodies (**Figure 7**).

DISCUSSION

Recently, select antibodies with the ability to neutralize diverse strains of HIV-1 have been identified in HIV-1 infected donors (Walker et al., 2009, 2011; Corti et al., 2010; Wu et al., 2010, 2011; Scheid et al., 2011). Like PGT135–137, antibodies from these donors often appear to be clonally related, to possess similar neutralization characteristics, and to cluster in a localized island (or islands) on identity/diversity plots. These islands observed in 454 pyrosequencing-derived analyses are often nearby but rarely overlap the few antibodies experimentally isolated from the same

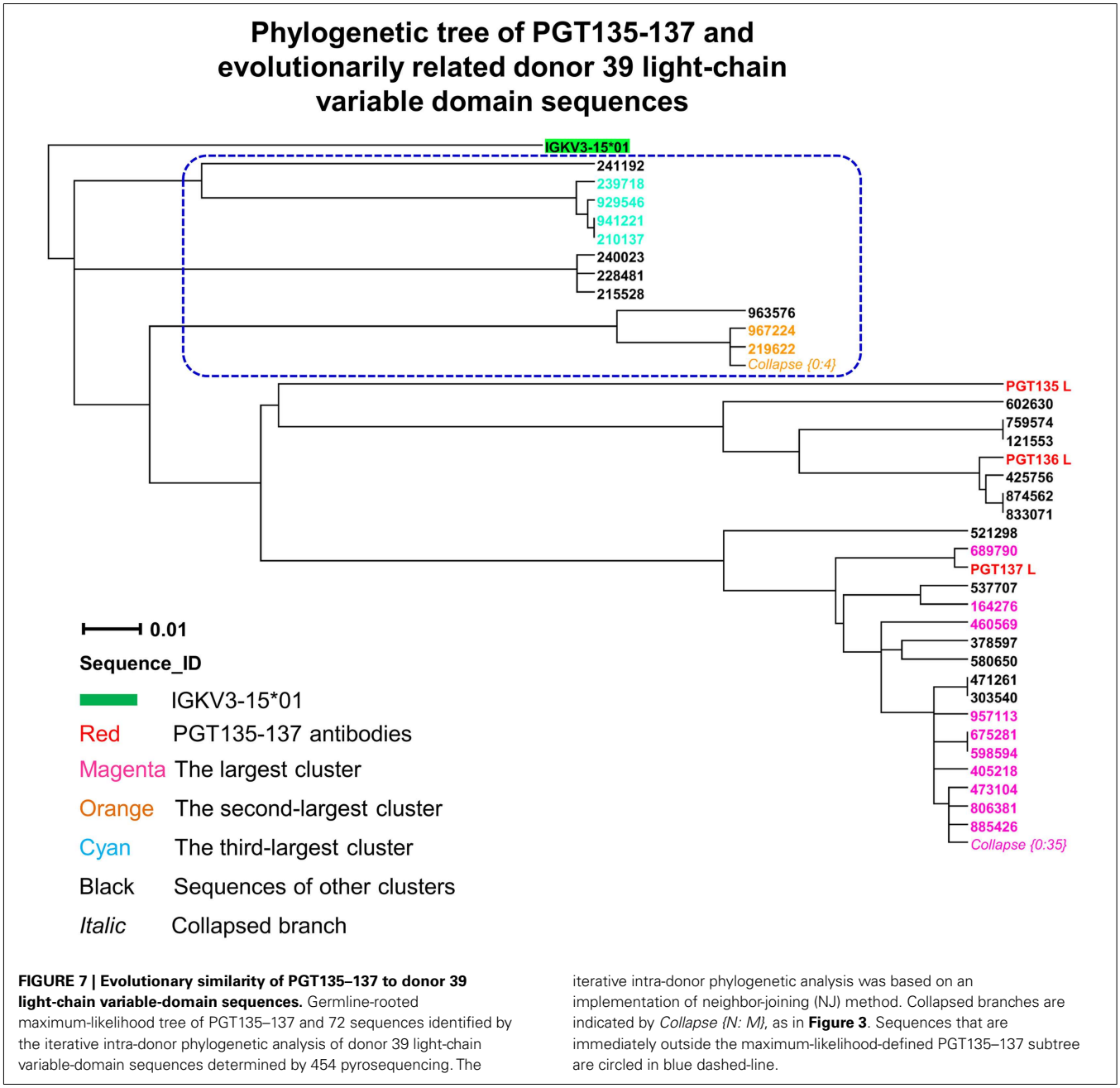
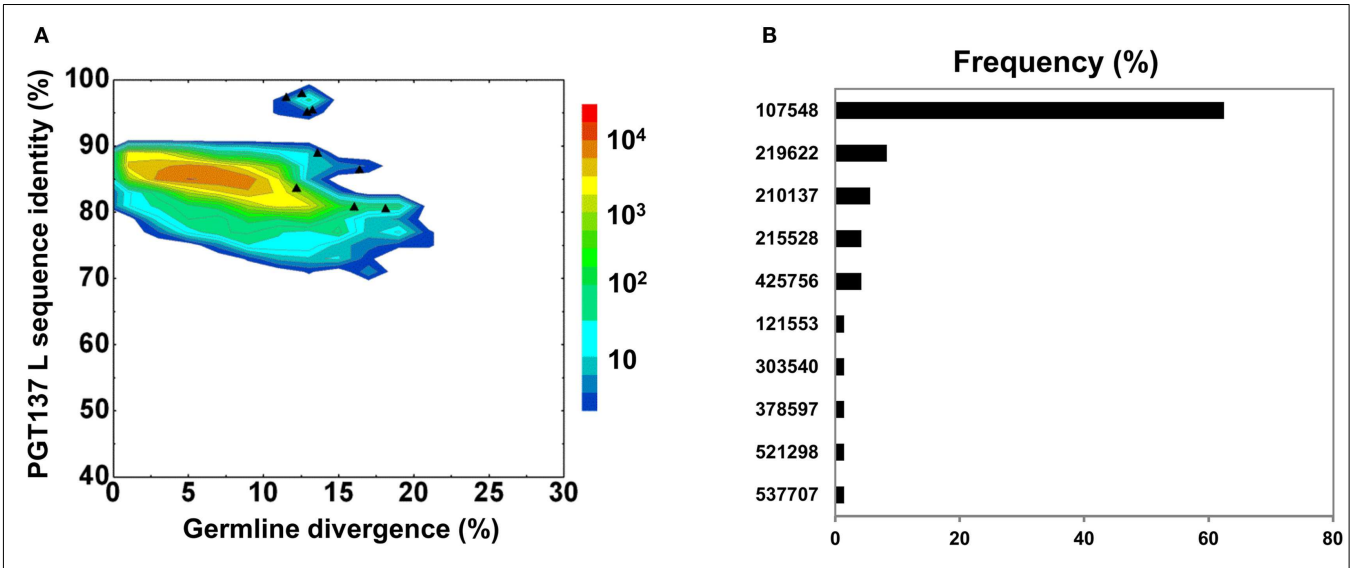


FIGURE 7 | Evolutionary similarity of PGT135–137 to donor 39 light-chain variable-domain sequences. Germline-rooted maximum-likelihood tree of PGT135–137 and 72 sequences identified by the iterative intra-donor phylogenetic analysis of donor 39 light-chain variable-domain sequences determined by 454 pyrosequencing. The iterative intra-donor phylogenetic analysis was based on an implementation of neighbor-joining (NJ) method. Collapsed branches are indicated by *Collapse {N: M}*, as in **Figure 3**. Sequences that are immediately outside the maximum-likelihood-defined PGT135–137 subtree are circled in blue dashed-line.

individual (even if they start with samples of exactly the same time point, as we have done here with donor 39). The differences between antibodies identified from sorting of memory B cells or by 454 pyrosequencing of B cell transcripts suggest that the experimental approaches may capture or sample different B cell population. In addition to exploring differences in phenotype of antibody identified by the two methods, we also explored differences related to the quantity of identified antibody. In particular, we ask whether the less-sparse view of the antibody repertoire provided by next-generation sequencing and systems-level bioinformatics might provide insight into the diversity of the antibody response.

With the heavy chains of PGT135–137, select sequences representing 15 distinct populations, showed dramatically different neutralization characteristics toward clade A and D viruses when reconstituted with the same light chain from PGT137. With the light chains of PGT135–137, select sequences representing 10 distinct populations were not localized to a discrete sequence island, indicating substantial differences in identity and diversity (**Figure 8**). Thus, even though these antibodies are somatically related, both their neutralization and sequence characteristics can diverge substantially (**Table 4**). These results demonstrate the utility of next-generation sequencing, which provides a more comprehensive sampling of sequences, and of



C

```
Light chain  ----------FR1----------____CDR1___------FR2------_CDR2__----------------FR3-------------___CDR3____---FR4----
IGKV3-15*01  EIVMTQSPATLSVSPGERATLSCRASQSVSSNLAWYQQKPGQAPRLLIYGASTRATGIPARFSGSGSGTEFTLTISSLQSEDFAVYYCQQYNNWPP-------------

PGT135 L     EIVMTQSPDTLSVSPGETVTLSCRASQNINKNLAWYQYKPGQSPRLVIFETYSKIAAFPARFVASGSGTEFTLTINNMQSEDVAVYYCQQYEEWPR--TFGQGTKVDIK
PGT136 L     EIVMTQSPPTLSVSPGETATLSCRASQNVKNNLAWYQLKPGQAPRLLIFDASSRAGGIPDRFSGSGYGTDFTLTVNSVQSEDFGDYFCQQYEEWPR--TFGQGTKVDIK
PGT137 L     EIMMTQSPAILSVSPGDRATLSCRASQSVKNNLAWYQKRPGQAPRLLIFDTSSRASGIPARFSGGGSGTEFTLTVNSMQSEDFATYYCQQYEEWPR--TFGQGTKVEIK

107548       EIVMTQSPAVLSVSPGDRATLSCRASQSVKNNLAWYQKRPGQAPRLLIFDTSSRASNIPARFSGGGSGTEFTLTVNSMQSEDFATYYCQQYEEWPR--TFGQGTKVEII
219622       QIILTQSPVTLSVSPGERVTLSCTASQSVRDRLVWLQQKPGQAPRVVIYGANGRATGVPDRFSGTGSGTEFSLTINNMQPEDSAIYYCQQYDQWPP-WAFGQGTKVEIK
210137       EIVMTQSPVTLSVSAGERVTLSCRASQNINRHLAWYQHKPPQAPRLLIYDTSRRATGIPARFSGSGSGTDFTLTIDSIQPEESAVYYCQQYDEWPPKTTFGLGTKLEIK
215528       ETILTQSPATLSVSPGDRATLSCKASQSLNNNLAWYQLKPGQAPRLVIYGAFTRATDIPVRFAGFGSGTEFSLTISSVQPEDFAIYYCQQYNEWPR--TFGQGTRVDIK
425756       EIVMTQSPPTLSVSPGETATLSCRASQNVKNNLAWYQLKPGQAPRLLIFDASSRAGGIPDRFSGSGYGTDFTLTVNSVQSEDFGDYFCQQYEEWPR--TFGQGTKVDLR
121553       EIVLTQSPPTLSVSPGEAATLSCRASQSVKNNLAWYHLKAGQAPRLLIFDASSRAGDAPARVSGSGYGTEFTLTVNSVQSEDFGDYYCQQYEEWPR--TFGQGTKVDLR
303540       EIVMTQSPAVLSVSPGDRATLSCRASQSVKNNLAWYQKRPGQAPRLLIFDTSSRASNIPARFSGGGSGTEFTLTVNSMQSEDFATYYCQQYEEWPR--TFGQGTKVEIN
378597       QIVLTQSPPTLSLSPGERATLSCRASQSVKNNLAWYQKRPGQAPRLLIFDTSSRASNIPARFSGGGSGTEFTLTVNSMQSEDFATYYCQQYEEWPR--TFGQGTKVEII
521298       EIMMTQSPAILSVSPGDRATLSCRASQSVKNNLAWYQKRPGQAPRLLIFDTSSRASGIPARFSGGGSGTEFTLTVNSMQSEDFATYYCQQYGSSPY--TFGQGTKLEIK
537707       EIVMTQSPAILSVSPGDRATFSCRASQSVKNNLAWYQKRPGQAPRLLIFDASSRASGIPARFSGGGSGTEFTLTVNSMQSEDFATYYCQQYEEWPR--TFGQGTKVEIM
```

FIGURE 8 | Sequence selection for functional characterization of light chains from donor 39. (A) Divergence/identity analysis of 10 light-chain variable-domain sequences obtained from the clustering analysis of 72 sequences identified by intra-donor phylogenetic analysis. Sequences of IGKV3-15*01 origin are plotted as a function of sequence identity to PGT137 light chain and sequence divergence from inferred germline allele, with 10 selected sequences shown as black triangles. **(B)** Percent population of 10 clusters obtained using a sequence identity cutoff of 95%. Each cluster is indicated by its representative sequence. **(C)** Protein sequences of 10 cluster representatives. Sequences are aligned to the inferred germline gene, IGKV3-15*01. Framework regions (FR) and complementarity-determining regions (CDRs) are based on Kabat nomenclature. Amino acids mutated from the germline gene are shown in red.

systems-level bioinformatics approaches, which enable these data to be mined effectively. Overall, data-intensive methods may be generally required to obtain true insight into questions of biological diversity such as the humoral immune response.

Prior next-generation sequencing and bioinformatics analyses have revealed the extraordinary genetic diversity of HIV-1 (Eriksson et al., 2008; Archer et al., 2009; Tsibris et al., 2009; Fischer et al., 2010). These same methods are now beginning to reveal the extraordinary diversity of antibodies generated in response to HIV-1 infection (Wu et al., 2011). Although this response appears to provide little benefit to the HIV-1-infected host (Poignard et al., 1999), if similar responses could be generated through vaccination, then in principle effective protection could be achieved in the setting of initial infection (Burton, 2002; Burton et al., 2004, 2005). The populations of antibodies we identify here may provide broader protection than a monoclonal member of the group. Furthermore, responses to infection or vaccination would be expected to generate diverse populations of antibodies, as we have shown here. Thus, population diversity, even within a single antibody clone or lineage, is likely to have a substantial impact on the effectiveness of the immune response.

DATA DEPOSITION

Next-generation sequencing data from donor 39 (heavy and light chains) and also for the 10 plasmid control have been deposited in the National Center for Biotechnology Information Short Reads Archives (SRA) under accession no. SRA055820. Information deposited with GenBank includes the heavy- and light-chain variable region sequences of genomically identified neutralizers: 10 heavy chains, gVRC-H1-10_{d39} (JX313021-30), amino-acid

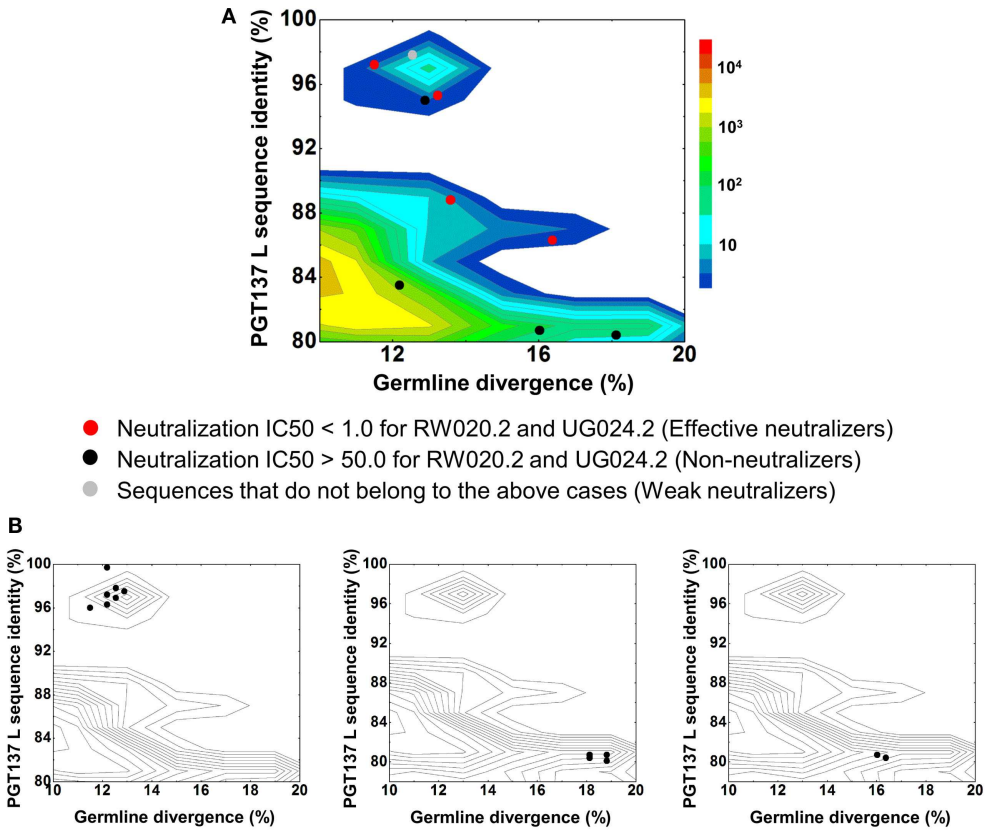


FIGURE 9 | Divergence/identity analysis of light-chain neutralization. (A) The expressed light-chain sequences color-coded based on the neutralization potency of reconstituted antibodies, with IC50 <1.0 for both viruses shown in red (effective neutralizers), IC50 >50.0 for both viruses in black (non-neutralizers), and other cases in gray (weak neutralizers). **(B)** The three largest clusters are displayed on the enlarged divergence/identity plot, with 45, 6, and 4 members, respectively.

consensus heavy-chain gVRC-H11$_{d39}$ (JX444560), and 6 light chains, gVRC-L1-6 $_{d39}$ (JX313030-36).

ACKNOWLEDGMENTS

We thank H. Coleman, M. Park, B. Schmidt, and A. Young for 454 pyrosequencing at the NIH Intramural Sequencing Center (NISC), J. Stuckey for assistance with figures. We also thank members of the Structural Biology Section and Structural Bioinformatics Core, Vaccine Research Center, for discussions or comments on the manuscript. We would like to thank all the study participants and research staff at each of the Protocol G clinical centers, and all of the Protocol G team members, the IAVI Human Immunology Laboratory, and all of the Protocol G clinical investigators, specifically, George Miiro, Anton Pozniak, Dale McPhee, Olivier Manigart, Etienne Karita, Andre Inwoley, Walter Jaoko, Jack DeHovitz, Linda-Gail Bekker, Punnee Pitisuttithum, Robert Paris, Jennifer Serwanga, and Susan Allen. Support for this work was provided by the Intramural Research Program of the Vaccine Research Center, National Institute of Allergy and Infectious Diseases and the National Human Genome Research Institute, National Institutes of Health, and by grants from the International AIDS Vaccine Initiative's Neutralizing Antibody Consortium.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found on line at http://www.frontiersin.org/Virology/10.3389/fmicb.2012.00315/abstract

Figure S1 | Pipeline processing of heavy-chain sequences of 10 plasmid antibodies determined by 454 pyrosequencing.

Figure S2 | Pyrosequencing-induced sequence variation for 10 plasmid antibodies after being processed by an error-correction procedure.

Figure S3 | Pipeline processing of donor 39 heavy-chain sequences determined by 454 pyrosequencing.

Figure S4 | Pipeline processing of donor 39 light-chain sequences determined by 454 pyrosequencing.

Table S1 | PCR primers and DNA polymerase systems used to prepare samples for 454 pyrosequencing.

Table S2 | Neutralization of reconstituted antibodies by pairing clustering-selected heavy-chain sequences from 454 pyrosequencing with PGT137 light chain.

Table S3 | Neutralization of reconstituted antibodies by pairing clustering-selected light-chain sequences from 454 pyrosequencing with PGT137 heavy chain.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Archer, J., Braverman, M. S., Taillon, B. E., Desany, B., James, I., Harrigan, P. R., Lewis, M., and Robertson, D. L. (2009). Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *AIDS* 23, 1209–1218.

Balazs, A. B., Chen, J., Hong, C. M., Rao, D. S., Yang, L., and Baltimore, D. (2011). Antibody-based protection against HIV infection by vectored immunoprophylaxis. *Nature* 481, 81–84.

Barouch, D. H., and Nabel, G. J. (2005). Adenovirus vector-based vaccines for human immunodeficiency virus type 1. *Hum. Gene Ther.* 16, 149–156.

Boyd, S. D., Gaeta, B. A., Jackson, K. J., Fire, A. Z., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., and Collins, A. M. (2010). Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* 184, 6986–6992.

Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36, W503–W508.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* 18, 763–770.

Burton, D. R. (2002). Antibodies, viruses and vaccines. *Nat. Rev. Immunol.* 2, 706–713.

Burton, D. R., Desrosiers, R. C., Doms, R. W., Koff, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. (2004). HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.* 5, 233–236.

Burton, D. R., Stanfield, R. L., and Wilson, I. A. (2005). Antibody vs. HIV in a clash of evolutionary titans. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14943–14948.

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D., Euskirchen, G., Lacroute, P., Bettinger, K., Boyle, A. P., Kasowski, M., Grubert, F., Seki, S., Garcia, M., Whirl-Carrillo, M., Gallardo, M., Blasco, M. A., Greenberg, P. L., Snyder, P., Klein, T. E., Altman, R. B., Butte, A. J., Ashley, E. A., Gerstein, M., Nadeau, K. C., Tang, H., and Snyder, M. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293–1307.

Corti, D., Langedijk, J. P., Hinz, A., Seaman, M. S., Vanzetta, F., Fernandez-Rodriguez, B. M., Silacci, C., Pinna, D., Jarrossay, D., Balla-Jhagjhoorsingh, S., Willems, B., Zekveld, M. J., Dreja, H., O'Sullivan, E., Pade, C., Orkin, C., Jeffs, S. A., Montefiori, D. C., Davis, D., Weissenhorn, W., McKnight, A., Heeney, J. L., Sallusto, F., Sattentau, Q. J., Weiss, R. A., and Lanzavecchia, A. (2010). Analysis of memory B cell responses and isolation of novel monoclonal antibodies with neutralizing breadth from HIV-1-infected individuals. *PLoS ONE* 5, e8805. doi:10.1371/journal.pone.0008805

Doria-Rose, N. A., Klein, R. M., Daniels, M. G., O'Dell, S., Nason, M., Lapedes, A., Bhattacharya, T., Migueles, S. A., Wyatt, R. T., Korber, B. T., Mascola, J. R., and Connors, M. (2010). Breadth of human immunodeficiency virus-specific neutralizing activity in sera: clustering analysis and association with clinical variables. *J. Virol.* 84, 1631–1636.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W., and Beerenwinkel, N. (2008). Viral population estimation using pyrosequencing. *PLoS Comput. Biol.* 4, e1000074. doi:10.1371/journal.pcbi.1000074

Fischer, W., Ganusov, V. V., Giorgi, E. E., Hraber, P. T., Keele, B. F., Leitner, T., Han, C. S., Gleasner, C. D., Green, L., Lo, C.-C., Nag, A., Wallstrom, T. C., Wang, S., McMichael, A. J., Haynes, B. F., Hahn, B. H., Perelson, A. S., Borrow, P., Shaw, G. M., Bhattacharya, T., and Korber, B. T. (2010). Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5, e12303. doi:10.1371/journal.pone.0012303

Gnanakaran, S., Daniels, M. G., Bhattacharya, T., Lapedes, A. S., Sethi, A., Li, M., Tang, H., Greene, K., Gao, H., Haynes, B. F., Cohen, M. S., Shaw, G. M., Seaman, M. S., Kumar, A., Gao, F., Montefiori, D. C., and Korber, B. (2010). Genetic signatures in the envelope glycoproteins of HIV-1 that associate with broadly neutralizing antibodies. *PLoS Comput. Biol.* 6, e1000955. doi:10.1371/journal.pcbi.1000955

Gonda, M. A., Wongstaal, F., Gallo, R. C., Clements, J. E., Narayan, O., and Gilden, R. V. (1985). Sequence homology and morphologic similarity of HTLV-III and Visna virus, a pathogenic lenti-virus. *Science* 227, 173–177.

Gray, E. S., Taylor, N., Wycuff, D., Moore, P. L., Tomaras, G. D., Wibmer, C. K., Puren, A., Decamp, A., Gilbert, P. B., Wood, B., Montefiori, D. C., Binley, J. M., Shaw, G. M., Haynes, B. F., Mascola, J. R., and Morris, L. (2009). Antibody specificities associated with neutralization breadth in plasma from human immunodeficiency virus type 1 subtype C-infected blood donors. *J. Virol.* 83, 8925–8937.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11, 476–486.

Hessell, A. J., Poignard, P., Hunter, M., Hangartner, L., Tehrani, D. M., Bleeker, W. K., Parren, P. W., Marx, P. A., and Burton, D. R. (2009a). Effective, low-titer antibody protection against low-dose repeated mucosal SHIV challenge in macaques. *Nat. Med.* 15, 951–954.

Hessell, A. J., Rakasz, E. G., Poignard, P., Hangartner, L., Landucci, G., Forthal, D. N., Koff, W. C., Watkins, D. I., and Burton, D. R. (2009b). Broadly neutralizing human anti-HIV antibody 2G12 is effective in protection against mucosal SHIV challenge even at low serum neutralizing titers. *PLoS Pathog.* 5, e1000433. doi:10.1371/journal.ppat.1000433

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8, 460. doi:10.1186/1471-2105-8-460

Kahn, S. D. (2011). On the future of genomic data. *Science* 331, 728–729.

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664.

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.

Kuhner, M. K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468.

Kwong, P. D., Mascola, J. R., and Nabel, G. J. (2011). Rational Design of vaccines to elicit broadly neutralizing antibodies to HIV-1. *Cold Spring Harb. Perspect. Med.* 1, a007278.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A.,

Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948.

Li, M., Gao, F., Mascola, J. R., Stamatatos, L., Polonis, V. R., Koutsoukos, M., Voss, G., Goepfert, P., Gilbert, P., Greene, K. M., Bilska, M., Kothe, D. L., Salazar-Gonzalez, J. F., Wei, X., Decker, J. M., Hahn, B. H., and Montefiori, D. C. (2005). Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* 79, 10108–10125.

Li, Y., Migueles, S. A., Welcher, B., Svehla, K., Phogat, A., Louder, M. K., Wu, X., Shaw, G. M., Connors, M., Wyatt, R. T., and Mascola, J. R. (2007). Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nat. Med.* 13, 1032–1034.

Mardis, E. R. (2008a). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141.

Mardis, E. R. (2008b). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402.

Mascola, J. R. (2003). Defining the protective antibody response for HIV-1. *Curr. Mol. Med.* 3, 209–216.

Mascola, J. R., Lewis, M. G., Stiegler, G., Harris, D., Vancott, T. C., Hayes, D., Louder, M. K., Brown, C. R., Sapan, C. V., Frankel, S. S., Lu, Y., Robb, M. L., Katinger, H., and Birx, D. L. (1999). Protection of macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. *J. Virol.* 73, 4009–4018.

Mascola, J. R., Stiegler, G., Vancott, T. C., Katinger, H., Carpenter, C. B., Hanson, C. E., Beary, H., Hayes, D., Frankel, S. S., Birx, D. L., and Lewis, M. G. (2000). Protection of macaques against vaginal transmission of a pathogenic HIV- 1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nat. Med.* 6, 207–210.

Parren, P. W., Marx, P. A., Hessell, A. J., Luckay, A., Harouse, J., Cheng-Mayer, C., Moore, J. P., and Burton, D. R. (2001). Antibody protects macaques against vaginal challenge with a pathogenic R5 simian/human immunodeficiency virus at serum levels giving complete neutralization in vitro. *J. Virol.* 75, 8340–8347.

Parren, P. W., Moore, J. P., Burton, D. R., and Sattentau, Q. J. (1999). The neutralizing antibody response to HIV-1: viral evasion and escape from humoral immunity. *AIDS* 13, S137–S162.

Poignard, P., Sabbe, R., Picchio, G. R., Wang, M., Gulizia, R. J., Katinger, H., Parren, P. W., Mosier, D. E., and Burton, D. R. (1999). Neutralizing antibodies have limited effects on the control of established HIV-1 infection in vivo. *Immunity* 10, 431–438.

Prabakaran, P., Streaker, E., Chen, W., and Dimitrov, D. S. (2011). 454 antibody sequencing – error characterization and correction. *BMC Res. Notes* 4, 404. doi:10.1186/1756-0500-4-404

Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science* 242, 1168–1171.

Reddy, S. T., Ge, X., Miklos, A. E., Hughes, R. A., Kang, S. H., Hoi, K. H., Chrysostomou, C., Hunicke-Smith, S. P., Iverson, B. L., Tucker, P. W., Ellington, A. D., and Georgiou, G. (2009). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28, U965–U969.

Reddy, S. T., and Georgiou, G. (2011). Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Curr. Opin. Biotechnol.* 22, 584–589.

Sather, D. N., Armann, J., Ching, L. K., Mavrantoni, A., Sellhorn, G., Caldwell, Z., Yu, X., Wood, B., Self, S., Kalams, S., and Stamatatos, L. (2009). Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *J. Virol.* 83, 757–769.

Scheid, J. F., Mouquet, H., Feldhahn, N., Seaman, M. S., Velinzon, K., Pietzsch, J., Ott, R. G., Anthony, R. M., Zebroski, H., Hurley, A., Phogat, A., Chakrabarti, B., Li, Y., Connors, M., Pereyra, F., Walker, B. D., Wardemann, H., Ho, D., Wyatt, R. T., Mascola, J. R., Ravetch, J. V., and Nussenzweig, M. C. (2009). Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* 458, 636–640.

Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D. R., Pereyra, F., Ho, D. D., Walker, B. D., Seaman, M. S., Bjorkman, P. J., Chait, B. T., and Nussenzweig, M. C. (2011). Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333, 1633–1637.

Seaman, M. S., Janes, H., Hawkins, N., Grandpre, L. E., Devoy, C., Giri, A., Coffey, R. T., Harris, L., Wood, B., Daniels, M. G., Bhattacharya, T., Lapedes, A., Polonis, V. R., McCutchan, F. E., Gilbert, P. B., Self, S. G., Korber, B. T., Montefiori, D. C., and Mascola, J. R. (2010). Tiered categorization of a diverse panel of HIV-1 Env pseudoviruses for neutralizing antibody assessment. *J. Virol.* 84, 1439–1452.

Simek, M. D., Rida, W., Priddy, F. H., Pung, P., Carrow, E., Laufer, D. S., Lehrman, J. K., Boaz, M., Tarragona-Fiol, T., Miiro, G., Birungi, J., Pozniak, A., McPhee, D. A., Manigart, O., Karita, E., Inwoley, A., Jaoko, W., Dehovitz, J., Bekker, L. G., Pitisuttithum, P., Paris, R., Walker, L. M., Poignard, P., Wrin, T., Fast, P. E., Burton, D. R., and Koff, W. C. (2009). Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. *J. Virol.* 83, 7337–7348.

Sonigo, P., Alizon, M., Staskus, K., Klatzmann, D., Cole, S., Danos, O., Retzel, E., Tiollais, P., Haase, A., and Wainhobson, S. (1985). Nucleotide-sequence of the visna lentivirus – relationship to the AIDS virus. *Cell* 42, 369–382.

Stamatatos, L., Morris, L., Burton, D. R., and Mascola, J. R. (2009). Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nat. Med.* 15, 866–870.

Starcich, B. R., Hahn, B. H., Shaw, G. M., McNeely, P. D., Modrow, S., Wolf, H., Parks, E. S., Parks, W. P., Josephs, S. F., Gallo, R. C., and Wongstaal, F. (1986). Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell* 45, 637–648.

Tsibris, A. M. N., Korber, B., Arnaout, R., Russ, C., Lo, C.-C., Leitner, T., Gaschen, B., Theiler, J., Paredes, R., Su, Z., Hughes, M. D., Gulick, R. M., Greaves, W., Coakley, E., Flexner, C., Nusbaum, C., and Kuritzkes, D. R. (2009). Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS ONE* 4, e5683. doi:10.1371/journal.pone.0005683

UNAIDS. (2010). *Joint United Nations Programme on HIV/AIDS. UN Report on the Global AIDS Epidemic 2010.* Available at: http://www.unaids.org/globalreport/Global_report.htm

Veazey, R. S., Shattock, R. J., Pope, M., Kirijan, J. C., Jones, J., Hu, Q., Ketas, T., Marx, P. A., Klasse, P. J., Burton, D. R., and Moore, J. P. (2003). Prevention of virus transmission to macaque monkeys by a vaginally applied monoclonal antibody to HIV-1 gp120. *Nat. Med.* 9, 343–346.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M.,

Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. Q. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J. H., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, C., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z. M., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W. M., Gong, F. C., Gu, Z. P., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z. X., Ketchum, K. A., Lai, Z. W., Lei, Y. D., Li, Z. Y., Li, J. Y., Liang, Y., Lin, X. Y., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B. X., Sun, J. T., Wang, Z. Y., Wang, A. H., Wang, X., Wang, J., Wei, M. H., Wides, R., Xiao, C. L., Yan, C. H., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science* 291, 1304–1351.

Walker, L. M., and Burton, D. R. (2010). Rational antibody-based HIV-1 vaccine design: current approaches and future directions. *Curr. Opin. Immunol.* 22, 358–366.

Walker, L. M., Huber, M., Doores, K. J., Falkowska, E., Pejchal, R., Julien, J. P., Wang, S. K., Ramos, A., Chan-Hui, P. Y., Moyle, M., Mitcham, J. L., Hammond, P. W., Olsen, O. A., Phung, P., Fling, S., Wong, C. H., Phogat, S., Wrin, T., Simek, M. D., Koff, W. C., Wilson, I. A., Burton, D. R., and Poignard, P. (2011). Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 477, 466–470.

Walker, L. M., Phogat, S. K., Chan-Hui, P. Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. A., Frey, S. M., Hammond, P. W., Kaminsky, S., Zamb, T., Moyle, M., Koff, W. C., Poignard, P., and Burton, D. R. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 326, 285–289.

Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X., Salazar-Gonzalez, J. F., Salazar, M. G., Kilby, J. M., Saag, M. S., Komarova, N. L., Nowak, M. A., Hahn, B. H., Kwong, P. D., and Shaw, G. M. (2003). Antibody neutralization and escape by HIV-1. *Nature* 422, 307–312.

Wu, X., Yang, Z. Y., Li, Y., Hogerkorp, C. M., Schief, W. R., Seaman, M. S., Zhou, T., Schmidt, S. D., Wu, L., Xu, L., Longo, N. S., McKee, K., O'Dell, S., Louder, M. K., Wycuff, D. L., Feng, Y., Nason, M., Doria-Rose, N., Connors, M., Kwong, P. D., Roederer, M., Wyatt, R. T., Nabel, G. J., and Mascola, J. R. (2010). Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* 329, 856–861.

Wu, X., Zhou, T., O'Dell, S., Wyatt, R. T., Kwong, P. D., and Mascola, J. R. (2009). Mechanism of human immunodeficiency virus type 1 resistance to monoclonal antibody B12 that effectively targets the site of CD4 attachment. *J. Virol.* 83, 10892–10907.

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N. S., Louder, M., McKee, K., O'Dell, S., Perfetto, S., Schmidt, S. D., Shi, W., Wu, L., Yang, Y., Yang, Z. Y., Yang, Z., Zhang, Z., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Doria-Rose, N. A., Connors, M., Mullikin, J. C., Nabel, G. J., Roederer, M., Shapiro, L., Kwong, P. D., and Mascola, J. R. (2011). Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333, 1593–1602.

Conflict of Interest Statement: 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.

Received: 12 June 2012; paper pending published: 04 July 2012; accepted: 13 August 2012; published online: 11 September 2012.

*Citation: Zhu J, O'Dell S, Ofek G, Pancera M, Wu X, Zhang B, Zhang Z, NISC Comparative Sequencing Program, Mullikin JC, Simek M, Burton DR, Koff WC, Shapiro L, Mascola JR and Kwong PD (2012) Somatic populations of PGT135–137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. Front. Microbio. **3**:315. doi: 10.3389/fmicb.2012.00315*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 19 July 2012
doi: 10.3389/fmicb.2012.00258

Molecular dynamics simulation in virus research

Hirotaka Ode[1,2]*, Masaaki Nakashima[1,3], Shingo Kitamura[1,3], Wataru Sugiura[1,4] and Hironori Sato[2]

[1] Clinical Research Center, National Hospital Organization Nagoya Medical Center, Nagoya, Aichi, Japan
[2] Pathogen Genomics Center, National Institute of Infectious Diseases, Musashimurayama, Tokyo, Japan
[3] Department of Biotechnology, Graduate School of Engineering, Nagoya University, Nagoya, Aichi, Japan
[4] Department of AIDS Research, Graduate School of Medicine, Nagoya University, Nagoya, Aichi, Japan

Edited by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan

Reviewed by:
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan
Yasuyuki Miyazaki, The University of Tokushima Graduate School, Japan

***Correspondence:**
Hirotaka Ode, Department of Infectious Diseases and Immunology, Clinical Research Center, National Hospital Organization Nagoya Medical Center, 4-1-1 Sannomaru, Naka-ku, Nagoya, Aichi, 460-0001, Japan.
e-mail: odehir@nih.go.jp

Virus replication in the host proceeds by chains of interactions between viral and host proteins. The interactions are deeply influenced by host immune molecules and anti-viral compounds, as well as by mutations in viral proteins. To understand how these interactions proceed mechanically and how they are influenced by mutations, one needs to know the structures and dynamics of the proteins. Molecular dynamics (MD) simulation is a powerful computational method for delineating motions of proteins at an atomic-scale via theoretical and empirical principles in physical chemistry. Recent advances in the hardware and software for biomolecular simulation have rapidly improved the precision and performance of this technique. Consequently, MD simulation is quickly extending the range of applications in biology, helping to reveal unique features of protein structures that would be hard to obtain by experimental methods alone. In this review, we summarize the recent advances in MD simulations in the study of virus–host interactions and evolution, and present future perspectives on this technique.

Keywords: MD simulation, viral protein, three-dimensional structure, protein dynamics, coarse-grained MD

INTRODUCTION

Proteins fluctuate spontaneously in solution (Ishima and Torchia, 2000). Accumulating evidence indicates that such fluctuations play key roles in the specific functions of proteins, such as catalytic reactions of enzymes (Nicholson et al., 1995; Lu et al., 1998; Eisenmesser et al., 2005; Henzler-Wildman et al., 2007; Abbondanzieri et al., 2008), interactions with other biomolecules (Thorpe and Brooks, 2007), and biomolecular motors and pumps (Astumian, 1997). Multiple experimental methods are available to characterize the protein dynamics (**Figure 1**). However, it is usually difficult to delineate motions of proteins at an atomic scale.

MD SIMULATION IN BIOLOGY

OUTLINE

Molecular dynamics (MD) simulation is a computational method to address the above issue (**Figure 1**) (Henzler-Wildman and Kern, 2007; Dror et al., 2010). This technique enables us to calculate movements of atoms in a molecular system, such as proteins in water, by numerically solving Newton's equations of motions (Karplus and Petsko, 1990; Adcock and McCammon, 2006). In a simple molecular system, all atoms and covalent bonds connecting the atoms are assumed to be the charged spheres and springs, respectively. Parameters of mathematical functions describing the potential energy of a system, termed the "force field," are set to simulate the movements of atoms and molecules. Frequently used force fields for proteins, such as the "AMBER" (Pearlman et al., 1995; Case et al., 2005) and "CHARMM" (Brooks et al., 2009) force fields, have the formulae of covalent bonds, angles, dihedrals, van der Waals, and electrostatic potentials.

PERFORMANCE AND CONSISTENCY WITH EXPERIMENTAL DATA

Application of MD simulation in the field of protein chemistry was first reported in 1977 (McCammon et al., 1977). Since then, the performance of this technique have been quickly improved quantitatively and qualitatively along with the rapid advances in hardware and software on biomolecular simulation (Lindorff-Larsen et al., 2012). The results of MD simulation are critically influenced by the force fields (Lindorff-Larsen et al., 2012). The qualities of parameters in the force fields, especially for dihedrals and electrostatic potentials, have been improved quantitatively and qualitatively over time by introducing improved approximation to the quantum ground-state potential energy surface. Recently, eight different protein force fields were evaluated on the basis of the consistency of simulations with the NMR data (Lindorff-Larsen et al., 2012). The study demonstrates that the most recent versions, while not perfect, provide results that are highly consistent with the experimental data (Lindorff-Larsen et al., 2012). In addition, explicit introduction of effects of the solvation has contributed to the qualitative improvement for the precision and performance of MD simulations (Adcock and McCammon, 2006).

MD IN STRUCTURAL BIOLOGY

MD simulation currently allows us to investigate the structural dynamics of proteins on timescales of nanoseconds to microseconds, and will probably allow investigation to milliseconds in the future (**Figure 1**) (Henzler-Wildman and Kern, 2007; Dror et al., 2010). This technique is widely used in the field of structural biology (Karplus and McCammon, 2002; Karplus and Kuriyan, 2005; Dodson et al., 2008). First, MD simulation is useful

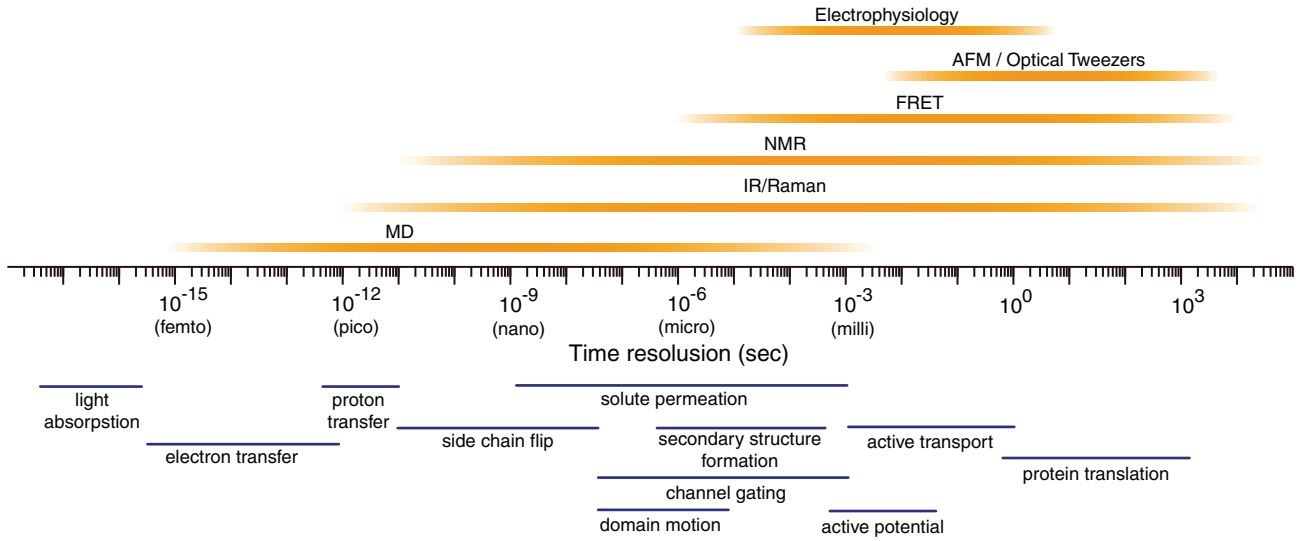


FIGURE 1 | Temporal resolution of various biophysical techniques. The timescales of some fundamental atom- or molecule-scale motions are shown below. AFM, atomic force microscopy; FRET, fluorescence resonance energy transfer; IR, infrared radiation; NMR, nuclear magnetic resonance.

for refining the experimentally determined three-dimensional (3-D) structures of proteins (Autore et al., 2010; Ozen et al., 2011). Second, MD simulation is beneficial for constructing previously undescribed 3-D structures of proteins in combination with homology modeling techniques (Marti-Renom et al., 2000; Sanchez et al., 2000; Baker and Sali, 2001), when a reported structure of a homolog is available. Third and most importantly, MD simulation provides a unique tool to address the structural dynamics of proteins, i.e., the time evolution of conformations in solution, at timescales of nanoseconds to microseconds (Henzler-Wildman and Kern, 2007; Dror et al., 2010). The structural snapshots obtained during MD simulation are helpful for depicting the unique structural features of proteins (Karplus and McCammon, 2002; Karplus and Kuriyan, 2005; Dodson et al., 2008).

MD SIMULATION IN VIROLOGY

To date, MD simulations have been applied in a range of virus researches, as shown in the following sections.

NEUTRALIZATION ESCAPE AND CELL TROPISM SWITCHING OF HIV-1 MEDIATED BY AN ELECTROSTATIC MECHANISM

It is very important to clarify how viruses evade neutralization antibodies in order to understand the viral life cycle and evolution, and to develop vaccines. MD simulation is used to address this issue as it pertains to human immunodeficiency virus type 1 (HIV-1). The third variable (V3) loop of the HIV-1 envelope gp120 protein constitutes the major antibody epitopes of HIV-1 and the major determinants for the entry coreceptor use of HIV-1. By analyzing the 40,000 structural snapshots obtained from 10–30 ns of MD simulations of the identical gp120 outer domain carrying a distinct V3 loop with net charge of +3 or +7, Yokoyama and colleagues showed that the change in V3 net charge alone is sufficient to induce global changes in fluctuation and conformation of the loops involved in binding to CD4, coreceptor, and neutralizing antibodies (Naganawa et al., 2008; Yokoyama et al., 2012). Structural changes caused by a reduction in the V3 net charge via V3 mutations are tightly linked to viral CCR5 coreceptor tropism (Naganawa et al., 2008), as well as to a reduction in viral neutralization sensitivity to anti-V3 antibodies (Naganawa et al., 2008) and anti-CD4 binding site monoclonal antibodies (Yokoyama et al., 2012). These findings suggest a hitherto unrecognized mechanism, V3-mediated electrostatic modulation of the structure and dynamics of the gp120 interaction surface, for adjusting the relative replication fitness and evolution of HIV-1 (Yokoyama et al., 2012). In addition, they partly explain a virological mystery, i.e., why HIV-1 variants using CCR5, which carries a V3 loop with a lower level of positive net charge, predominantly persist before the onset of AIDS.

MECHANISMS OF VIRAL ESCAPE FROM HOST DEFENSE SYSTEMS

Viruses also evade host defense systems other than neutralization antibodies (**Figure 2**). MD simulation is used to clarify the structural basis for viral escape from host defense systems by mutations. Mutations at the 120th amino acid in the HIV-2 capsid protein play a key role in evading tripartite motif-containing protein 5α (TRIM5α), an anti-retroviral cellular protein induced by interferon, both *in vivo* (Onyango et al., 2010) and *in vitro* (Song et al., 2007). An MD simulation study has revealed that the mutations could extensively influence the conformation and fluctuation of the interaction surface of capsid proteins by altering the probability of hydrogen bond formation between helices 4 and 5 (Miyamoto et al., 2011).

HIV-1 Vpu antagonizes an antiviral cellular protein termed tetherin, also known as BST-2/CD317/HM1.24, by interaction with the transmembrane (TM) domain of tetherin and subsequent degradation (Douglas et al., 2010; Kobayashi et al., 2011). An MD simulation suggests that alignment of the four

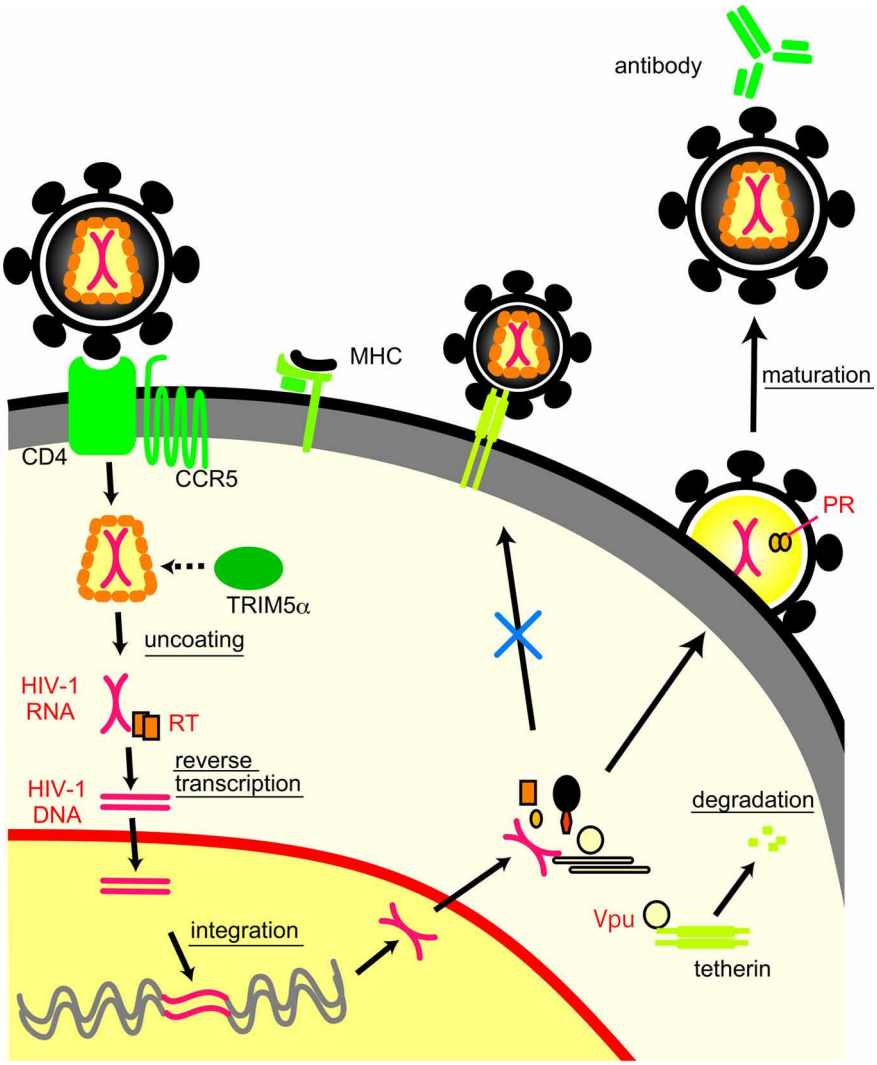


FIGURE 2 | Life cycle of HIV-1 and interactions between viral proteins and host immune molecules.

amino acid residues (I34, L37, L41, and T45) on the same helical face in the human tetherin TM domain is crucial for the Vpu-mediated antagonism against human tetherin (Kobayashi et al., 2011). The interface structure of the tetherin TM for the antagonism was also predicted by the MD simulation of another group (Zhou et al., 2012) and experimentally confirmed by an NMR study (Skasko et al., 2012).

MD simulation is also used to study the mechanisms of functional interactions between cytotoxic T lymphocyte (CTL) epitope and major histocompatibility complex (MHC) molecules (Reboul et al., 2012). An MD simulation study has revealed that a 13-mer epitope peptide from Epstein-Barr virus has the low structural flexibility in an MHC molecule that induces a CTL response but exhibits high flexibility in another MHC molecule that cannot induce a CTL response (Reboul et al., 2012). Thus, structural flexibility of CTL epitope region seems to be critical for the specific recognition by MHC molecules, and mutations that alter the flexibility may influence CTL response. There are other viral proteins and immune molecules involved in viral evasion from host defense systems (Neil and Bieniasz, 2009; Malim and Bieniasz, 2012). MD simulations should also be applicable for the studies of these molecules.

STRUCTURE AND FUNCTION OF VIRAL ENZYMES

Viral enzymes are essential for viral replications and thus are important targets for anti-viral drug development. MD simulations are used to study the basis of the structural dynamics that allow the viral enzyme and its drug to function properly. Viral polymerase (Pol) is essential for viral genome replication in the viral life cycle. The Pol is composed of the fingers, palm, and thumb subdomains, which form a cavity for the substrate binding, as in eukaryotic Pol (Joyce and Steitz, 1994; Lamers et al., 2006; Cameron et al., 2009). MD simulations suggest that the finger and thumb domains of HIV-1 reverse transcriptase (RT) are especially mobile among the various regions of this enzyme (Zhou et al., 2005; Kirmizialtin et al., 2012). The mobility

is severely attenuated by binding of allosteric non-nucleotide RT inhibitors (NNRTIs) (Zhou et al., 2005). Interestingly, a large conformational change of RT subdomains during millisecond timescale simulations can lock the correct nucleotide at the active site but promotes release of a mismatched nucleotide (Kirmizialtin et al., 2012). Furthermore, conformational dynamics leading to opening and closing motions of the substrate binding cleft are highly conserved among four RNA Pols in the picornavirus family, despite the amino acid identity being as low as 30–74% (Moustafa et al., 2011). These findings are consistent with each other and strongly suggest that the structural dynamics of viral Pol play a key role in the polymerization.

Viral protease (PR) plays a key role in viral propagation by catalyzing cleavages of viral precursor proteins (Pettit et al., 1994, 2002; Steven et al., 2005). HIV-1 PR and other retroviral PRs have unique regions termed the "flaps" outside the substrate binding clefts (Dunn et al., 2002). MD simulation studies suggest that the PR flaps in HIV-1 are intrinsically mobile, undergoing conversions between the "semiopen," "open," and "closed" conformations (Hornak et al., 2006; Deng et al., 2011). This movement is severely attenuated upon placement of the substrate or PR inhibitor in the binding cleft (Karthik and Senapati, 2011), suggesting that flap movement plays a critical role in PR function.

MD simulations are also used to study the structural dynamics of the substrates of viral PR. Peptides corresponding to cleavage junctions of viral precursor proteins of HIV-1 are intrinsically unstructured in aqueous solution (Datta et al., 2011; Ode et al., 2011). However, the folding preference of the junction peptides may be different among the junctions and related to the efficiency of substrate binding and cleavage reaction by PR (Ode et al., 2011). Furthermore, peptides at the capsid-p2 junction can adopt a helical conformation when the polarity of the environment is reduced (Datta et al., 2011). The MD simulation of PR and its substrates will help to clarify how the viral precursor is processed orderly during viral maturation.

DRUG-RESISTANCE MECHANISMS

Antiviral drug resistance is a major clinical problem for the treatment of virus-infected individuals (Cortez and Maldarelli, 2011; van der Vries et al., 2011). Viral resistance to antiviral drugs is primarily caused by genetic mutations that eventually lead to a reduction in the drug affinity of drug target viral proteins. MD simulations are used to examine how viral mutations cause the drug resistance at the atomic level.

A reduction in the binding affinity of the PR inhibitors to HIV-1 PR can be caused by a reduction in hydrophobic interactions (Kagan et al., 2005; Wittayanarakul et al., 2005; Sadiq et al., 2007; Chen et al., 2010; Dirauf et al., 2010), reduction in electrostatic interactions (Ode et al., 2005, 2006, 2007a; Chen et al., 2010), changes in flexibility at the flap of the PR (Piana et al., 2002; Perryman et al., 2004; Chang et al., 2006; Foulkes-Murzycki et al., 2007), and changes in the shape of the inhibitor-binding pocket (Ode et al., 2005, 2006, 2007b). Reduction in binding affinity of the nucleotide/nucleoside RT inhibitors (NRTIs) to HIV-1 RT can be caused by a distinct conformational preference of NRTIs in the substrate/NRTI-binding site compared to normal substrates (Carvalho et al., 2006) or enhancement of ATP-mediated excision of misincorporated nucleotide analogs via increased accessibility of ATP to the terminus of extending DNA (White et al., 2004; Carvalho et al., 2007). Reduction in the binding affinity of the NNRTIs to HIV-1 RT can be attained by occlusion of the NNRTI-entry pathway (Rodriguez-Barrios and Gago, 2004; Rodriguez-Barrios et al., 2005) or restoration of the proper flexibility of the RT even with NNRTIs (Zhou et al., 2005).

A change in volume of the binding site of influenza virus (IFV) M2 channel blockers has been shown to reduce the blockers' binding affinity (Gu et al., 2011; Leonov et al., 2011; Wang et al., 2011). Disruption of the proper guidance of IFV neuraminidase (NA) inhibitors into their binding pocket is proposed as a possible mechanism for the reduction in the binding affinity of the inhibitors (Le et al., 2010; Kasson, 2012). MD simulations are also used to study how the genetic differences of HIV variants around the world can influence the efficacy of antiviral inhibitors (Batista et al., 2006; Ode et al., 2007a; Matsuyama et al., 2010; Soares et al., 2010; Kar and Knecht, 2012). Thus, MD simulation will be valuable to assist in the study of drug efficacy when genetic information on the drug target proteins is available (Shenderovich et al., 2003; Stoica et al., 2008; Sadiq et al., 2010; Wright and Coveney, 2011).

ANTIVIRAL DRUG DISCOVERY AND DEVELOPMENT

MD simulations are used to assist in the discovery and development of antiviral drugs (Durrant and McCammon, 2011; Borhani and Shaw, 2012). MD simulations allow sampling snapshots of fluctuated protein structures, which include their short-lived conformations as well as stable conformations. This is beneficial for searching conformations of a protein on ligand-binding, since ligand-binding can stabilize conformation of a protein that is not the most stable at ligand-free state (Tobi and Bahar, 2005; Xu et al., 2008). Thus, the MD simulations are used to improve the enrichment performance of molecular docking during *in silico* drug screening by taking accounts of multiple docking poses (Okimoto et al., 2009). The method is also applied for identifying concealed drug-binding sites, which are apparently masked and not evident from the X-ray crystal structures, by considering the structural flexibility of proteins. For example, MD simulations have been used to find a trench adjacent to the active site of HIV-1 integrase (Schames et al., 2004). A site-directed mutagenesis study provided evidence that the trench indeed plays key roles in ligand-binding (Lee and Robinson, 2006). These findings have been used to design HIV-1 integrase inhibitors with potent antiviral effects (Durrant and McCammon, 2011).

Likewise, MD simulations are used to assist in the development of antiviral drugs against IFV. Using this method, a universal cavity adjacent to the binding site of natural substrate has been reported with NA proteins of human 2009 pandemic H1N1, avian H5N1, and human H2N2 strains (Amaro et al., 2011). MD simulations were also used to construct a 3-D structure model of CCR5, a major coreceptor of HIV-1 (Maeda et al., 2008; Da and Wu, 2011).

VIRION STRUCTURE

It is essential to clarify the structure of virions in order to understand the mechanisms of viral infection and assembly.

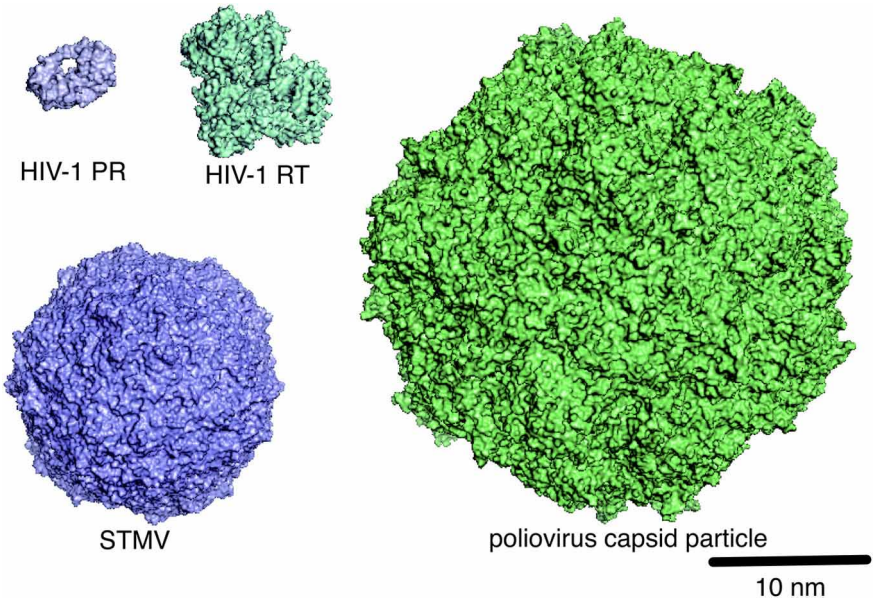


FIGURE 3 | Molecular scales of viral proteins and capsid particles. The structures of HIV-1 PR (PDB code: 1HHP), HIV-1 RT (PDB code: 1RTD), STMV (PDB code: 1A34), and poliovirus capsid particle (PDB code: 1HXS), which are deposited in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do) or the VIPERdb (http://viperdb.scripps.edu/), are shown in surface representation.

MD simulation is used to address this issue. Using a super computer, Freddolino et al. performed 50-nanosecond-timescale MD simulations of the symmetric structure of a complete satellite tobacco mosaic virus (STMV) particle containing about 1 million atoms (Arkhipov et al., 2006) (**Figure 3**). Thus, far, this is one of the largest systems among the MD simulations reported in all biological fields. Notably, the virion with viral RNA was stable during the simulations, whereas the one without the RNA was unstable, suggesting that viral RNA plays a key role in stabilizing the STMV virion (Arkhipov et al., 2006). The study is consistent with the experimental data (Day et al., 2001) and therefore provides a set of rationale conditions for performing the MD simulation of virion. Likewise, Larsson et al. reported about 1-microsecond-timescale MD simulations of the satellite tobacco necrosis virus (STNV) (Larsson et al., 2012). Their study reproduced the biochemical phenomenon of the STNV virion in solution (Unge et al., 1986), i.e., the swelling of capsid upon Ca^{2+} removal by EDTA treatment. These findings will provide a structural basis for identifying the key regulators of assembly and infections and for illustrating how they function mechanically. Although MD simulation of virions composed of very large numbers of atoms is still difficult in most cases, progress in the hardware and software for the simulation, together with the accumulation of biological and physicochemical information on virions, will help us to overcome these limitations in the MD simulation of virions.

PERSPECTIVE

Since the processing speed of computers is still doubling approximately every two years according to Moore's law, MD studies will be extended to simulations of larger and more complex system at longer timescales. This will then lead to a better understanding of the structures and dynamics of macromolecules involved in virus–host interactions.

COARSE-GRAINED (CG) MD SIMULATIONS

MD simulations of macromolecules consisting of large molecular systems, such as oligomeric proteins, macromolecular complexes, and membrane proteins in a lipid bilayer, and virions are desired to better understand viruses. However, such simulations require unrealistically long analytical times and high-performance computers at present, and thereby are still limited mostly to the small molecules (Henzler-Wildman and Kern, 2007; Dror et al., 2010). To cope with this issue and to improve the practicability of long timescale MD simulation, a "coarse-grained (CG) MD" simulation has been developed (Merchant and Madura, 2011; Takada, 2012). The CG-MD simulation employs "pseudo-atoms" that consist of several atoms in a group and calculates the movement of these "pseudo-atoms" rather than the movement of "individual atoms," thereby greatly reducing the calculation time (Merchant and Madura, 2011; Takada, 2012). CG-MD simulations have been used to study helicases of hepatitis C virus (HCV) and simian virus 40 and have successfully reproduced enzyme motions, such as "ratcheting inchworm translocation" and "spring-loaded DNA unwinding" (Flechsig and Mikhailov, 2010; Yoshimoto et al., 2010). Briefly, the ratcheting inchworm translocation is the unidirectional motion of the HCV NS3 helicase during translocation that occurs by the step size of one base per ATP hydrolysis cycle (Gu and Rice, 2010). Meanwhile, the spring-loaded DNA unwinding is the discrete steps of unwinding of DNA by the HCV NS3 helicase that occurs periodically via a burst of 3-bp unwinding during NS3 translocation consuming ATPs (Myong et al., 2007).

CG-MD has also been applied to the study of the structural characteristics and stabilities of the capsid particle and virion (**Figure 3**). Such studies have been used to investigate small plant viruses (∼28 nanometer in diameter), such as the three satellite plant viruses STMV, STNV, and the satellite panicum mosaic virus (SPMV), as well as the brome mosaic virus (BMV) (Arkhipov et al., 2006, 2009), and more complex capsids such as poliovirus (Arkhipov et al., 2006, 2009), asymmetric, conical-shaped HIV-1 capsid particles (Krishna et al., 2010), and the immature HIV-1 virion (Ayton and Voth, 2010). These studies have predicted various molecular interactions that can be tested experimentally. Thus, CG-MD may play a pivotal role in the MD study of micrometer-sized systems at millisecond timescale (Merchant and Madura, 2011; Takada, 2012) and therefore may uncover novel characteristics of the interactions in virus–host relationships.

INTRINSICALLY DISORDERED PROTEINS

Some eukaryotic proteins have no stable 3-D structure under physiological conditions (Dunker et al., 2002, 2008; Dyson and Wright, 2005). These proteins are referred to as intrinsically disordered, natively unfolded, or intrinsically unstructured proteins. They undergo structural transition from a disordered to an ordered state upon binding to target molecules such as proteins, DNA, and small molecules (Dunker et al., 2005; Sandhu and Dash, 2007). They are often related to the "hub proteins" that have many binding partners and control important biological processes (Iakoucheva et al., 2002; Haynes et al., 2006; Sandhu, 2009). Interestingly, viral proteins or portions of viral proteins are often intrinsically disordered. These include genome-linked protein VPg protein of plant viruses (Grzela et al., 2008; Rantalainen et al., 2008; Jiang and Laliberte, 2011; Rantalainen et al., 2011), HIV-1 Tat (Shojania and O'Neil, 2010), and Vif proteins (Reingewertz et al., 2010), and paramyxovirus nucleoproteins and phosphoproteins (Habchi and Longhi, 2012). It has been proposed that the disordered structure is beneficial for viruses to gain multiple functions in the viral life cycle with limited genome size (Rantalainen et al., 2011; Habchi and Longhi, 2012; Xue et al., 2012). Clarifying the folding landscape of viral proteins by standard MD and CG-MD simulations may help in understanding the structural principles by which viral proteins execute multiple functions in the viral life cycle.

ACKNOWLEDGMENTS

The MD studies of our laboratory described in this article are supported by grants from the Ministry of Health, Labor and Welfare for HIV/AIDS research and by a Grant-in-Aid for JSPS Fellows.

REFERENCES

Abbondanzieri, E. A., Bokinsky, G., Rausch, J. W., Zhang, J. X., Le Grice, S. F., and Zhuang, X. (2008). Dynamic binding orientations direct activity of HIV reverse transcriptase. *Nature* 453, 184–189.

Adcock, S. A., and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* 106, 1589–1615.

Amaro, R. E., Swift, R. V., Votapka, L., Li, W. W., Walker, R. C., and Bush, R. M. (2011). Mechanism of 150-cavity formation in influenza neuraminidase. *Nat. Commun.* 2, 388.

Arkhipov, A., Freddolino, P. L., and Schulten, K. (2006). Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure* 14, 1767–1777.

Arkhipov, A., Roos, W. H., Wuite, G. J., and Schulten, K. (2009). Elucidating the mechanism behind irreversible deformation of viral capsids. *Biophys. J.* 97, 2061–2069.

Astumian, R. D. (1997). Thermodynamics and kinetics of a brownian motor. *Science* 276, 917–922.

Autore, F., Bergeron, J. R., Malim, M. H., Fraternali, F., and Huthoff, H. (2010). Rationalisation of the differences between APOBEC3G structures from crystallography and NMR studies by molecular dynamics simulations. *PLoS ONE* 5:e11515. doi: 10.1371/journal.pone.0011515

Ayton, G. S., and Voth, G. A. (2010). Multiscale computer simulation of the immature HIV-1 virion. *Biophys. J.* 99, 2757–2765.

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. *Science* 294, 93–96.

Batista, P. R., Wilter, A., Durham, E. H., and Pascutti, P. G. (2006). Molecular dynamics simulations applied to the study of subtypes of HIV-1 protease common to Brazil, Africa, and Asia. *Cell Biochem. Biophys.* 44, 395–404.

Borhani, D. W., and Shaw, D. E. (2012). The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided. Mol. Des.* 26, 15–26.

Brooks, B. R., Brooks, C. L. 3rd., Mackerell, A. D. Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30, 1545–1614.

Cameron, C. E., Moustafa, I. M., and Arnold, J. J. (2009). Dynamics: the missing link between structure and function of the viral RNA-dependent RNA polymerase? *Curr. Opin. Struct. Biol.* 19, 768–774.

Carvalho, A. T., Fernandes, P. A., and Ramos, M. J. (2006). Insights on resistance to reverse transcriptase: the different patterns of interaction of the nucleoside reverse transcriptase inhibitors in the deoxyribonucleotide triphosphate binding site relative to the normal substrate. *J. Med. Chem.* 49, 7675–7682.

Carvalho, A. T., Fernandes, P. A., and Ramos, M. J. (2007). The excision mechanism in reverse transcriptase: pyrophosphate leaving and fingers opening are uncoupled events with the analogues AZT and d4T. *J. Phys. Chem. B* 111, 12032–12039.

Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M. Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The amber biomolecular simulation programs. *J. Comput. Chem.* 26, 1668–1688.

Chang, C. E., Shen, T., Trylska, J., Tozzini, V., and McCammon, J. A. (2006). Gated binding of ligands to HIV-1 protease: brownian dynamics simulations in a coarse-grained model. *Biophys. J.* 90, 3880–3885.

Chen, J., Zhang, S., Liu, X., and Zhang, Q. (2010). Insights into drug resistance of mutations D30N and I50V to HIV-1 protease inhibitor TMC-114, free energy calculation and molecular dynamic simulation. *J. Mol. Model.* 16, 459–468.

Cortez, K. J., and Maldarelli, F. (2011). Clinical management of HIV drug resistance. *Viruses* 3, 347–378.

Da, L. T., and Wu, Y. D. (2011). Theoretical studies on the interactions and interferences of HIV-1 glycoprotein gp120 and its coreceptor CCR5. *J. Chem. Inf. Model.* 51, 359–369.

Datta, S. A., Temeselew, L. G., Crist, R. M., Soheilian, F., Kamata, A., Mirro, J., Harvin, D., Nagashima, K., Cachau, R. E., and Rein, A. (2011). On the role of the SP1 domain in HIV-1 particle assembly: a molecular switch? *J. Virol.* 85, 4111–4121.

Day, J., Kuznetsov, Y. G., Larson, S. B., Greenwood, A., and McPherson, A. (2001). Biophysical studies on the RNA cores of satellite tobacco mosaic virus. *Biophys. J.* 80, 2364–2371.

Deng, N. J., Zheng, W., Gallicchio, E., and Levy, R. M. (2011). Insights into the dynamics of HIV-1 protease: a kinetic network

model constructed from atomistic simulations. *J. Am. Chem. Soc.* 133, 9387–9394.

Dirauf, P., Meiselbach, H., and Sticht, H. (2010). Effects of the V82A and I54V mutations on the dynamics and ligand binding properties of HIV-1 protease. *J. Mol. Model.* 16, 1577–1583.

Dodson, G. G., Lane, D. P., and Verma, C. S. (2008). Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep.* 9, 144–150.

Douglas, J. L., Gustin, J. K., Viswanathan, K., Mansouri, M., Moses, A. V., and Fruh, K. (2010). The great escape: viral strategies to counter BST-2/tetherin. *PLoS Pathog.* 6:e1000913. doi: 10.1371/journal.ppat.1000913

Dror, R. O., Jensen, M. O., Borhani, D. W., and Shaw, D. E. (2010). Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J. Gen. Physiol.* 135, 555–562.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry* 41, 6573–6582.

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 272, 5129–5148.

Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 18, 756–764.

Dunn, B. M., Goodenow, M. M., Gustchina, A., and Wlodawer, A. (2002). Retroviral proteases. *Genome Biol.* 3, Reviews3006.

Durrant, J. D., and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biol.* 9, 71.

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.* 6, 197–208.

Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438, 117–121.

Flechsig, H., and Mikhailov, A. S. (2010). Tracing entire operation cycles of molecular motor hepatitis C virus helicase in structurally resolved dynamical simulations. *Proc. Natl. Acad. Sci. U.S.A.* 107, 20875–20880.

Foulkes-Murzycki, J. E., Scott, W. R., and Schiffer, C. A. (2007). Hydrophobic sliding: a possible mechanism for drug resistance in human immunodeficiency virus type 1 protease. *Structure* 15, 225–233.

Grzela, R., Szolajska, E., Ebel, C., Madern, D., Favier, A., Wojtal, I., Zagorski, W., and Chroboczek, J. (2008). Virulence factor of potato virus Y, genome-attached terminal protein VPg, is a highly disordered protein. *J. Biol. Chem.* 283, 213–221.

Gu, M., and Rice, C. M. (2010). Three conformational snapshots of the hepatitis C virus NS3 helicase reveal a ratchet translocation mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 107, 521–528.

Gu, R. X., Liu, L. A., Wei, D. Q., Du, J. G., Liu, L., and Liu, H. (2011). Free energy calculations on the two drug binding sites in the M2 proton channel. *J. Am. Chem. Soc.* 133, 10817–10825.

Habchi, J., and Longhi, S. (2012). Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. *Mol. Biosyst.* 8, 69–81.

Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., Uversky, V. N., Vidal, M., and Iakoucheva, L. M. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* 2:e100. doi: 10.1371/journal.pcbi.0020100

Henzler-Wildman, K., and Kern, D. (2007). Dynamic personalities of proteins. *Nature* 450, 964–972.

Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M., and Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450, 913–916.

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.

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 323, 573–584.

Ishima, R., and Torchia, D. A. (2000). Protein dynamics from NMR. *Nat. Struct. Biol.* 7, 740–743.

Jiang, J., and Laliberte, J. F. (2011). The genome-linked protein VPg of plant viruses-a protein with many partners. *Curr. Opin. Virol.* 1, 347–354.

Joyce, C. M., and Steitz, T. A. (1994). Function and structure relationships in DNA polymerases. *Annu. Rev. Biochem.* 63, 777–822.

Kagan, R. M., Shenderovich, M. D., Heseltine, P. N., and Ramnarayan, K. (2005). Structural analysis of an HIV-1 protease I47A mutant resistant to the protease inhibitor lopinavir. *Protein Sci.* 14, 1870–1878.

Kar, P., and Knecht, V. (2012). Origin of decrease in potency of darunavir and two related antiviral inhibitors against HIV-2 compared to HIV-1 protease. *J. Phys. Chem. B* 116, 2605–2614.

Karplus, M., and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6679–6685.

Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9, 646–652.

Karplus, M., and Petsko, G. A. (1990). Molecular dynamics simulations in biology. *Nature* 347, 631–639.

Karthik, S., and Senapati, S. (2011). Dynamic flaps in HIV-1 protease adopt unique ordering at different stages in the catalytic cycle. *Proteins* 79, 1830–1840.

Kasson, P. M. (2012). Receptor binding by influenza virus: using computational techniques to extend structural data. *Biochemistry* 51, 2359–2365.

Kirmizialtin, S., Nguyen, V., Johnson, K. A., and Elber, R. (2012). How conformational dynamics of DNA polymerase select correct substrates: experiments and simulations. *Structure* 20, 618–627.

Kobayashi, T., Ode, H., Yoshida, T., Sato, K., Gee, P., Yamamoto, S. P., Ebina, H., Strebel, K., Sato, H., and Koyanagi, Y. (2011). Identification of amino acids in the human tetherin transmembrane domain responsible for HIV-1 Vpu interaction and susceptibility. *J. Virol.* 85, 932–945.

Krishna, V., Ayton, G. S., and Voth, G. A. (2010). Role of protein interactions in defining HIV-1 viral capsid shape and stability: a coarse-grained analysis. *Biophys. J.* 98, 18–26.

Lamers, M. H., Georgescu, R. E., Lee, S. G., O'Donnell, M., and Kuriyan, J. (2006). Crystal structure of the catalytic alpha subunit of E. coli replicative DNA polymerase III. *Cell* 126, 881–892.

Larsson, D. S., Liljas, L., and van der Spoel, D. (2012). Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLoS Comput. Biol.* 8:e1002502. doi: 10.1371/journal.pcbi.1002502

Le, L., Lee, E. H., Hardy, D. J., Truong, T. N., and Schulten, K. (2010). Molecular dynamics simulations suggest that electrostatic funnel directs binding of Tamiflu to influenza N1 neuraminidases. *PLoS Comput. Biol.* 6:e1000939. doi: 10.1371/journal.pcbi.1000939

Lee, D. J., and Robinson, W. E. Jr. (2006). Preliminary mapping of a putative inhibitor-binding pocket for human immunodeficiency virus type 1 integrase inhibitors. *Antimicrob. Agents Chemother.* 50, 134–142.

Leonov, H., Astrahan, P., Krugliak, M., and Arkin, I. T. (2011). How do aminoadamantanes block the influenza M2 channel, and how does resistance develop? *J. Am. Chem. Soc.* 133, 9903–9911.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Systematic validation of protein force fields against experimental data. *PLoS ONE* 7:e32131. doi: 10.1371/journal.pone.0032131

Lu, H. P., Xun, L., and Xie, X. S. (1998). Single-molecule enzymatic dynamics. *Science* 282, 1877–1882.

Maeda, K., Das, D., Yin, P. D., Tsuchiya, K., Ogata-Aoki, H., Nakata, H., Norman, R. B., Hackney, L. A., Takaoka, Y., and Mitsuya, H. (2008). Involvement of the second extracellular loop and transmembrane residues of CCR5 in inhibitor binding and HIV-1 fusion: insights into the mechanism of allosteric inhibition. *J. Mol. Biol.* 381, 956–974.

Malim, M. H., and Bieniasz, P. D. (2012). HIV restriction factors and mechanisms of evasion. *Cold Spring Harb. Perspect. Med.* 2, a006940.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325.

Matsuyama, S., Aydan, A., Ode, H., Hata, M., Sugiura, W., and Hoshino, T. (2010). Structural and energetic analysis on the complexes of clinically isolated subtype C HIV-1 proteases and approved inhibitors by molecular dynamics simulation. *J. Phys. Chem. B* 114, 521–530.

McCammon, J. A., Gelin, B. R., and Karplus, M. (1977). Dynamics of folded proteins. *Nature* 267, 585–590.

Merchant, B. A., and Madura, J. D. (2011). A review of coarse-grained molecular dynamics techniques to access extended spatial and temporal scales in biomolecular simulations. *Annu. Rep. Comput. Chem.* 7, 67–87.

Miyamoto, T., Yokoyama, M., Kono, K., Shioda, T., Sato, H., and Nakayama, E. E. (2011). A single amino acid of human immunodeficiency virus type 2 capsid protein affects conformation of two external loops and viral sensitivity to TRIM5alpha. *PLoS ONE* 6:e22779. doi: 10.1371/journal.pone.0022779

Moustafa, I. M., Shen, H., Morton, B., Colina, C. M., and Cameron, C. E. (2011). Molecular dynamics simulations of viral RNA polymerases link conserved and correlated motions of functional elements to fidelity. *J. Mol. Biol.* 410, 159–181.

Myong, S., Bruno, M. M., Pyle, A. M., and Ha, T. (2007). Spring-loaded mechanism of DNA unwinding by hepatitis C virus NS3 helicase. *Science* 317, 513–516.

Naganawa, S., Yokoyama, M., Shiino, T., Suzuki, T., Ishigatsubo, Y., Ueda, A., Shirai, A., Takeno, M., Hayakawa, S., Sato, S., Tochikubo, O., Kiyoura, S., Sawada, K., Ikegami, T., Kanda, T., Kitamura, K., and Sato, H. (2008). Net positive charge of HIV-1 CRF01_AE V3 sequence regulates viral sensitivity to humoral immunity. *PLoS ONE* 3:e3206. doi: 10.1371/journal.pone.0003206

Neil, S., and Bieniasz, P. (2009). Human immunodeficiency virus, restriction factors, and interferon. *J. Interferon Cytokine Res.* 29, 569–580.

Nicholson, L. K., Yamazaki, T., Torchia, D. A., Grzesiek, S., Bax, A., Stahl, S. J., Kaufman, J. D., Wingfield, P. T., Lam, P. Y. S., Jadhav, P. K., Hodge, C. N., Domaille, P. J., and Chang, C.-H. (1995). Flexibility and function in HIV-1 protease. *Nat. Struct. Biol.* 2, 274–280.

Ode, H., Matsuyama, S., Hata, M., Hoshino, T., Kakizawa, J., and Sugiura, W. (2007a). Mechanism of drug resistance due to N88S in CRF01_AE HIV-1 protease, analyzed by molecular dynamics simulations. *J. Med. Chem.* 50, 1768–1777.

Ode, H., Matsuyama, S., Hata, M., Neya, S., Kakizawa, J., Sugiura, W., and Hoshino, T. (2007b). Computational characterization of structural role of the non-active site mutation M36I of human immunodeficiency virus type 1 protease. *J. Mol. Biol.* 370, 598–607.

Ode, H., Neya, S., Hata, M., Sugiura, W., and Hoshino, T. (2006). Computational simulations of HIV-1 proteases - multi-drug resistance due to nonactive site mutation L90M. *J. Am. Chem. Soc.* 128, 7887–7895.

Ode, H., Ota, M., Neya, S., Hata, M., Sugiura, W., and Hoshino, T. (2005). Resistant mechanism against nelfinavir of human immunodeficiency virus type 1 proteases. *J. Phys. Chem. B* 109, 565–574.

Ode, H., Yokoyama, M., Kanda, T., and Sato, H. (2011). Identification of folding preferences of cleavage junctions of HIV-1 precursor proteins for regulation of cleavability. *J. Mol. Model.* 17, 391–399.

Okimoto, N., Futatsugi, N., Fuji, H., Suenaga, A., Morimoto, G., Yanai, R., Ohno, Y., Narumi, T., and Taiji, M. (2009). High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS Comput. Biol.* 5:e1000528. doi: 10.1371/journal.pcbi.1000528

Onyango, C. O., Leligdowicz, A., Yokoyama, M., Sato, H., Song, H., Nakayama, E. E., Shioda, T., de Silva, T., Townend, J., Jaye, A., Whittle, H., Rowland-Jones, S., and Cotten, M. (2010). HIV-2 capsids distinguish high and low virus load patients in a West African community cohort. *Vaccine* 28(Suppl. 2), B60–B67.

Ozen, A., Haliloglu, T., and Schiffer, C. A. (2011). Dynamics of preferential substrate recognition in HIV-1 protease: redefining the substrate envelope. *J. Mol. Biol.* 410, 726–744.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E. I., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.* 91, 1–41.

Perryman, A. L., Lin, J. H., and McCammon, J. A. (2004). HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci.* 13, 1108–1123.

Pettit, S. C., Henderson, G. J., Schiffer, C. A., and Swanstrom, R. (2002). Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. *J. Virol.* 76, 10226–10233.

Pettit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. V., Klein, C. A., and Swanstrom, R. (1994). The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J. Virol.* 68, 8017–8027.

Piana, S., Carloni, P., and Rothlisberger, U. (2002). Drug resistance in HIV-1 protease: flexibility-assisted mechanism of compensatory mutations. *Protein Sci.* 11, 2393–2402.

Rantalainen, K. I., Eskelin, K., Tompa, P., and Makinen, K. (2011). Structural flexibility allows the functional diversity of potyvirus genome-linked protein VPg. *J. Virol.* 85, 2449–2457.

Rantalainen, K. I., Uversky, V. N., Permi, P., Kalkkinen, N., Dunker, A. K., and Makinen, K. (2008). Potato virus A genome-linked protein VPg is an intrinsically disordered molten globule-like protein with a hydrophobic core. *Virology* 377, 280–288.

Reboul, C. F., Meyer, G. R., Porebski, B. T., Borg, N. A., and Buckle, A. M. (2012). Epitope flexibility and dynamic footprint revealed by molecular dynamics of a pMHC-TCR complex. *PLoS Comput. Biol.* 8:e1002404. doi: 10.1371/journal.pcbi.1002404

Reingewertz, T. H., Shalev, D. E., and Friedler, A. (2010). Structural disorder in the HIV-1 Vif protein and interaction-dependent gain of structure. *Protein Pept. Lett.* 17, 988–998.

Rodriguez-Barrios, F., Balzarini, J., and Gago, F. (2005). The molecular basis of resilience to the effect of the Lys103Asn mutation in non-nucleoside HIV-1 reverse transcriptase inhibitors studied by targeted molecular dynamics simulations. *J. Am. Chem. Soc.* 127, 7570–7578.

Rodriguez-Barrios, F., and Gago, F. (2004). Understanding the basis of resistance in the irksome Lys103Asn HIV-1 reverse transcriptase mutant through targeted molecular dynamics simulations. *J. Am. Chem. Soc.* 126, 15386–15387.

Sadiq, S. K., Wan, S., and Coveney, P. V. (2007). Insights into a mutation-assisted lateral drug escape mechanism from the HIV-1 protease active site. *Biochemistry* 46, 14865–14877.

Sadiq, S. K., Wright, D. W., Kenway, O. A., and Coveney, P. V. (2010). Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model.* 50, 890–905.

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000). Protein structure modeling for structural genomics. *Nat. Struct. Biol.* 7(Suppl.), 986–990.

Sandhu, K. S. (2009). Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. *J. Mol. Recognit.* 22, 1–8.

Sandhu, K. S., and Dash, D. (2007). Dynamic alpha-helices: conformations that do not conform. *Proteins* 68, 109–122.

Schames, J. R., Henchman, R. H., Siegel, J. S., Sotriffer, C. A., Ni, H., and McCammon, J. A. (2004). Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* 47, 1879–1881.

Shenderovich, M. D., Kagan, R. M., Heseltine, P. N., and Ramnarayan, K. (2003). Structure-based phenotyping predicts HIV-1 protease inhibitor resistance. *Protein Sci.* 12, 1706–1718.

Shojania, S., and O'Neil, J. D. (2010). Intrinsic disorder and function of the HIV-1 Tat protein. *Protein Pept. Lett.* 17, 999–1011.

Skasko, M., Wang, Y., Tian, Y., Tokarev, A., Munguia, J., Ruiz, A., Stephens, E. B., Opella, S. J., and Guatelli, J. (2012). HIV-1 Vpu antagonizes the innate restriction factor BST-2 via lipid-embedded helix-helix interactions. *J. Biol. Chem.* 287, 58–67.

Soares, R. O., Batista, P. R., Costa, M. G., Dardenne, L. E., Pascutti, P. G., and Soares, M. A. (2010). Understanding the HIV-1 protease nelfinavir resistance mutation D30N in subtypes B and C through molecular dynamics simulations. *J. Mol. Graph. Model.* 29, 137–147.

Song, H., Nakayama, E. E., Yokoyama, M., Sato, H., Levy, J. A., and Shioda, T. (2007). A single amino acid of the human immunodeficiency virus type 2 capsid affects its replication in the presence of cynomolgus monkey and human TRIM5alphas. *J. Virol.* 81, 7280–7285.

Steven, A. C., Heymann, J. B., Cheng, N., Trus, B. L., and Conway, J. F. (2005). Virus maturation: dynamics and mechanism of a stabilizing structural transition that leads to infectivity. *Curr. Opin. Struct. Biol.* 15, 227–236.

Stoica, I., Sadiq, S. K., and Coveney, P. V. (2008). Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *J. Am. Chem. Soc.* 130, 2639–2648.

Takada, S. (2012). Coarse-grained molecular simulations of large

biomolecules. *Curr. Opin. Struct. Biol.* 22, 130–137.

Thorpe, I. F., and Brooks, C. L. 3rd. (2007). Molecular evolution of affinity and flexibility in the immune system. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8821–8826.

Tobi, D., and Bahar, I. (2005). Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U.S.A.* 102, 18908–18913.

Unge, T., Montelius, I., Liljas, L., and Ofverstedt, L. G. (1986). The EDTA-treated expanded satellite tobacco necrosis virus: biochemical properties and crystallization. *Virology* 152, 207–218.

van der Vries, E., Schutten, M., and Boucher, C. A. (2011). The potential for multidrug-resistant influenza. *Curr. Opin. Infect. Dis.* 24, 599–604.

Wang, J., Ma, C., Fiorin, G., Carnevale, V., Wang, T., Hu, F., Lamb, R. A., Pinto, L. H., Hong, M., Klein, M. L., and DeGrado, W. F. (2011). Molecular dynamics simulation directed rational design of inhibitors targeting drug-resistant mutants of influenza A virus M2. *J. Am. Chem. Soc.* 133, 12834–12841.

White, K. L., Chen, J. M., Margot, N. A., Wrin, T., Petropoulos, C. J., Naeger, L. K., Swaminathan, S., and Miller, M. D. (2004). Molecular mechanisms of tenofovir resistance conferred by human immunodeficiency virus type 1 reverse transcriptase containing a diserine insertion after residue 69 and multiple thymidine analog-associated mutations. *Antimicrob. Agents Chemother.* 48, 992–1003.

Wittayanarakul, K., Aruksakunwong, O., Saen-oon, S., Chantratita, W., Parasuk, V., Sompornpisut, P., and Hannongbua, S. (2005). Insights into saquinavir resistance in the G48V HIV-1 protease: quantum calculations and molecular dynamic simulations. *Biophys. J.* 88, 867–879.

Wright, D. W., and Coveney, P. V. (2011). Resolution of discordant HIV-1 protease resistance rankings using molecular dynamics simulations. *J. Chem. Inf. Model.* 51, 2636–2649.

Xu, Y., Colletier, J. P., Jiang, H., Silman, I., Sussman, J. L., and Weik, M. (2008). Induced-fit or preexisting equilibrium dynamics? lessons from protein crystallography and MD simulations on acetylcholinesterase and implications for structure-based drug design. *Protein Sci.* 17, 601–605.

Xue, B., Mizianty, M. J., Kurgan, L., and Uversky, V. N. (2012). Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell. Mol. Life Sci.* 69, 1211–1259.

Yokoyama, M., Naganawa, S., Yoshimura, K., Matsushita, S., and Sato, H. (2012). Structural dynamics of HIV-1 envelope gp120 outer domain with V3 loop. *PLoS ONE* 7:e37530. doi: 10.1371/journal.pone.0037530

Yoshimoto, K., Arora, K., and Brooks, C. L. 3rd. (2010). Hexameric helicase deconstructed: interplay of conformational changes and substrate coupling. *Biophys. J.* 98, 1449–1457.

Zhou, J., Zhang, Z., Mi, Z., Wang, X., Zhang, Q., Li, X., Liang, C., and Cen, S. (2012). Characterization of the interface of the bone marrow stromal cell antigen 2-Vpu protein complex via computational chemistry. *Biochemistry* 51, 1288–1296.

Zhou, Z., Madrid, M., Evanseck, J. D., and Madura, J. D. (2005). Effect of a bound non-nucleoside RT inhibitor on the dynamics of wild-type and mutant HIV-1 reverse transcriptase. *J. Am. Chem. Soc.* 127, 17253–17260.

Conflict of Interest Statement: 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.

Received: 08 June 2012; paper pending published: 20 June 2012; accepted: 02 July 2012; published online: 19 July 2012.
*Citation: Ode H, Nakashima M, Kitamura S, Sugiura W and Sato H (2012) Molecular dynamics simulation in virus research. Front. Microbio. **3**:258. doi: 10.3389/fmicb.2012.00258*
This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

MINI REVIEW ARTICLE
published: 21 December 2012
doi: 10.3389/fmicb.2012.00428

Toward a three-dimensional view of protein networks between species

Eric A. Franzosa[1], Sara Garamszegi[1] and Yu Xia[1,2,3,4]*

[1] *Bioinformatics Program, Boston University, Boston, MA, USA*
[2] *Department of Chemistry, Boston University, Boston, MA, USA*
[3] *Department of Biomedical Engineering, Boston University, Boston, MA, USA*
[4] *Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan
Jens Von Einem, Institute of Virology, Ulm University Hospital, Germany
Pascal Braun, Technical University of Munich, Germany

***Correspondence:**
Yu Xia, Bioinformatics Program, Boston University, Boston, MA 02215, USA.
e-mail: yuxia@bu.edu

General principles governing biomolecular interactions between species are expected to differ significantly from known principles governing the interactions within species, yet these principles remain poorly understood at the systems level. A key reason for this knowledge gap is the lack of a detailed three-dimensional (3D), atomistic view of biomolecular interaction networks between species. Recent progress in structural biology, systems biology, and computational biology has enabled accurate and large-scale construction of 3D structural models of nodes and edges for protein–protein interaction networks within and between species. The resulting within- and between-species structural interaction networks have provided new biophysical, functional, and evolutionary insights into species interactions and infectious disease. Here, we review the nascent field of between-species structural systems biology, focusing on interactions between host and pathogens such as viruses.

Keywords: structural systems biology, protein–protein interaction, host–pathogen interaction, bioinformatics and computational biology, network biology

INTRODUCTION

Protein–protein interactions (PPIs) can be divided into two fundamentally different classes. The first class of PPIs involves interactions between two proteins encoded within the genome of a single species, where the two proteins cooperate with each other to achieve cellular function in a coordinated fashion. The second class of PPIs involves interactions between two proteins from different species, for example between host proteins and microbial proteins, or between proteins from two different microbial species. These between-species PPIs play key roles in host–microbe and microbe–microbe interactions. Unlike the cooperative PPIs within the host, the interactions between host and microbes are driven by a wide spectrum of co-evolutionary mechanisms, ranging from parasitic to mutualistic (Dethlefsen et al., 2007). General principles of the PPI networks between microbes and their host may differ significantly from known principles governing the cooperative PPI network encoded within the host, yet these principles are not well understood. Here, we review recent progress toward constructing a high-resolution, three-dimensional (3D) structural view of host–pathogen and within-host PPI networks. The resulting host–pathogen and within-host structural interaction networks enable the discovery of new principles of host–pathogen interactions that are otherwise hidden in the binary PPI network. This review focuses on high-throughput mapping and large-scale analysis of host–pathogen PPI networks, which reveal global trends and patterns in host–pathogen interactions that are minimally confounded by investigator biases.

HOST–PATHOGEN PROTEIN–PROTEIN INTERACTION NETWORKS

The first step toward building host–pathogen structural interaction networks is to map the networks of physical interactions between host proteins and pathogen proteins. Host–pathogen PPIs have traditionally been studied one at a time. Recently, systems biology approaches have been applied to host–pathogen interaction research. Significant progress has been made in genome-wide mapping of host–pathogen PPI networks ("interactomes") for many pathogens, especially viruses. Using high-throughput methods such as the yeast two-hybrid system (Fields and Song, 1989) and affinity purification followed by mass spectrometry identification (Rigaut et al., 1999), experimental host–pathogen interactome maps now exist for many viruses (von Schwedler et al., 2003; Uetz et al., 2006; Calderwood et al., 2007; de Chassey et al., 2008; Shapira et al., 2009; Zhang et al., 2009; Khadka et al., 2011; Jager et al., 2012; Pichlmair et al., 2012; Rozenblatt-Rosen et al., 2012). Since viruses are obligate intracellular parasites with small genomes, many, but not all, physical interactions between viral proteins and host proteins have functional importance. Thus, it is essential to complement physical interactome mapping with functional assays that identify host proteins whose perturbation significantly affects viral infection and replication (Brass et al., 2008; Konig et al., 2008, 2010; Krishnan et al., 2008; Karlas et al., 2010). In addition to host–virus interactome maps, limited host–pathogen interactome data exist for bacterial and eukaryotic pathogens (Dyer et al., 2008, 2010; Mukhtar et al., 2011; Schleker et al., 2012). Since most proteins in bacterial and eukaryotic pathogens do not directly interact with host proteins, a key challenge is to identify pathogen effector proteins that act directly on the host cell to enable infection (Tobe et al., 2006).

Experimental host–pathogen interactome datasets are expected to continue to expand in the near future. The many thousands of experimentally detected host–pathogen PPIs are collected in databases such as VirusMINT (Chatr-aryamontri et al., 2009), VirHostNet (Navratil et al., 2009), IntAct (Aranda et al., 2010),

PIG (Driscoll et al., 2009), and NCBI HIV-1 protein interaction database (Fu et al., 2009). These databases typically rely heavily on manual curation to maintain standards of quality, and there is a great need to complement manual curation with automated literature mining of host–pathogen PPIs (Thieu et al., 2012).

Because of the challenges associated with experimental determination of host–pathogen PPIs, it is desirable to develop computational methods to predict host–pathogen PPIs. Prediction of host–pathogen PPIs is usually based on sequence homology with known PPIs (Uetz et al., 2006; Davis et al., 2007; Doolittle and Gomez, 2011; Wuchty, 2011), the presence of known or predicted interacting domain pairs (Dyer et al., 2007), as well as the presence of other predictive sequence and functional features (Tastan et al., 2009; Qi et al., 2010; Dyer et al., 2011). Computational predictions of host–pathogens PPIs are most effective as a means to prioritize subsequent experimental validations, which are often time-consuming (Uetz et al., 2006). Other areas where computational methods play an increasingly important role include genomic data integration of diverse host–pathogen physical, genetic, and functional interactions (Shapira et al., 2009; Konig et al., 2010; Rozenblatt-Rosen et al., 2012), and network-based prediction of host proteins important in host–pathogen interaction (Navratil et al., 2010; Murali et al., 2011).

Experimental host–pathogen PPI networks are useful in many ways. They not only help generate hypotheses regarding the function of specific pathogen proteins and the biology of specific pathogens, but also provide insights into principles governing host–pathogen interactions at the systems level. Global analyses of host–pathogen PPI networks have revealed that viruses and other microbial pathogens tend to interact with host proteins that are hubs (i.e., proteins with many interaction partners in the host network) and bottlenecks (i.e., proteins whose removal would disrupt many shortest paths in the host network; Calderwood et al., 2007; de Chassey et al., 2008; Dyer et al., 2008; Wuchty et al., 2010; Pichlmair et al., 2012). Host proteins that interact with pathogens tend to be conserved among closely related species (Jager et al., 2012; Pichlmair et al., 2012), although many of them are also under positive selection (Bozek and Lengauer, 2010). Host proteins that interact with pathogens tend to form densely connected network modules by clustering into biological pathways and physical complexes (Dyer et al., 2008; Bushman et al., 2009; MacPherson et al., 2010). In addition, host–pathogen PPI networks are enriched for certain network motifs (e.g., mutual inhibition; van Dijk et al., 2010). Furthermore, pathogens tend to target host proteins involved in common biological processes essential to pathogen infection and replication in general, such as host defense and immune response (Dyer et al., 2008; Pichlmair et al., 2012), often through convergent evolution (Mukhtar et al., 2011). At the same time, different classes of pathogens (e.g., DNA viruses versus RNA viruses, or viruses versus bacteria) also target distinct host pathways due to class-specific differences in infection and replication mechanisms (Durmus Tekir et al., 2012; Pichlmair et al., 2012). Finally, host proteins targeted by pathogens tend to be in network proximity to other proteins implicated in diseases associated with pathogen infections (Navratil et al., 2011; Gulbahce et al., 2012). It is clear that much can be learned by taking a global and network perspective on host–pathogen interactions.

HOST–PATHOGEN STRUCTURAL INTERACTION NETWORKS

The mapping of host–pathogen PPI networks lays the foundation for and constitutes the first step toward constructing host–pathogen structural interaction networks. Despite experimental and computational advances in the global analysis of host–pathogen PPI networks, the utility of PPI networks is ultimately limited by their low-resolution nature (i.e., proteins represented as nodes and PPIs represented as edges). A high-resolution view of the host–pathogen PPI network can be achieved by building accurate 3D structural models for nodes and edges in the network (**Figure 1A**). Is it feasible to construct such a host–pathogen structural interaction network in a global and accurate way? And if so, does this 3D structural view provide new insights into host–pathogen interactions that are not apparent in the binary PPI network?

Although the 3D structure of proteins and PPIs can in principle be predicted from sequence without resorting to homology [using template-free structure prediction (Moult, 2005) and macromolecular docking (Gray, 2006)], in practice homology modeling remains the most successful and reliable 3D structure prediction method on a genomic scale for both proteins and PPIs (Marti-Renom et al., 2000; Russell et al., 2004). To build a homology model for a query protein or a query pair of interacting proteins, the query protein or protein pair is searched against a template library consisting of proteins or PPIs of known 3D structure deposited in the Protein Data Bank (PDB; Berman et al., 2000). The most significantly matched 3D template is then used to construct a homology model for the query protein or PPI. Despite the obvious limitations that good homology models cannot be built for proteins with entirely new folds or PPIs with entirely new modes of interaction, and that the conformation of proteins and PPIs is not always conserved during evolution, homology modeling has been highly successful in practice, thanks to major advances in structural biology and computational biology. Proteins and PPIs are composed of a limited number of domains and domain–domain interactions (Chothia, 1992; Aloy and Russell, 2004), and certain domains and domain–domain interactions are significantly overrepresented in proteomes and interactomes (Qian et al., 2001). Thus, homology models for many proteins and PPIs can be built based on a relatively small number of representative domains and domain–domain interactions of known 3D structure, stored in databases such as SUPERFAMILY (Madera et al., 2004), iPfam (Finn et al., 2005), and 3did (Stein et al., 2005). Indeed, it is estimated that ∼60% of all query proteins share significant sequence similarity with at least one template protein of known 3D structure (Madera et al., 2004). For the vast majority of these cases, the query protein shares significant structural similarity with the template protein, an accurate sequence alignment can be constructed, and an accurate homology model (∼3 Å RMSD) can be built for at least a part of the query protein (typically a domain; Marti-Renom et al., 2000; Dalton and Jackson, 2007). Compared to homology modeling of single proteins, the coverage of accurate homology models for within-species PPIs is smaller but still considerable (∼20%; Kim et al., 2006). Indeed, it was recently argued that 3D templates exist for most known within-species PPIs, provided that good homology models can be built for the protein components (Kundrotas et al., 2012). The coverage of accurate template-based

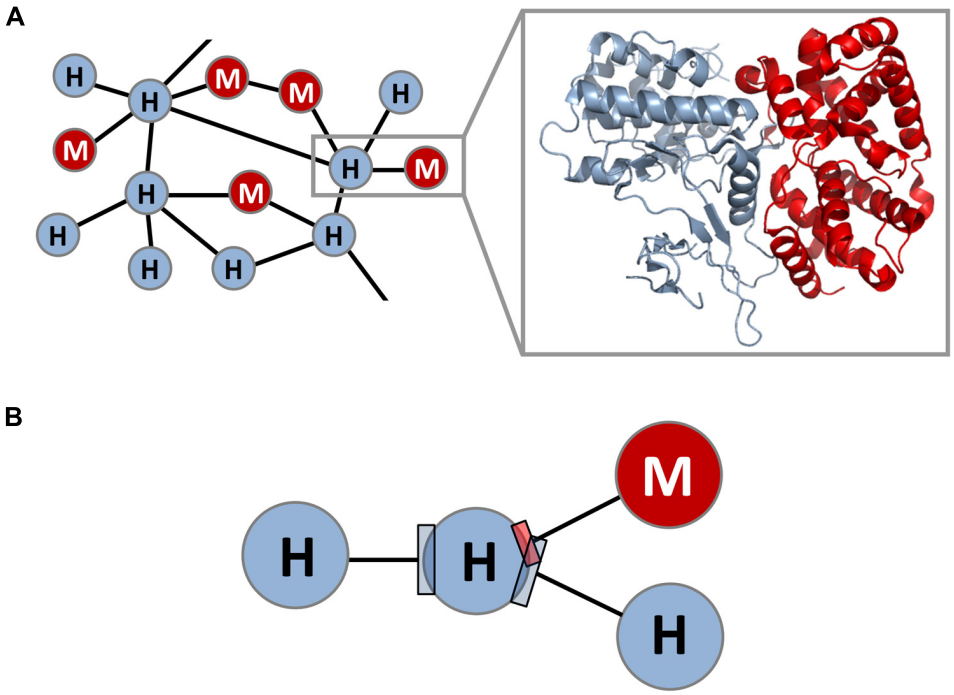


FIGURE 1 | Structural interaction network between species. (A) Shown is a high-resolution, 3D structural view of the PPI network between host and microbial pathogens, where each within-host and host–microbe protein–protein interaction (PPI) edge is associated with an accurate 3D structural model; one such interaction (gray box) and its structural model are highlighted. Interactions can be within human or within microbe (within-species interactions), or between human and microbe (between-species interactions). **(B)** The resulting host–microbe structural interaction network reveals high-resolution geometrical relationships between exogenous interfaces (between-species interfaces) and endogenous interfaces (within-species interfaces) that are otherwise hidden in the binary PPI network. In this example, a microbial protein is seen to bind to a target protein in the host at the same site as another host protein, albeit using a smaller interface.

models for PPIs can be further improved by identifying additional 3D templates that are structurally similar to the query proteins in the absence of sequence similarity (Zhang et al., 2012).

Homology modeling has been successfully used to construct within-species structural interaction networks, where 3D structural models are built for known within-species PPIs (Aloy et al., 2004; Kim et al., 2006). Despite the caveat that 3D homology models are biased toward soluble, stable, and structurally well-ordered proteins and PPIs, structural interaction networks can be viewed as high-quality subsets of binary PPI networks with much higher spatial resolution. Computational analyses of the within-species structural interaction networks have provided significant insights into a wide range of topics including biophysics, evolution, disease biology, and drug design (Kim et al., 2006, 2008; Franzosa and Xia, 2008, 2009; Kar et al., 2009; Xie et al., 2011; Wang et al., 2012). Such structural systems biology approaches are highly valuable as a unifying framework that integrates molecular biophysics with cell systems biology.

Most recently, structural systems biology was applied to between-species interactions, and an integrated map of human–virus and within-human structural interaction networks was constructed (Franzosa and Xia, 2011). The structural interaction networks consist of 53 human-virus PPIs and >3,000 human–human PPIs in the form of either experimental 3D structures or homology models. Here, instead of predicting new host–pathogen PPIs (Davis et al., 2007), homology modeling is used to annotate known host–pathogen PPIs with 3D structural information, thus providing a structural map of the binary PPI network in much higher spatial resolution. For example, the binary PPI network indicates that the human CDK6 protein interacts with both human proteins and the cyclin D homolog protein from herpesvirus. The structural interaction network further reveals that these interactions largely occur at two distinct, non-overlapping interfaces on the human CDK6 protein: one interface mediating the interactions with the viral protein as well as the human cyclin D protein, and a second interface mediating the interactions with various human CDK inhibitor proteins (Russo et al., 1998; Pratt et al., 2006). Such a high-resolution map enables the detailed analysis of the geometrical properties and relationships of human–virus PPI interfaces (exogenous interfaces) and human–human PPI interfaces (endogenous interfaces) that is otherwise inaccessible in the binary PPI network (**Figure 1B**; Franzosa and Xia, 2011). For example, although binary PPI network analysis revealed that viral proteins tend to interact with host protein hubs participating in many endogenous interactions, the precise spatial relationships among these exogenous and endogenous interactions are not known. On the other hand, structural interaction analysis further revealed that exogenous interfaces, although smaller in size, tend to overlap significantly with and mimic endogenous interfaces, often in the absence of sequence or structural

similarity. In addition, the endogenous interfaces that are mimicked by viral proteins tend to participate in multiple endogenous interactions which are transient and regulatory in nature. A case in point is the interaction between the UL36 protein from the HSV-1 virus and the human ubiquitin protein, an important regulator of protein function and cell behavior (Schlieker et al., 2007). The endogenous interface of the human ubiquitin protein mimicked by the virus mediates as many as 30 interactions with other human proteins. On average an endogenous interface mimicked by virus mediates more than three interactions with other human proteins in the structural interaction network, whereas a generic endogenous interface only mediates ∼1.5 interactions with other human proteins. These observations demonstrate that viral proteins tend to mimic and hijack high-level regulatory components of the host cellular circuitry, by efficiently binding to existing endogenous interfaces rather than creating entirely new interfaces. Furthermore, endogenous interfaces mimicked by viral proteins tend to evolve more quickly than other endogenous interfaces, suggesting an evolutionary "arms race" between host and pathogen. Overall, 3D structural analysis revealed, in a systematic and statistically rigorous way, distinct principles governing antagonism versus cooperation in host–pathogen and within-host PPI networks (Franzosa and Xia, 2011).

Protein–protein interactions can be divided into two classes: the first class involves PPIs mediated by interactions between two globular domains, and the second class involves PPIs mediated by short linear motifs interacting with globular domains. Both classes are important mediators of host–pathogen interactions (Davey et al., 2011; Franzosa and Xia, 2011). A recent survey revealed extensive mimicry of host short linear motifs by viruses (Davey et al., 2011). Viral mimicry of host linear motifs was found for 52 of the ∼150 motif classes in the Eukaryotic Linear Motif (ELM) database (Gould et al., 2010), 13 of which have solved 3D structures involving viral motifs in complex with their host targets. For example, there are many cases of viral proteins targeting the SH3, SH2, or PDZ domains of host proteins using mimicked motifs. These observations are in agreement with the requirements for viral proteins to extensively hijack and manipulate diverse host proteins and pathways, despite the severe spatial constraints imposed by their small genomes (Davey et al., 2011). These motifs tend to cluster into hotspots in the viral genome (Sarmady et al., 2011), and they may be important determinants of virulence (Yang, 2012). While motifs play an important role in the biology of viruses and viruses use motifs extensively, it is not known if viruses use motifs more often than the host (Davey et al., 2011). These findings collectively highlight the feasibility and importance of structural systems biology in host–pathogen interaction research.

CONCLUSION

Despite being a relatively new field, between-species structural systems biology has already provided major insights into species interactions and infectious disease. We expect to see rapid growth in between-species structural systems biology over the next few years on the following fronts. First, host–pathogen physical, genetic, and functional interaction datasets will continue to accumulate for more pathogens, and with higher coverage and accuracy. The impact of these interactions on host and pathogen physiology will continue to be systematically evaluated. In addition to interaction data, small-scale experiments and large-scale technologies such as genome-wide association studies (Khor and Hibberd, 2012) have generated large amounts of data describing mutations that affect host–pathogen interaction and pathogenicity. A key computational challenge is the development of unified, predictive models of how host and pathogens interact through integration of these datasets. Second, the success of homology modeling depends critically on the availability of 3D structural templates for representative proteins and PPIs solved by experimental structural biologists. The power of homology modeling is especially limited for fast-evolving pathogens such as viruses, where experimental structural biology plays a central role. It is encouraging that the number of 3D structures of human–virus PPIs have doubled in the past 5 years (Franzosa and Xia, 2012), and we expect a significant expansion in the number of 3D structures for host–pathogen PPIs in the next few years. Structural genomics has been highly successful by focusing primarily on structure determination of single proteins (Chandonia and Brenner, 2006). It will be fascinating to investigate if high-throughput structural biology can be applied to within- and between-species PPIs as well. Finally, new methods will be developed to integrate interaction datasets with 3D structure datasets. Computational analysis of the resulting structural interaction networks will uncover new system-level insights into host–pathogen interactions.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation (CCF-1219007) to Yu Xia. Sara Garamszegi was supported by a National Science Foundation Graduate Research Fellowship (DGE-0741448).

REFERENCES

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., et al. (2004). Structure-based assembly of protein complexes in yeast. *Science* 303, 2026–2029.

Aloy, P., and Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* 22, 1317–1321.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., et al. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 38, D525–D531.

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.

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

Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., et al. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921–926.

Bushman, F. D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T. L., et al. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 5:e1000437. doi: 10.1371/journal.ppat.1000437

Calderwood, M. A., Venkatesan, K., Xing, L., Chase, M. R., Vazquez, A., Holthaus, A. M., et al. (2007). Epstein–Barr virus and virus human protein interaction maps. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7606–7611.

Chandonia, J. M., and Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science* 311, 347–351.

Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res.* 37, D669–D673.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544.

Dalton, J. A., and Jackson, R. M. (2007). An evaluation of automated homology modelling methods at low target template sequence similarity. *Bioinformatics* 23, 1901–1908.

Davey, N. E., Trave, G., and Gibson, T. J. (2011). How viruses hijack cell regulation. *Trends Biochem. Sci.* 36, 159–169.

Davis, F. P., Barkan, D. T., Eswar, N., Mckerrow, J. H., and Sali, A. (2007). Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.* 16, 2585–2596.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugue, S., et al. (2008). Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230.

Dethlefsen, L., Mcfall-Ngai, M., and Relman, D. A. (2007). An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449, 811–818.

Doolittle, J. M., and Gomez, S. M. (2011). Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Negl. Trop. Dis.* 5:e954. doi: 10.1371/journal.pntd.0000954

Driscoll, T., Dyer, M. D., Murali, T. M., and Sobral, B. W. (2009). PIG – the pathogen interaction gateway. *Nucleic Acids Res.* 37, D647–D650.

Durmus Tekir, S., Cakir, T., and Ulgen, K. O. (2012). Infection strategies of bacterial and viral pathogens through pathogen–human protein–protein interactions. *Front. Microbiol.* 3:46. doi: 10.3389/fmicb.2012.00046

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2007). Computational prediction of host–pathogen protein–protein interactions. *Bioinformatics* 23, i159–i166.

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.* 4:e32. doi: 10.1371/journal.ppat.0040032

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2011). Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect. Genet. Evol.* 11, 917–923.

Dyer, M. D., Neff, C., Dufford, M., Rivera, C. G., Shattuck, D., Bassaganya-Riera, J., et al. (2010). The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PLoS ONE* 5:e12089. doi: 10.1371/journal.pone.0012089

Fields, S., and Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246.

Finn, R. D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21, 410–412.

Franzosa, E. A., and Xia, Y. (2008). Structural perspectives on protein evolution. *Annu. Rep. Comput. Chem.* 4, 3–21.

Franzosa, E. A., and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* 26, 2387–2395.

Franzosa, E. A., and Xia, Y. (2011). Structural principles within the human–virus protein–protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10538–10543.

Franzosa, E. A., and Xia, Y. (2012). Structural models for host–pathogen protein–protein interactions: assessing coverage and bias. *Pac. Symp. Biocomput.* 287–298.

Fu, W., Sanders-Beer, B. E., Katz, K. S., Maglott, D. R., Pruitt, K. D., and Ptak, R. G. (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.* 37, D417–D422.

Gould, C. M., Diella, F., Via, A., Puntervoll, P., Gemund, C., Chabanis-Davidson, S., et al. (2010). ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.* 38, D167–D180.

Gray, J. J. (2006). High-resolution protein–protein docking. *Curr. Opin. Struct. Biol.* 16, 183–193.

Gulbahce, N., Yan, H., Dricot, A., Padi, M., Byrdsong, D., Franchi, R., et al. (2012). Viral perturbations of host networks reflect disease etiology. *PLoS Comput. Biol.* 8:e1002531. doi: 10.1371/journal.pcbi.1002531

Jager, S., Cimermancic, P., Gulbahce, N., Johnson, J. R., Mcgovern, K. E., Clarke, S. C., et al. (2012). Global landscape of HIV-human protein complexes. *Nature* 481, 365–370.

Kar, G., Gursoy, A., and Keskin, O. (2009). Human cancer protein–protein interaction network: a structural perspective. *PLoS Comput. Biol.* 5:e1000601. doi: 10.1371/journal.pcbi.1000601

Karlas, A., Machuy, N., Shin, Y., Pleissner, K. P., Artarini, A., Heuer, D., et al. (2010). Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature* 463, 818–822.

Khadka, S., Vangeloff, A. D., Zhang, C., Siddavatam, P., Heaton, N. S., Wang, L., et al. (2011). A physical interaction network of dengue virus and human proteins. *Mol. Cell. Proteomics* 10, M111.012187.

Khor, C. C., and Hibberd, M. L. (2012). Host–pathogen interactions revealed by human genome-wide surveys. *Trends Genet.* 28, 233–243.

Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314, 1938–1941.

Kim, P. M., Sboner, A., Xia, Y., and Gerstein, M. (2008). The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.* 4, 179.

Konig, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H. H., Bhattacharyya, S., et al. (2010). Human host factors required for influenza virus replication. *Nature* 463, 813–817.

Konig, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M., Irelan, J. T., et al. (2008). Global analysis of host–pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 135, 49–60.

Krishnan, M. N., Ng, A., Sukumaran, B., Gilfoy, F. D., Uchil, P. D., Sultana, H., et al. (2008). RNA interference screen for human genes associated with West Nile virus infection. *Nature* 455, 242–245.

Kundrotas, P. J., Zhu, Z., Janin, J., and Vakser, I. A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9438–9441.

MacPherson, I., Dickerson, J. E., Pinney, J. W., and Robertson, D. L. (2010). Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput. Biol.* 6:e1000863. doi: 10.1371/journal.pcbi.1000863

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C., and Gough, J. (2004). The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* 32, D235–D239.

Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* 29, 291–325.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* 15, 285–289.

Mukhtar, M. S., Carvunis, A. R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., et al. (2011). Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* 333, 596–601.

Murali, T. M., Dyer, M. D., Badger, D., Tyler, B. M., and Katze, M. G. (2011). Network-based prediction and analysis of HIV dependency factors. *PLoS Comput. Biol.* 7:e1002164. doi: 10.1371/journal.pcbi.1002164

Navratil, V., De Chassey, B., Combe, C. R., and Lotteau, V. (2011). When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Syst. Biol.* 5:13. doi: 10.1186/1752-0509-5-13

Navratil, V., De Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., Andre, P., et al. (2009). VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks. *Nucleic Acids Res.* 37, D661–D668.

Navratil, V., De Chassey, B., Meyniel, L., Pradezynski, F., Andre, P., Rabourdin-Combe, C., et al. (2010). System-level comparison of protein–protein interactions between viruses and the human type I interferon system network. *J. Proteome Res.* 9, 3527–3536.

Pichlmair, A., Kandasamy, K., Alvisi, G., Mulhern, O., Sacco, R., Habjan, M., et al. (2012). Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* 487, 486–490.

Pratt, D. J., Bentley, J., Jewsbury, P., Boyle, F. T., Endicott, J. A., and Noble, M. E. (2006). Dissecting the determinants of cyclin-dependent kinase 2 and cyclin-dependent kinase 4 inhibitor selectivity. *J. Med. Chem.* 49, 5470–5477.

Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J., and Weston, J. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics* 26, i645–i652.

Qian, J., Luscombe, N. M., and Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* 313, 673–681.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* 17, 1030–1032.

Rozenblatt-Rosen, O., Deo, R. C., Padi, M., Adelmant, G., Calderwood, M. A., Rolland, T., et al. (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* 487, 491–495.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud,

M., et al. (2004). A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.* 14, 313–324.

Russo, A. A., Tong, L., Lee, J. O., Jeffrey, P. D., and Pavletich, N. P. (1998). Structural basis for inhibition of the cyclin-dependent kinase Cdk6 by the tumour suppressor p16INK4a. *Nature* 395, 237–243.

Sarmady, M., Dampier, W., and Tozeren, A. (2011). HIV protein sequence hotspots for crosstalk with host hub proteins. *PLoS ONE* 6:e23293. doi: 10.1371/journal.pone.0023293

Schleker, S., Sun, J., Raghavan, B., Srnec, M., Muller, N., Koepfinger, M., et al. (2012). The current *Salmonella*–host interactome. *Proteomics Clin. Appl.* 6, 117–133.

Schlieker, C., Weihofen, W. A., Frijns, E., Kattenhorn, L. M., Gaudet, R., and Ploegh, H. L. (2007). Structure of a herpesvirus-encoded cysteine protease reveals a unique class of deubiquitinating enzymes. *Mol. Cell* 25, 677–687.

Shapira, S. D., Gat-Viks, I., Shum, B. O., Dricot, A., De Grace, M. M., Wu, L., et al. (2009). A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 139, 1255–1267.

Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.* 33, D413–D417.

Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. *Pac. Symp. Biocomput.* 516–527.

Thieu, T., Joshi, S., Warren, S., and Korkin, D. (2012). Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* 28, 867–875.

Tobe, T., Beatson, S. A., Taniguchi, H., Abe, H., Bailey, C. M., Fivian, A., et al. (2006). An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14941–14946.

Uetz, P., Dong, Y. A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., et al. (2006). Herpesviral protein networks and their interaction with the human proteome. *Science* 311, 239–242.

van Dijk, D., Ertaylan, G., Boucher, C. A. B., and Sloot, P. M. A. (2010). Identifying potential survival strategies of HIV-1 through virus–host protein interaction networks. *BMC Syst. Biol.* 4:96. doi: 10.1186/1752-0509-4-96

von Schwedler, U. K., Stuchell, M., Muller, B., Ward, D. M., Chung, H. Y., Morita, E., et al. (2003). The protein network of HIV budding. *Cell* 114, 701–713.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164.

Wuchty, S. (2011). Computational prediction of host–parasite protein interactions between *P. falciparum* and *H. sapiens*. *PLoS ONE* 6:e26960. doi: 10.1371/journal.pone.0026960

Wuchty, S., Siwo, G., and Ferdig, M. T. (2010). Viral organization of human proteins. *PLoS ONE* 5:e11796. doi: 10.1371/journal.pone.0011796

Xie, L., Xie, L., and Bourne, P. E. (2011). Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.* 21, 189–199.

Yang, C. W. (2012). A comparative study of short linear motif compositions of the influenza A virus ribonucleoproteins. *PLoS ONE* 7:e38637. doi: 10.1371/journal.pone.0038637

Zhang, L., Villa, N. Y., Rahman, M. M., Smallwood, S., Shattuck, D., Neff, C., et al. (2009). Analysis of vaccinia virus–host protein–protein interactions: validations of yeast two-hybrid screenings. *J. Proteome Res.* 8, 4311–4318.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., et al. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 490, 556–560.

Conflict of Interest Statement: 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.

Received: 09 September 2012; paper pending published: 28 September 2012; accepted: 06 December 2012; published online: 21 December 2012.

Citation: Franzosa EA, Garamszegi S and Xia Y (2012) Toward a three-dimensional view of protein networks between species. Front. Microbio. ***3****:428. doi: 10.3389/fmicb.2012.00428*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

PERSPECTIVE ARTICLE
published: 06 November 2012
doi: 10.3389/fmicb.2012.00387

In silico 3D structure analysis accelerates the solution of a real viral structure and antibodies docking mechanism

Motohiro Miki [1,2] ***and Kazuhiko Katayama*** [1] *******

[1] *Department of Virology II, National Institute of Infectious Diseases, Tokyo, Japan*
[2] *Denka-Seiken Co., Ltd, Niigata, Japan*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan
Sam-Yong Park, Yokohama City University, Japan

****Correspondence:***
Kazuhiko Katayama, Department of Virology II, National Institute of Infectious Diseases, Tokyo 208-0011, Japan.
e-mail: katayama@nih.go.jp

Norwalk virus (NoV) is responsible for most outbreaks of non-bacterial gastroenteritis. NoV is genetically diverse and show antigenically variable. Recently, we produced a monoclonal antibody called 5B-18 that reacts broadly with NoV genogroup II (GII). We suspected the 5B-18 binds to a conformational epitope on 3D structure of virion. X-ray crystallography showed us that 5B-18 binds to NoV at the P domain, which protrudes from the capsid surface of the virion. However, there seems to be no space that would allow the IgG to approach the virion. To solve this problem, we used cryo-electron microscopy to examine NoV GII virus-like particles (VLPs). The P domain rises up higher in NoV GII than in NoV GI, and it seems to form an outer layer around the virion. Finally, using *in silico* modeling we found the 5B-18 Fab arms and NoV P region are quite flexible, so that 5B-18 can bind the NoV virion from bottom of P domain. This study demonstrates the shortcomings of studying biological phenomenon by only one technique. Each method has limitations. Multiple methods and modeling *in silico* are the keys to solving structural problems.

Keywords: Norwalk virus, monoclonal antibody, x-ray crystallography, *in silico* modeling, cryo-electron microscopy

THE BASICS OF NORWALK VIRUS

Norwalk virus (NoV) is responsible for most of the outbreaks of non-bacterial gastroenteritis in developed countries and, it is thought, in developing countries as well. Yet, although NoV was identified more than 30 years ago, we know little about their pathogenicity and basic virology (Guix et al., 2007). Studies of NoV have been hampered by the lack of a cell-culture system or a small animal model in which the virus will grow, except murine norovirus that is classified as NoV genogroup V (Wobus et al., 2006).

NoV belongs to the family Caliciviridae. The genus *Norovirus* has only one species, *Norwalk viruses,* with five genogroups (GI–GV). Genogroup GI and II cause most human infections, and they are further subdivided into numerous genotypes (GI.1–8 and GII.1–17; Zheng et al., 2006). The NoV genome is a 7.3 to 7.7-kb positive-sense, polyadenylated, single-stranded RNA molecule. It contains three open reading frames (ORFs): ORF1 encodes a non-structural polyprotein, and ORF2 and ORF3 encode the major and minor capsid proteins, VP1 and VP2, respectively (Jiang et al., 1992; Lambden et al., 1993).

Without an *in vitro* system for propagating the virus, the antigenicity of NoV has been inferred from studies of virus-like particles (VLPs). Nucleic acid-free VLPs self-assemble when the capsid protein is expressed in a baculovirus expression system (**Figure 1A**). The VLPs are assumed to have a similar morphology and, thus, antigenicity as that of the native virions (Jiang et al., 1992). Cryo-electron microscopy (cryo-EM) and x-ray crystal structures of the prototype norovirus VLP (GI.1, Norwalk virus) showed that the VLPs form a T = 3 icosahedral structure (Prasad et al., 1994, 1999).

However, structures of large protein complexes are difficult to determine by x-ray analysis. We sought to understand the structure of the virion and how it interacts with antibodies by combining data from x-ray diffraction, cryo-EM, and *in silico* modeling.

A MONOCLONAL ANTIBODY REACTS BROADLY WITH NoV GII

NoV is generally detected by RT-PCR with degenerate primers or an ELISA with NoV-specific antibodies. Many polyclonal and monoclonal antibodies used in the ELISA kits were developed in mice or rabbit immunized with norovirus VLPs (Hansman et al., 2011).

Recently, we produced a monoclonal antibody called 5B-18 that reacts broadly with NoV GII (Hansman et al., 2012). In fact, 5B-18 is used as a NoV GII broad-range capture antibody in a commercial ELISA kit [NV-AD(III) SEIKEN NoV antigen ELISA] and in an immunochromatography (IC) kit (Quick naviNoro IC kit, both from Denka-Seiken, Japan).

The 5B-18 monoclonal antibody was produced by immunizing a mouse with norovirus VLPs. Several monoclonal antibodies bind to the shell (S) domain (Yoda et al., 2003; Li et al., 2010), and others bind to the protruding (P) domain (Lindesmith et al., 2012). We suspect that 5B-18 also binds to S or P domain on the surface of the NoV virion. However, no high-resolution structural details of the antibody binding to the VLPs, S domain or P domain are available.

X-RAY CRYSTALLOGRAPHY OF THE BINDING SITE

The 5B-18 binds major NoV genotypes, such as GII.4 and GII.3, and the minor NoV genotypes GII.10 and GII.12 strongly. We suspect 5B-18 binds to a conserved epitope on the NoV capsid surface. We wanted to define the recognition site of 5B-18 and the NoV minor genotype GII.10 P domain, and we began with x-ray

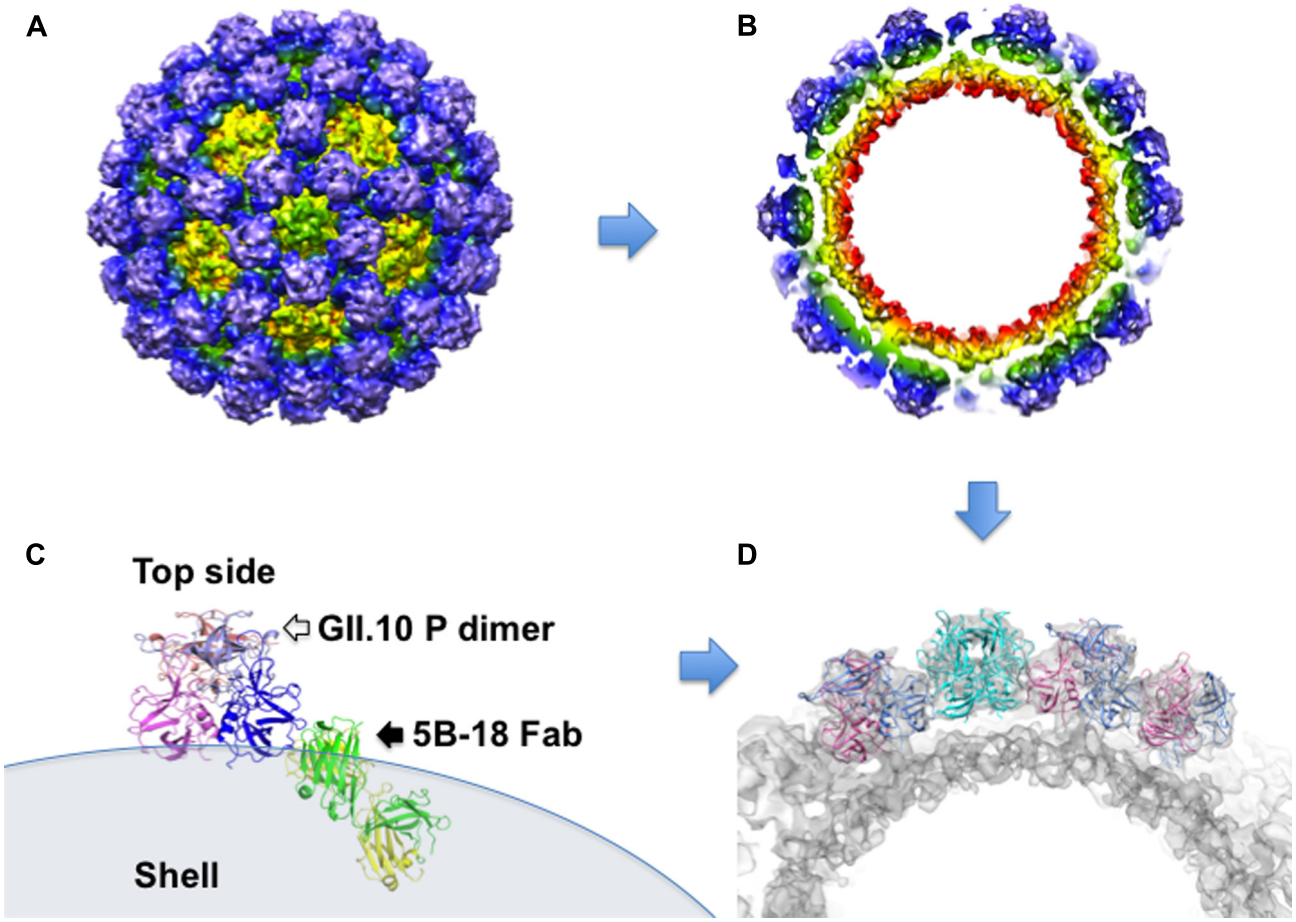


FIGURE 1 | The cryo-EM image of GII.10 VLP and the P dimer structure fit to the cryo-EM image by *in silico* modeling. The cryo-EM structure of NoV GII.10 VLP was constructed with 180 molecules of VP1 protein **(A)**. The cross-section of GII.10 VLP **(B)**. P domains (blue and green) rise up from shell domain (yellow), and it seemed to construct an outer layer of the capsid. Diagram based on the x-ray crystal structure of the GII.10 P domain-Fab complex shows Fab bound to the lower side of the P1 sub domain. Top side indicates the outside of the VLP, and the shell side contains the VLP core that is called shell. A white arrow indicates the GII.10 P dimer; the monomer is colored blue and purple. Black arrow shows MoAb 5B-18 Fab region. Green represents the heavy chain, and yellow represents the kappa chain. The part of oval shadow represents VLP shell **(C)**. The x-ray crystallography was fit into the cryo-EM structure by *in silico* modeling **(D)**. The GII.10 P domain (apo P domain structure) into the A/B dimer subunit (light blue and pink, respectively) and C/C dimer subunit (cyan).

crystallography, one of the gold standard for protein structural studies. We expressed the NoV GII.10 P domain in the *Escherichia coli* strain BL21 (DE3). The P domain was purified and stored in gel filtration buffer. Next we prepared of 5B-18 Fab fragment by immunizing a mouse with NoV GII.4-strain 445 VLPs (GenBank accession number DQ093064; Denka-Seiken, Japan). To prepare crystals of the bound complex, purified GII.10 P domain and Fab were mixed in a 1.4:1 ratio. Crystals were grown by the hanging-drop vapor-diffusion method, mixing the protein and reservoir solution (40% [vol/vol] polyethylene glycol [PEG] 400, 5% [wt/vol] PEG 3350, and 0.1 M acetic acid, pH 5.5) in a 1:1 ratio. Crystals grew over 1 week at 20°C.

One GII.10 P domain-Fab complex crystal diffracted x-rays to a resolution 3.3Å, and we solved the structure by molecular replacement with a GII.10 P domain monomer (PDB ID 3ONU) and a mouse Fab (PDB ID 1WEJ) as search models. Molecular replacement indicated an asymmetrical unit contained two P domain monomers and two 5B-18 Fabs, each with a kappa and a heavy chain (**Figure 1C**; Hansman et al., 2012).

The binding of the P domain and the Fab involved nine hydrogen bonds. Of these, eight linked the P1 subdomain to the kappa chain, and one linked the P1 subdomain and the heavy chain. More specifically, the amino acids in the P1 subdomain amino acids that interacted with the 5B-18 Fab were as follows (in each pair, the amino acids are for the P1 domain and Fab, respectively): Tyr533 and Tyr92 (one bond), Thr534 and Gly93 (three bonds), Thr534 and Trp97 (one bond), Leu535 and Tyr32 (one bond), Glu496 and Tyr92 (one bond), and Asn530 and Ser94 (one bond). Finally, Val433 and Asn52 in the heavy chain formed one hydrogen bond (Hansman et al., 2012).

CONFIRMATION OF 5B-18 BINDING

With the x-ray crystallographic analysis, we found the 5B-18 antibody bound to a hidden site on the P domain that is located inside

of the shell of NoV particle. However, in a previous study, the NoV GI structure indicated that bottom of the P domain was completely covered by the shell of NoV particle (**Figure 1C**). If the structure of GII is the same as GI, then 5B-18 could not bind GII. These results presented an apparent paradox for the 5B-18 binding mechanism. To resolve the paradox, we set out to identify the binding residue in the capsid.

From the crystallographic analysis, we knew that the 5B-18 Fab formed hydrogen bonds with residues at three sites in the P1 subdomain, called A, B, and C (**Figure 2A**). By aligning the amino acid sequences of representatives from NoV GII genotypes, we discovered that Val433 (site A) was the most variable. Other genotypes had threonine, serine, asparagine, leucine, or methionine at this position. Thr534 (site C) was mostly conserved: the only other amino acid at this position was a serine. Glu496 (site B), Asn530 (site C), Tyr533 (site C), and Leu535 (site C) were all highly conserved among the representative GII genotypes.

To confirm that 5B-18 binds the A, B, and C regions, we divided the GII.10 capsid domain into three major subdomains: N, S, P1-1 P2, and P1-2. We prepared five constructs (1–5), expressed them in an *E. coli* expression system, and identified a liner epitope of 5B-18 by western blotting (**Figure 2B**). Construct 3, a P1-2 region (i.e., A, B, and C), showed the strongest band signal, and construct 5 (i.e., B and C) showed a positive band. The intensity of the band from construct 5 was only about half the strength of construct 3 because it did not contain epitope A. However, construct 4 included only A, and constructs 1 and 2 also were not detected. Thus, the three 5B-18 epitopes A, B, and C were confirmed to be part of the binding epitope.

Next, we determined if 5B-18 binds to other NoV GII VLPs (**Figure 2A**). We prepared and purified six kinds of GII VLPs that were 809 (GII.3), 104 (GII.4), 445 (GII.6), 026 (GII.10), Hiro (GII.12), and GII.13 VLPs as aligned in **Figure 2A**. The GII VLPs that had all 5B-18 epitopes A, B, and C were captured by the anti-GII VLPs rabbit serum that was pre-coated on ELISA plate and detected with 5B-18 and horseradish peroxidase (HRP)-labeled anti-mouse IgG secondary antibodies. When the cut-off value was under 0.2, 5B-18 detected all kinds of GII VLPs in a dose-dependent manner (data not shown). These results suggested that 5B-18 binds to a variety of GII VLPs. In fact, the commercial ELISA and IC kits use 5B-18 (Denka-Seiken, Japan), and we have practical results showing that 5B-18 detects various infectious NoV GII in stool samples.

COMBINING CRYO-EM AND *IN SILICO* MODELING TO SOLVE A PARADOX

We had a simple question. Are the structures of NoV GI VLP and GII VLP the same or not? For G1 VLP, there is no space where the 5B-18 can access and bind the bottom of P domain. If the GII VLP had same conformation as the GI VLP, the lower part of the P domain would be buried under the virion shell (**Figure 1C**). However, 5B-18 binds and detects GII VLPs and GII infectious viruses. These conflicting facts suggested that the structures of the GII VLPs and infectious GII virions were different than the GI VLP structure. However, structure determinations by x-ray crystallography have many challenges and limitations, and we suspected this might be one of those cases.

To answer the question, we turned to cryo-EM and *in silico* modeling. We reconstructed the overall structure of GII.10 VLPs and 5B-18 Fabs from the x-ray structural data. To determine if the GII VLP had enough space to allow binding, we also used *in silico* modeling to fit the P and 5B-18 Fabs structures that had been derived by x-ray crystallography.

The GII.10 VLPs formed homogeneous, monodisperse particles in ice. By reference-free class averages and at 10Å resolution (0.5 FSC criterion), these icosahedral particles had several notable features, including spike-like structures extending from the vertices (**Figure 1B**), and at the three- and fivefold axes, significant amounts of the surface of the S domain were exposed (**Figure1A**). The GII.10 VLP P domain formed a second outer shell that seems to be separated from the S domain by about 15Å (**Figure 1B**). Thus, unlike the GI VLPs and virions, GII VLPs and virions seem to have a space between the shell and bottom of P domain, indicating that the two genotypes have different structures. Furthermore, the electron density was much weaker at the tip of the P domain (the P2 subdomain) than at the base. This observation is consistent with published reconstructions of calicivirus particles (Bhella et al., 2008; Bhella and Goodfellow, 2011) and indicates that the P domains have considerable heterogeneity.

Next we attempted to fit the GII.10 P domain and P domain-Fab complex structures into the GII.10 VLP cryo-EM structure. At 10Å resolution, the GII.10 P domain monomers in the VLP were easily distinguished. We manually fitted the crystal structures of the GII.10 P domain and P domain-Fab complex into the GII.10 VLP cryo-EM map, using published reports of GV.1 P domain dimers and the GV.1 cryo-EM map (Taube et al., 2010) as guides.

We refined the approximate alignment with the Fit-in-Map function in UCSF Chimera (Pettersen et al., 2004) to a cross-correlation coefficient of 0.94 (**Figure 1D**) with excellent results. The x-ray structure of the GII.10 P domain dimer (PDB ID 3ONU) unambiguously fitted the corresponding density in the cryo-EM map (**Figure 1D**). Only some loops of the P2 subdomain did not fit. They had only weak electron density, and their tips were less ordered than the S domain and P1 domains in the cryo-EM reconstruction. These subdomains are probably more flexible. P1, but not the P2, subdomains in the VLP appeared to be connected to the P domain dimers.

Next we fitted the x-ray structure of the P domain from the P domain-Fab complex into the reconstructed A/B dimer subunit and found that the 5B-18 binding site was close to an adjacent dimer of P domain (**Figure 1C**). At the twofold axes, the 5B-18 Fab was hindered by the S domain, which also provided an obstacle to assembly of the neighbor P domain dimer. However, when the P domain was fitted into the C/C dimer subunit, the 5B-18 Fab was in contact with the P domain dimer and slightly interfered with part of the S domain at the fivefold axes. Thus, the antibody binding site overlapped with part of the P1 subdomain.

Thus, this model predicted an unstable structure in which the VLP could not bind with the 5B-18 antibodies. How could this be? There are several possibilities. First, 5B-18 might bind at sites on the P domain that are only transiently exposed. Second, 5B-18 might bind to defects in the P domain. Finally, the Fab arms of 5B-18 might be very flexible.

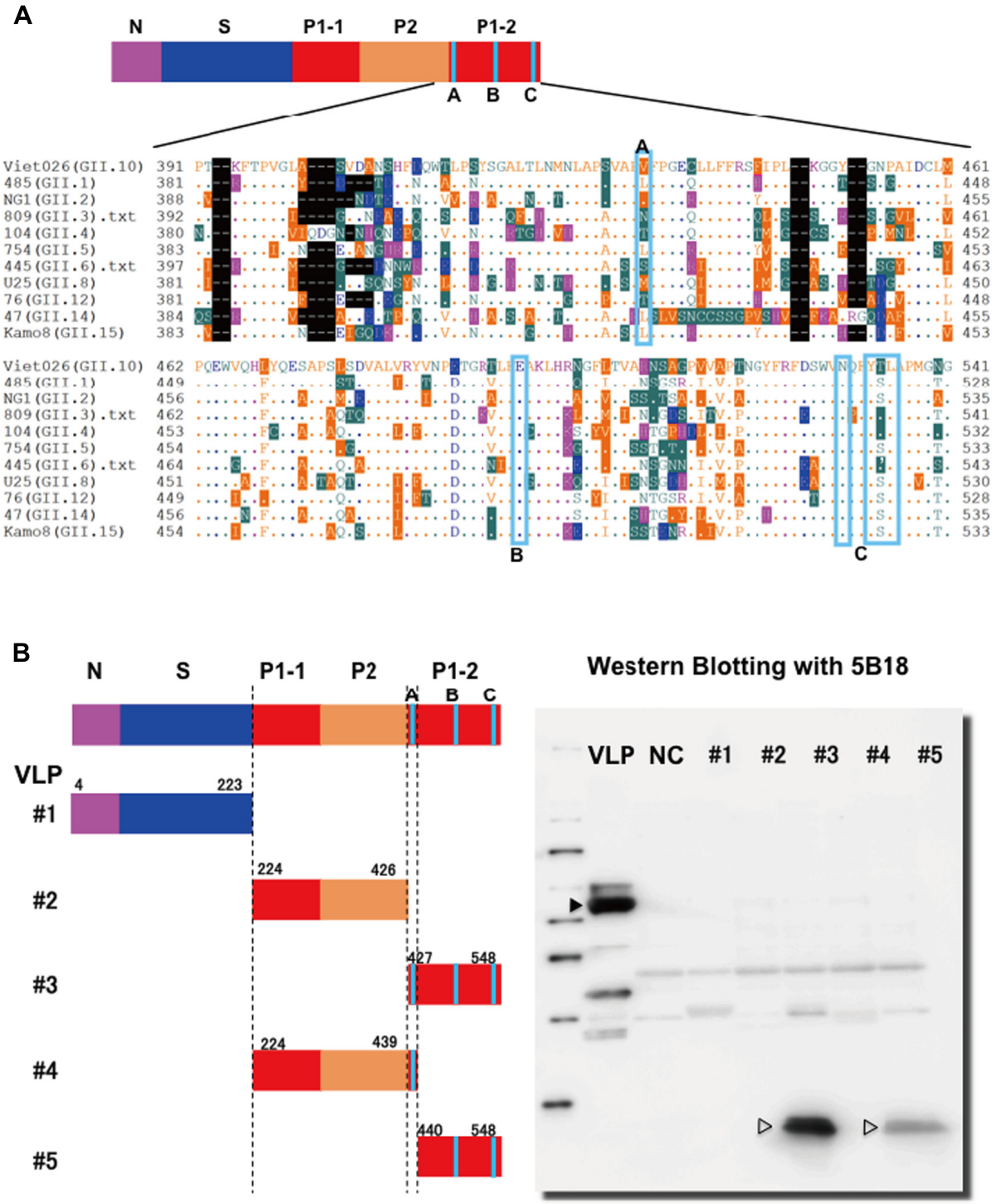


FIGURE 2 | Schematic diagram of NoV VP1 protein and the amino acid alignment of NoV GII VP1 sequences. Capsid (VP1) sequences from 11 GII genotypes were aligned, and the GII.10 capsid sequence was used as the consensus. The GII.10 P domain residues that interacted with the 5B-18 Fab involved three sites on the P domain termed A, B, and C. The six GII.10 P domain residues that interacted with the 5B-18 Fab are indicated by light blue rectangles **(A)**. Left panel of **(B)**: a schematic of the VP1 deletion mutants. The construct 1 includes the N-terminal region and shell domain of VP1, and amino acids 4–223 from the N-terminal methionine of VP1. The P1-1 and P2 domain constructs without binding site A are construct 2 (amino acids 224–426), with binding site A is construct 4 (amino acids 224–439). Construct 3 has all binding sites A, B and C (amino acids 427–548). Construct 5 deletes binding site A from construct 3 (amino acids 440–548). Right panel of **(B)**: western blotting results with 5B-18. Samples indicated as VLP, NC (negative control), mutant construct 1 (#1), #2, #3, #4 and #5. The black arrow represents VLP band at 58 kDa, and the white arrows represent #3 and #5 products at 13 kDa.

IgG flexibility is not unknown. For example, a neutralizing antibody 9C12 binds to hexon, the major coat protein of adenovirus, at a ratio of 240 antibody molecules to one virus particle or one antibody per hexon trimer (Varghese et al., 2004). By dynamic light scattering and negative-stain EM, electron-dense material coats the virus, but it was not aggregated at neutralizing ratios. In images reconstructed from cryo-EM, the viral surface was covered by electron density from the 9C12 antibody. Two Fab arms bridge two peripentonal hexons. One has a normal Fab shape and fitted the models well (Harris et al., 1998). The other arm has a somewhat distorted structure. A low-density tail extends to a third hexon that forms a minor alternate binding site. The normal arm binds to a unique site in the asymmetric unit of the virus. It has no alternate binding sites because a penton, rather than a hexon, is positioned at the icosahedral fivefold axis. In addition, the angle between the long axes of the Fabs was $<115°$ that was found in the uncomplexed IgG1 (Harris et al., 1998). Thus, flexibility is important for the bivalent binding of 9C12.

ESTIMATING THE FLEXIBILITY IN THE STRUCTURE

The findings from the 9C12 study were informative for our 5B-18 paradox. 5B-18 could reach the bottom of the P domain if the Fab domain could bend and escape the P1 subdomain or star-like structure on the shell. 5B-18 IgG bound equally well with intact and partially broken GII.10 VLPs. To determine if 5B-18 binds to intact or broken particles, we took advantage of a characteristic of norovirus VLPs: they are less stable and appear to be broken at high pHs (Ausar et al., 2006). Therefore, we looked at 5B-18 binding at different pHs. At low and neutral pHs (5.3, 6.3, and 7.3), the GII.10 VLPs were mostly homogenous in size and unbroken, but at higher pHs (8.3 and 9.3), they were less homogenous and partially broken. 5B-18 IgG bound to GII.10 VLPs at different pH values with nearly identical efficacies, regardless of the fraction of damaged particles. At pH 5.3, 6.3, and 8.3, the titer was 512,000. At pH 9.3, it was 1,024,000, and at pH 7.3, it was 2,048,000 (optical density cutoff of 0.2; Hansman et al., 2006). We also determined size distribution of the VLPs by dynamic light scattering in each pH conditions. VLPs were shown single peak on diameter 38 to 50 nm (data not shown). These results suggest that 5B-18 appears detects nominally intact GII.10 VLPs.

We studied the 5B-18 binding mechanism by x-ray crystallography, molecular virology, and cryo-EM. We combined the results in *in silico* modeling that simulates molecular dynamics and is a reliable method for revealing fluctuations in protein structure. Each technique complemented the other by filling in for data that was lacking from the others. Interestingly, the 5B-18 study suggests that VLP and viral virion have structure flexibility and that IgG molecule have flexible arms. They co-work each other and bind. *In silico* modeling is clearly a powerful tool for enhancing our understanding of basic viral processes.

ACKNOWLEDGMENTS

We thank P. Kwong and his lab members for their guidance in this work and for critical discussions about structural analysis and K. Nagayama and K. Murata for assistance with the cryo-electron microscopy and for discussions. We also thank H. Sato for giving us this great opportunity to publish our studies.

REFERENCES

Ausar, S. F., Foubert, T. R., Hudson, M. H., Vedvick, T. S., and Middaugh, C. R. (2006). Conformational stability and disassembly of Norwalk virus-like particles. Effect of pH and temperature. *J. Biol. Chem.* 281, 19478–19488.

Bhella, D., Gatherer, D., Chaudhry, Y., Pink, R., and Goodfellow, I. G. (2008). Structural insights into calicivirus attachment and uncoating. *J. Virol.* 82, 8051–8058.

Bhella, D., and Goodfellow, I. G. (2011). The cryo-electron microscopy structure of feline calicivirus bound to junctional adhesion molecule A at 9-angstrom resolution reveals receptor-induced flexibility and two distinct conformational changes in the capsid protein VP1. *J. Virol.* 85, 11381–11390.

Guix, S., Asanaka, M., Katayama, K., Crawford, S. E., Neill, F. H., Atmar, R. L., et al. (2007). Norwalk virus RNA is infectious in mammalian cells. *J. Virol.* 81, 12238–12248.

Hansman, G. S., Biertumpfel, C., Georgiev, I., McLellan, J. S., Chen, L., Zhou, T., et al. (2011). Crystal structures of GII.10 and GII.12 norovirus protruding domains in complex with histo-blood group antigens reveal details for a potential site of vulnerability. *J. Virol.* 85, 6687–6701.

Hansman, G. S., Natori, K., Shirato-Horikoshi, H., Ogawa, S., Oka, T., Katayama, K., et al. (2006). Genetic and antigenic diversity among noroviruses. *J. Gen. Virol.* 87(pt 4), 909–919.

Hansman, G. S., Taylor, D. W., McLellan, J. S., Smith, T. J., Georgiev, I., Tame, J. R., et al. (2012). Structural basis for broad detection of genogroup II noroviruses by a monoclonal antibody that binds to a site occluded in the viral particle. *J. Virol.* 86, 3635–3646.

Harris, L. J., Skaletsky, E., and McPherson, A. (1998). Crystallographic structure of an intact IgG1 monoclonal antibody. *J. Mol. Biol.* 275, 861–872.

Jiang, X., Graham, D. Y., Wang, K. N., and Estes, M. K. (1990). Norwalk virus genome cloning and characterization. *Science* 250, 1580–1583.

Jiang, X., Wang, M., Graham, D. Y., and Estes, M. K. (1992). Expression, self-assembly, and antigenicity of the Norwalk virus capsid protein. *J. Virol.* 66, 6527–6532.

Lambden, P. R., Caul, E. O., Ashley, C. R., and ClarkeI, N. (1993). Sequence and genome organization of a human small round-structured (Norwalk-like) virus. *Science* 259, 516–519.

Li, X., Zhou, R., Tian, X., Li, H., and Zhou, Z. (2010). Characterization of a cross-reactive monoclonal antibody against Norovirus genogroups I, II, III and V. *Virus Res.* 151, 142–147.

Lindesmith, L. C., Beltramello, M., Donaldson, E. F., Corti, D., Swanstrom, J., Debbink, K., et al. (2012) Immunogenetic mechanisms driving norovirus GII.4 antigenic variation. *PLoS Pathog.* 8, e1002705. doi: 10.1371/journal.ppat.1002705

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Mengm, E. C., et al. (2004). UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612.

Prasad, B. V., Hardy, M. E., Dokland, T., Bella, J., Rossmann, M. G., and Estes, M. K. (1999). X-ray crystallographic structure of the Norwalk virus capsid. *Science* 286, 287–290.

Prasad, B. V., Matson, D. O., and Smith, A. W. (1994). Three-dimensional structure of calicivirus. *J. Mol. Biol.* 240, 256–64.

Taube, S., Rubin, J. R., Katpally, U., Smith, T. J., Kendall, A., Stuckey, J. A., et al. (2010). High-resolution x-ray structure and functional analysis of the murine norovirus 1 capsid protein protruding domain. *J. Virol.* 84, 5695–705.

Varghese, R., Mikyas, Y., Stewart, P. L., and Ralston, R. (2004). Postentry neutralization of adenovirus type 5 by an antihexon antibody. *J. Virol.* 78, 12320–12332.

Wobus, C. E., Thackray, L. B., and Virgin, H. W. 4th. (2006). Murine norovirus: a model system to study norovirus biology and pathogenesis. *J. Virol.* 80, 5104–5112.

Yoda, T., Suzuki, Y., Terano, Y., Yamazaki, K., Sakon, N., Kuzuguchi, T., et al. (2003). Precise characterization of norovirus (Norwalk-like virus)-specific monoclonal antibodies with broad reactivity. *J. Clin. Microbiol.* 41, 2367–2371.

Zheng, D. P., Ando, T., Fankhauser, R. L., Beard, R. S., Glass, R. I., and Monroe, S. S. (2006). Norovirus classification and proposed strain nomenclature. *Virology* 346, 312–323.

Conflict of Interest Statement: 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.

Received: 14 August 2012; paper pending published: 24 August 2012; accepted: 18 October 2012; published online: 06 November 2012.

Citation: Miki M and Katayama K (2012) In silico 3D structure analysis accelerates the solution of a real viral structure and antibodies docking mechanism. Front. Microbio. ***3****:387. doi: 10.3389/fmicb.2012.00387*
This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 June 2012
doi: 10.3389/fmicb.2012.00206

Electrostatic potential of human immunodeficiency virus type 2 and rhesus macaque simian immunodeficiency virus capsid proteins

Katarzyna Bozek*[1], *Emi E. Nakayama*[2], *Ken Kono*[2] *and Tatsuo Shioda*[2]

[1] *Max Planck Institute for Informatics, Saarbrücken, Germany*
[2] *Department of Viral Infections, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka, Japan*

Edited by:
Masaru Yokoyama, National Institute of Infectious Diseases, Japan

Reviewed by:
Masako Nomaguchi, The University of Tokushima Graduate School, Japan
Masaru Yokoyama, National Institute of Infectious Diseases, Japan

***Correspondence:**
Tatsuo Shioda, Department of Viral Infections, Research Institute for Microbial Diseases, Osaka University, 3-1, Yamada-oka, Suita, Osaka 565-0871, Japan.
e-mail: shioda@biken.osaka-u.ac.jp

Human immunodeficiency virus type 2 (HIV-2) and simian immunodeficiency virus isolated from a macaque monkey (SIVmac) are assumed to have originated from simian immunodeficiency virus isolated from sooty mangabey (SIVsm). Despite their close similarity in genome structure, HIV-2 and SIVmac show different sensitivities to TRIM5α, a host restriction factor against retroviruses. The replication of HIV-2 strains is potently restricted by rhesus (Rh) monkey TRIM5α, while that of SIVmac strain 239 (SIVmac239) is not. Viral capsid protein is the determinant of this differential sensitivity to TRIM5α, as the HIV-2 mutant carrying SIVmac239 capsid protein evaded Rh TRIM5α-mediated restriction. However, the molecular determinants of this restriction mechanism are unknown. Electrostatic potential on the protein-binding site is one of the properties regulating protein–protein interactions. In this study, we investigated the electrostatic potential on the interaction surface of capsid protein of HIV-2 strain GH123 and SIVmac239. Although HIV-2 GH123 and SIVmac239 capsid proteins share more than 87% amino acid identity, we observed a large difference between the two molecules with the HIV-2 GH123 molecule having predominantly positive and SIVmac239 predominantly negative electrostatic potential on the surface of the loop between α-helices 4 and 5 (L4/5). As L4/5 is one of the major determinants of Rh TRIM5α sensitivity of these viruses, the present results suggest that the binding site of the Rh TRIM5α may show complementarity to the HIV-2 GH123 capsid surface charge distribution.

Keywords: HIV-2, SIVmac, capsid, TRIM5α, electrostatic potential, APBS, SAS

INTRODUCTION

The host range of human immunodeficiency virus type 1 (HIV-1) is narrow, limited to humans and chimpanzees (Gao et al., 1999). HIV-1 fails to replicate in activated CD4-positive T lymphocytes from Old World monkeys (OWM), such as rhesus (Rh; Shibata et al., 1995; Himathongkham and Luciw, 1996) and cynomolgus (CM) monkeys (Akari et al., 1996, 1999). On the other hand, simian immunodeficiency virus (SIV) isolated from sooty mangabey (SIVsm) and SIV isolated from African green monkey (SIVagm) replicate well in their natural hosts (VandeWoude and Apetrei, 2006). SIV isolated from a macaque monkey (SIVmac) evolved from SIVsm in captive macaques, and replicates efficiently in Rh (Shibata et al., 1995; Himathongkham and Luciw, 1996) and CM (Akari et al., 1996, 1999) monkeys. Human immunodeficiency virus type 2 (HIV-2) is assumed to have originated from SIVsm as the result of zoonotic events involving monkeys and humans (Hahn et al., 2000). Previous studies have shown that HIV-2 strains vary widely in their ability to grow in cells of OWM (Castro et al., 1990, 1991; Locher et al., 1998, 2003; Fujita et al., 2003).

TRIM5α was identified as an anti-HIV-1 host restriction factor in Rh monkey cells (Stremlau et al., 2004). TRIM5 proteins are members of the tripartite motif family containing RING, B-box, and coiled-coil domains. The α isoform of TRIM5 has an additional C-terminal PRYSPRY domain (Reymond et al., 2001). TRIM5α recognizes the multimerized capsid (viral core) of an incoming virus by its PRYSPRY domain and causes degradation of the core (Sebastian and Luban, 2005; Stremlau et al., 2006). In CM monkey, TRIM5α has also been shown to restrict HIV-1 infection (Nakayama et al., 2005).

We previously evaluated the sensitivity of HIV-2 and SIVmac to Rh and CM TRIM5αs, and found that HIV-2 strain GH123 carrying P at position 120 of the capsid protein (CA) was potently restricted by CM TRIM5α, while the HIV-2 GH123 mutant in which P was replaced with Q was resistant to CM TRIM5α (Song et al., 2007). In contrast, Rh TRIM5α potently restricted the replication of both viruses (Kono et al., 2008). Three amino acid residues, TFP, at positions 339–341 in the PRYSPRY domain of Rh TRIM5α were necessary for restricting HIV-2 strains that were resistant to CM TRIM5α (Kono et al., 2008). Although SIVmac239 CA possesses Q at position 118 corresponding to position 120 of GH123, SIVmac239 was resistant to both of CM and Rh TRIM5αs (Kono et al., 2008, 2010). Therefore, we attempted to identify the viral determinant of SIVmac239 underlying evasion from Rh TRIM5α-mediated restriction, and found that multiple regions including the N-terminal loop, a loop between α-helices 4 and 5 (L4/5), and a loop between α-helices 6 and 7 (L6/7) in

the N-terminal half of SIVmac239 CA are necessary for complete evasion of Rh TRIM5α restriction (Kono et al., 2010).

Apart from the sequence and structural characteristics regulating protein–protein interaction, the electrostatic potential at the binding site is an important factor allowing molecular interactions. The electrostatic potential on the protein surface is generated through redistribution of electrons according to local electrical fields. It is defined as the potential energy of a proton at a particular location near a molecule. Negative electrostatic potential results in attraction of the proton by the concentrated electron density. Positive electrostatic potential results in repulsion of the proton by the atomic nuclei in regions where low electron density exists and nuclear charge is incompletely shielded. Electrostatic effects were shown to be a major factor in determining the nature and strength of the interactions between protein surfaces (Dong and Zhou, 2002; Kortemme and Baker, 2002). A complementary charge on the binding site of both proteins may result in an attractive force allowing binding to occur.

In the present study, we analyzed the electrostatic potentials of the surface regions of the CA loop. We analyzed two CA variants, HIV-2 GH123 and SIVmac239 CAs, showing opposite restriction phenotypes. We first modeled the 3-D structures of the proteins by homology modeling and next calculated the electrostatic potentials in the regions of interest based on Adaptive Poisson–Boltzmann Solver and non-local electrostatic method. We found a large difference in the electrostatic potentials of the loop surface between the HIV-2 GH123 and SIVmac239 CAs, potentially responsible for the differential TRIM5α sensitivity of these two viruses.

MATERIALS AND METHODS

MODELING

The structure of the N-terminal domain of the HIV-1 CA (PDB number 1GWP; Tang et al., 2002) was used as a template for building the corresponding domain models of HIV-2 GH123 and SIVmac239 CAs. The models were built using Modeller 9v4 (Eswar et al., 2007) and visualized with PyMOL (http://www.pymol.org).

CALCULATION OF ELECTROSTATIC POTENTIALS

As the initial step preceding electrostatic potential modeling, we added missing hydrogen atoms and estimated the ionization (protonation) of the molecules. We used H++ server (Gordon et al., 2005) http://biophysics.cs.vt.edu/H++, which adds protons to the input structure according to the calculated ionization states at the specified pH of the solvent. The H++ method models molecules as a low dielectric medium ε_{in} in a solvent with a high dielectric constant ε_{out}. It additionally allows the user to define the salt concentration of the medium and its pH. We used the most biologically relevant parameters of human cells: pH = 7.2, salinity 1%, molecule dielectric $\varepsilon_{in} = 10$, and medium dielectric $\varepsilon_{in} = 80$. The dielectric parameters were chosen according to the suggestions of the authors of the H++ method as appropriate for modeling protonation of surface residues. We also inspected electrostatic potential profiles resulting from several other parameter combinations. Other parameter regimes did not produce markedly different electrostatic potentials in the region of interest. Therefore, we chose the initial parameters as the most relevant for biological settings.

We next applied two methods of electrostatic potential calculation: Adaptive Poisson–Boltzmann Solver (APBS; Baker et al., 2001) and non-local electrostatic method (Hildebrandt et al., 2007). In both methods, electrostatic properties are described by the Poisson–Boltzmann equation, a second-order non-linear partial differential equation. APBS method solves the equation using finite element techniques based on parameter discretization and iterative parallel refinement of the equation solution. The non-local electrostatic method allows inclusion of the structure of water molecules in the calculation and describes the system as a continuum. This method captures the effects of the dipole polarization of water molecules and the effects of the surrounding hydrogen bond network, and is therefore a more accurate model of the electrostatic potential estimations close to the molecule-solvent interface.

We used two different surface approximations: solvent-accessible surface (SAS) of two different sizes. SAS is the surface of a molecule that is accessible to a solvent. It is estimated using a "rolling ball" approach (Shrake and Rupley, 1973) in which a sphere of solvent of a particular radius is used to probe the surface of the molecule, the surface is then described by the center of the probing sphere. We used the approximate radius of a water molecule of 1.4 Å and an additional 3 Å to determine how the electrostatic potential changes with distance from the molecule.

RESULTS

THE 3-D STRUCTURAL MODELS OF HIV-2 GH123 AND SIVmac239 CA N-TERMINAL DOMAINS

Previously, we evaluated the sensitivity of HIV-2 GH123 and SIVmac239 to Rh and CM TRIM5αs, and found that HIV-2 GH123 was sensitive to CM and Rh TRIM5αs (Song et al., 2007; Kono et al., 2008; **Figure 1A**). In contrast, SIVmac239 was resistant to CM and Rh TRIM5αs (Kono et al., 2008, 2010; **Figure 1A**). CA is the determinant for this differential sensitivity to TRIM5α between HIV-2 GH123 and SIVmac239, as the HIV-2 GH123 mutant carrying SIVmac239 CA (HIV-2 GH/SCA) was also resistant to CM and Rh TRIM5αs (**Figure 1A**; Kono et al., 2010). Despite this marked difference in TRIM5α sensitivity between HIV-2 GH123 and SIVmac239, CA of these two viruses share more than 87% amino acid identity (**Figure 1B**). Therefore, we compared the structural properties of HIV-2 GH123 CA with those of SIVmac239.

We first constructed 3-D models of HIV-2 GH123 and SIVmac239 CA N-terminal domains by homology modeling. In the constructed models, HIV-2 GH123 and SIVmac239 CA N-terminal domains showed the most striking differences in shape of surface exposed loops (**Figure 2**). SIVmac239 CA is characterized by a more contracted shape as compared to the expanded loop structure of HIV-2 GH123. To confirm that this shape difference is not due to modeling noise, we remodeled both proteins using each one as a template for the other. The remodeled structures showed similar shape differences (data not shown), suggesting that the real structures differ.

ELECTROSTATIC POTENTIALS OF HIV-2 GH123 AND SIVmac239 CA N-TERMINAL DOMAINS

Figure 3 shows the distributions of calculated electrostatic potentials of HIV-2 GH123 and SIVmac239 CA N-terminal domains. We observed strong differences between the two molecules on

A

	CM TRIM5α	Rh TRIM5α
HIV-2 GH123	Sensitive	Sensitive
HIV-2 GH/SCA	Resistant	Resistant
SIVmac239	Resistant	Resistant

B

```
                N-terminal
HIV-2 GH123   1 PVQQTGGGNYIHVPLSPRTLNAWVKLVEDKKFGAEVVPGFQALSEGCTPYDINQMLNCVG    60
SIVmac239     1 ....I..-..V.L.............I.E.............................    59

                                                 L4/5
HIV-2 GH123  61 DHQAAMQIIREIINDEAADWDAQHPIPGPLPAGQLRDPRGSDIAGTTSTVEEQIQWMYRP   120
SIVmac239    60 ..........D...E......L...Q.A.-QQ....E.S.........S.D........Q   118

              L6/7
HIV-2 GH123 121 QNPVPVGNIYRRWIQIGLQKCVRMYNPTNILDVKQGPKEPFQSYVDRFYKSLRAEQTDPA   180
SIVmac239   119 ...I..........L.........................................A.   178


HIV-2 GH123 181 VKNWMTQTLLIQNANPDCKLVLKGLGMNPTLEEMLTACQGVGGPGQKARLM           231
SIVmac239   179 ........................V.......................              229
```

FIGURE 1 | (A) Sensitivities of HIV-2 GH123, HIV-2 GH123 mutant carrying SIVmac239 capsid protein (HIV-2 GH/SCA), and SIVmac239 to cynomolgus (CM) and rhesus (Rh) monkey TRIM5α. The replication of HIV-2 GH123 was potently restricted by CM and Rh TRIM5α (sensitive), while that of SIVmac239 and the HIV-2 GH123 mutant carrying SIVmac239 capsid was not (resistant). **(B)** Alignment of amino acid sequences of HIV-2 GH123 and SIVmac239 capsid proteins. Positions of the N-terminal loop (N-terminal), a loop between α-helices 4 and 5 (L4/5), and a loop between α-helices 6 and 7 (L6/7) are indicated above the amino acid sequences.

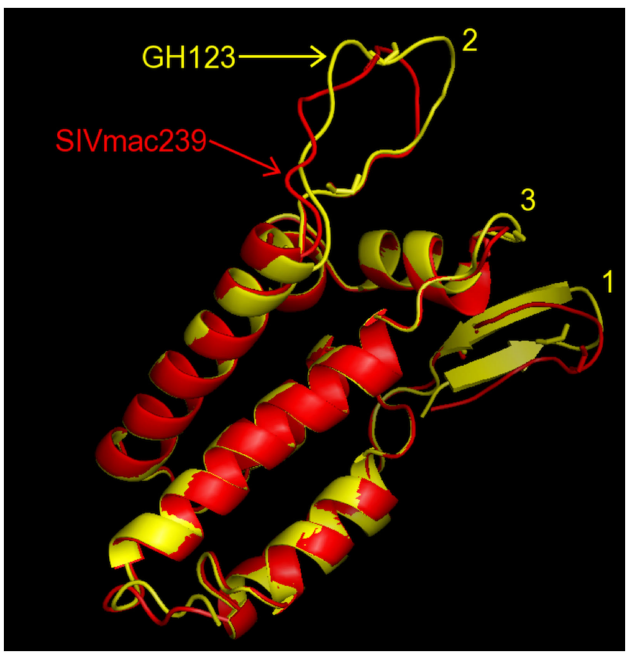


FIGURE 2 | Superposition of modeled structures of the N-terminal domains of HIV-2 GH123 (GH123, yellow) and SIVmac239 (red) capsid proteins. The three loops containing sites important for the TRIM5α interaction are numbered as follows: (1) N-terminal loop, (2) loop between α-helices 4 and 5 (L4/5), (3) loop between α-helices 6 and 7 (L6/7).

the surface of the loops with the GH123 molecule having predominantly positive and SIVmac239 predominantly negative electrostatic potential on this part of the surface (**Figure 3**).

To quantify this observation and obtain further insight into the specific region where the electrostatic potential differences are strong, we extracted the electrostatic potential values on the surfaces of the two molecules. From the electrostatic potential values estimated in a grid covering the entire space around the molecules, we extracted grid points neighboring the points of triangulation of each surface type. We grouped these electrostatic potential values according to the atoms of the closest loop residues. This comparison of grouped electrostatic potential values of corresponding residues in the two analyzed molecules allowed us to quantitatively confirm the differences in electrostatic potential in the region of interest and to point to specific residues around which the differences were stronger. The strongest difference in electrostatic potential between HIV-2 GH123 and SIVmac239 CAs was observed on the surface of L4/5, with HIV-2 GH123 and SIVmac239 showing positive and negative electrostatic potential, respectively. Eight of nine residues in this loop showed significant differences in mean electrostatic potential and clear separation of the electrostatic potential values on the grid neighboring to the loop residues by both local ABPS and non-local electrostatic methods (**Table 1**).

Residues in L6/7 showed weak but similar electrostatic potential differences to those of L4/5 by the local ABPS method, but these differences were not confirmed by the non-local electrostatic method (**Table 1**). The N-terminal loop showed the opposite pattern, with HIV-2 GH123 and SIVmac239 having negative and positive electrostatic potentials, respectively, according to the local APBS method (**Table 1**). However, the differences were smaller and were not confirmed by the non-local electrostatic method (**Table 1**).

Similar electrostatic potential differences, although spanning a narrower range of values than those described above, were

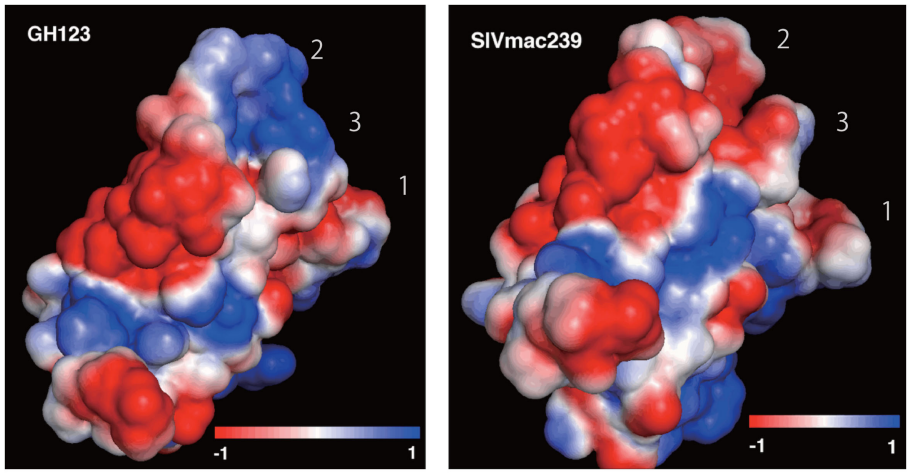


FIGURE 3 | Electrostatic potential on the surface of HIV-2 GH123 (GH123) and SIVmac239 capsid protein N-terminal domains. Structures are positioned as in **Figure 2** with the loops directed toward the upper right of the image. Electrostatic potential was calculated and visualized using the APBS plugin in PyMOL. The three loops containing sites important for the TRIM5α interaction are numbered as follows: (1) N-terminal loop, (2) loop between α-helices 4 and 5 (L4/5), (3) loop between α-helices 6 and 7 (L6/7).

Table 1 | Mean electrostatic potential on the surface surrounding residues of the N-terminal loop (N-terminal), the loop between α-helices 4 and 5 (L4/5), and the loop between α-helices 6 and 7 (L6/7) of HIV-2 GH123 and SIVmac239 CAs calculated using the local Adaptive Poisson–Boltzmann Solver (APBS) and non-local electrostatic methods.

	Residue (GH123/SIVmac239)	APBS			Non-local		
		HIV-2 GH123	SIVmac239	*p*-Value	HIV-2 GH123	SIVmac239	*p*-Value
N-terminal	5 THR/5 ILE	−0.206	0.064	<0.001	−1.049	−0.096	<0.001
	6 GLY/6 GLY	0.025	−0.196	0.006	0.787	−0.805	<0.001
	7 GLY/7 GLY	−0.315	0.024	<0.001	−1.283	−0.854	<0.001
	8 GLY/8 ASN	−0.420	0.066	<0.001	0.058	−1.092	0.406
	9 ASN/9 TYR	−0.463	−0.241	0.741	−5.668	2.697	<0.001
	10 TYR/10 VAL	−0.782	0.021	<0.001	−8.827	−1.367	<0.001
L4/5	88 GLY/87 ALA	0.147	−0.248	<0.001	2.906	−1.700	<0.001
	89 PRO/88 PRO	0.355	−0.522	<0.001	2.879	−0.524	<0.001
	90 LEU/–	−0.426	–	–	6.567	–	–
	91 PRO/89 GLN	0.603	−0.133	<0.001	6.543	−0.673	<0.001
	92 ALA/90 GLN	0.047	−0.051	<0.001	1.282	−0.418	<0.001
	93 GLY/91 GLY	−0.230	−0.269	0.076	−2.761	3.070	<0.001
	94 GLN/92 GLN	0.895	−0.735	<0.001	7.148	0.820	<0.001
	95 LEU/93 LEU	−0.958	−1.433	0.046	−6.661	2.234	<0.001
	96 ARG/94 ARG	0.090	−0.227	<0.001	5.805	−3.992	<0.001
	97 ASP/95 GLU	−0.045	−1.599	<0.001	−8.336	−3.481	0.001
L6/7	117 MET/115 MET	−0.765	0.799	<0.001	−6.437	−9.665	0.078
	118 TYR/116 TYR	0.070	−0.069	0.167	−5.037	0.055	<0.001
	119 ARG/117 ARG	1.022	0.405	<0.001	6.785	−2.802	<0.001
	120 PRO/118 GLN	−0.094	−0.706	<0.001	−5.178	3.904	<0.001
	121 GLN/119 GLN	0.802	−0.260	<0.001	4.308	0.340	<0.001
	122 ASN/120 ASN	0.119	−0.674	<0.001	−4.078	−6.824	0.003
	123 PRO/121 PRO	−0.782	−0.235	<0.001	−17.281	−11.590	<0.001
	124 VAL/122 ILE	−1.200	−1.906	<0.001	−6.233	−8.141	0.003
	125 PRO/123 PRO	−0.250	0.455	<0.001	−4.804	−12.468	<0.001

Color indicates significant difference ($p < 0.05$, Wilcoxon test) between the electrostatic potentials of the two molecules with positive electrostatic potential marked in blue and negative marked in red.

observed on the SAS of the 3 Å probe radius (data not shown). These observations reflect the electrostatic potential decrease with distance from the molecule surface.

DISCUSSION

In the present study, we constructed 3-D models of HIV-2 GH123 and SIVmac239 CA N-terminal domains by homology modeling and analyzed the electrostatic potential distributions on the SASs of these molecules. We observed a large difference between the HIV-2 GH123 and SIVmac239 CA N-terminal domains, with the HIV-2 GH123 molecule having predominantly positive and SIVmac239 predominantly negative electrostatic potential on the surface of L4/5. This result may be relevant to the previous findings that CA L4/5 was one of the major determinants for the differential sensitivity to Rh TRIM5α between HIV-2 and SIVmac239 (Ylinen et al., 2005; Kono et al., 2010).

Precise calculation of the interaction electrostatics is challenging due to the large surfaces involved and the large structural changes that can occur upon binding. Here, our quantitative approach based on two different methods for calculation of electrostatic potential indicated negative electrostatic potential on the surface of the resistant CA variant SIVmac239 and positive electrostatic potential of the non-resistant HIV-2 GH123 variant. The presence of positive electrostatic potential on the surface of L4/5 may therefore be a prerequisite for the interactions with Rh TRIM5α. This loop is the most outward pointing part of the CA protein. Complementarity to the HIV-2 GH123 surface charge distribution at the binding site of the host protein may be necessary for binding. Therefore, similar studies of TRIM5α surface electrostatic potentials could help to point to the specific site of this interaction, although the 3-D structural analysis of TRIM5α PRYSPRY domain is required for this goal.

It was recently reported that a recombinant TRIM5α protein carrying TRIM21 RING domain (TRIM5-21R) assembled to form 2-D paracrystalline hexagonal arrays *in vitro* (Ganser-Pornillos et al., 2011). This assembly requires RING and B-box 2 domains, and the hexagonal lattices of HIV-1 CA that mimic the surface of core act as template for stabilization of TRIM5-21R arrays in a PRYSPRY-dependent manner (Ganser-Pornillos et al., 2011). As the interaction between individual CA monomers and TRIM5α is very weak, CA recognition by TRIM5α is thought to be a synergistic combination of direct binding interactions with the PRYSPRY domain, higher-order assembly of TRIM5α, template-based assembly, and lattice complementarity. Therefore, the electrostatic potential might be the crucial determinant of this binding allowing TRIM5α for recognition of a broader range of CA sequence variants.

In addition to L4/5, our previous study revealed that the N-terminal loop and L6/7 in the N-terminal half of SIVmac239 CA are also necessary for complete evasion of Rh TRIM5α restriction (Kono et al., 2010). Electrostatic potentials of these 2 loops did not show large differences between HIV-2 GH123 and SIVmac239. Therefore, it is possible that a certain interaction other than the electrostatic interaction would be involved in binding of Rh TRIM5α PRYSPRY domain with the N-terminal loop and L6/7 of HIV-2 GH123.

On sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis, SIVmac239 CA is known to migrate at a molecular weight of 27 kDa, while HIV-2 GH123 CA migrates at a molecular weight of 25 kDa (Kono et al., 2010). However, the number of amino acid residues in SIVmac239 CA is smaller than that in HIV-2 GH123 CA (**Figure 1B**), and the molecular weight of SIVmac239 CA is therefore smaller than that of HIV-2 GH123. We reported previously that the amino acid sequences of L4/5 determined this differential electrophoretic mobility of CAs (Kono et al., 2010). The difference seems to be attributable to the presence of non-polar P and A residues at positions 91 and 92, respectively, in L4/5 of HIV-2 GH123 CA, where two more hydrophilic Q residues are located in SIVmac239 CA L4/5 (**Figure 1**; **Table 1**). In addition, HIV-2 GH123 CA L4/5 has a hydrophobic L insertion at position 90 (**Figure 1**; **Table 1**). Therefore, L4/5 of HIV-2 GH123 CA is more hydrophobic and would attract larger numbers of SDS molecules than that of SIVmac239 leading to accelerated electrophoretic speed of the CA. It is therefore possible that hydrophobic interactions between Rh TRIM5α and viral CAs would also be involved in determining the anti-viral specificity of TRIM5α in addition to the electrostatic interactions discussed above. Further biochemical studies of TRIM5α and viral CAs are necessary to address this question.

ACKNOWLEDGMENTS

We thank Dr. Thomas Lengauer for his support and Ms. Noriko Teramoto for her help. This work was supported by grants from the Ministry of Education, Culture, Sports, Science, and Technology, and the Ministry of Health, Labour and Welfare, Japan.

REFERENCES

Akari, H., Mori, K., Terao, K., Otani, I., Fukasawa, M., Mukai, R., and Yoshikawa, Y. (1996). In vitro immortalization of old world monkey T lymphocytes with *Herpesvirus saimiri*: its susceptibility to infection with simian immunodeficiency viruses. *Virology* 218, 382–388.

Akari, H., Nam, K. H., Mori, K., Otani, I., Shibata, H., Adachi, A., Terao, K., and Yoshikawa, Y. (1999). Effects of SIVmac infection on peripheral blood CD4+CD8+ T lymphocytes in *Cynomolgus macaques*. *Clin. Immunol.* 91, 321–329.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10037–10041.

Castro, B. A., Barnett, S. W., Evans, L. A., Moreau, J., Odehouri, K., and Levy, J. A. (1990). Biologic heterogeneity of human immunodeficiency virus type 2 (HIV-2) strains. *Virology* 178, 527–534.

Castro, B. A., Nepomuceno, M., Lerche, N. W., Eichberg, J. W., and Levy, J. A. (1991). Persistent infection of baboons and rhesus monkeys with different strains of HIV-2. *Virology* 184, 219–226.

Dong, F., and Zhou, H. X. (2002). Electrostatic contributions to T4 lysozyme stability: solvent-exposed charges versus semi-buried salt bridges. *Biophys. J.* 83, 1341–1347.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* Chap. 2, Unit 2.9.

Fujita, M., Yoshida, A., Sakurai, A., Tatsuki, J., Ueno, F., Akari, H., and Adachi, A. (2003). Susceptibility of HVS-immortalized lymphocytic HSC-F cells to various strains and mutants of HIV/SIV. *Int. J. Mol. Med.* 11, 641–644.

Ganser-Pornillos, B. K., Chandrasekaran, V., Pornillos, O., Sodroski, J. G., Sundquist, W. I., and Yeager, M. (2011). Hexagonal assembly of a restricting TRIM5alpha protein. *Proc. Natl. Acad. Sci. U.S.A.* 108, 534–539.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 397, 436–441.

Gordon, J. C., Myers, J. B., Folta, T., Shoja, V., Heath, L. S., and Onufriev, A. (2005). H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* 33, W368–W371.

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.

Hildebrandt, A., Blossey, R., Rjasanow, S., Kohlbacher, O., and Lenhof, H. P. (2007). Electrostatic potentials of proteins in water: a structured continuum approach. *Bioinformatics* 23, e99–e103.

Himathongkham, S., and Luciw, P. A. (1996). Restriction of HIV-1 (subtype B) replication at the entry step in rhesus macaque cells. *Virology* 219, 485–488.

Kono, K., Song, H., Shingai, Y., Shioda, T., and Nakayama, E. E. (2008). Comparison of anti-viral activity of rhesus monkey and cynomolgus monkey TRIM5alphas against human immunodeficiency virus type 2 infection. *Virology* 373, 447–456.

Kono, K., Song, H., Yokoyama, M., Sato, H., Shioda, T., and Nakayama, E. E. (2010). Multiple sites in the N-terminal half of simian immunodeficiency virus capsid protein contribute to evasion from rhesus monkey TRIM5alpha-mediated restriction. *Retrovirology* 7, 72.

Kortemme, T., and Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14116–14121.

Locher, C. P., Blackbourn, D. J., Herndier, B. G., Reyes-Teran, G., Barnett, S. W., Murthy, K. K., and Levy, J. A. (1998). Transient virus infection and pathogenesis of a new HIV type 2 isolate, UC12, in baboons. *AIDS Res. Hum. Retroviruses* 14, 79–82.

Locher, C. P., Witt, S. A., Herndier, B. G., Abbey, N. W., Tenner-Racz, K., Racz, P., Kiviat, N. B., Murthy, K. K., Brasky, K., Leland, M., and Levy, J. A. (2003). Increased virus replication and virulence after serial passage of human immunodeficiency virus type 2 in baboons. *J. Virol.* 77, 77–83.

Nakayama, E. E., Miyoshi, H., Nagai, Y., and Shioda, T. (2005). A specific region of 37 amino acid residues in the SPRY (B30.2) domain of African green monkey TRIM5 alpha determines species-specific restriction of simian immunodeficiency virus SIVmac infection. *J. Virol.* 79, 8870–8877.

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., Riganelli, D., Zanaria, E., Messali, S., Cainarca, S., Guffanti, A., Minucci, S., Pelicci, P. G., and Ballabio, A. (2001). The tripartite motif family identifies cell compartments. *EMBO J.* 20, 2140–2151.

Sebastian, S., and Luban, J. (2005). TRIM5alpha selectively binds a restriction-sensitive retroviral capsid. *Retrovirology* 2, 40.

Shibata, R., Sakai, H., Kawamura, M., Tokunaga, K., and Adachi, A. (1995). Early replication block of human immunodeficiency virus type 1 in monkey cells. *J. Gen. Virol.* 76(Pt 11), 2723–2730.

Shrake, A., and Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79, 351–371.

Song, H., Nakayama, E. E., Yokoyama, M., Sato, H., Levy, J. A., and Shioda, T. (2007). A single amino acid of the human immunodeficiency virus type 2 capsid affects its replication in the presence of cynomolgus monkey and human TRIM5alphas. *J. Virol.* 81, 7280–7285.

Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P., and Sodroski, J. (2004). The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in old world monkeys. *Nature* 427, 848–853.

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., Diaz-Griffero, F., Anderson, D. J., Sundquist, W. I., and Sodroski, J. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5514–5519.

Tang, C., Ndassa, Y., and Summers, M. F. (2002). Structure of the N-terminal 283-residue fragment of the immature HIV-1 Gag polyprotein. *Nat. Struct. Biol.* 9, 537–543.

VandeWoude, S., and Apetrei, C. (2006). Going wild: lessons from naturally occurring T-lymphotropic lentiviruses. *Clin. Microbiol. Rev.* 19, 728–762.

Ylinen, L. M., Keckesova, Z., Wilson, S. J., Ranasinghe, S., and Towers, G. J. (2005). Differential restriction of human immunodeficiency virus type 2 and simian immunodeficiency virus SIVmac by TRIM5 alpha alleles. *J. Virol.* 79, 11580–11587.

Conflict of Interest Statement: 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.

Received: 19 April 2012; paper pending published: 18 May 2012; accepted: 21 May 2012; published online: 05 June 2012.

Citation: Bozek K, Nakayama EE, Kono K and Shioda T (2012) Electrostatic potential of human immunodeficiency virus type 2 and rhesus macaque simian immunodeficiency virus capsid proteins. Front. Microbio. ***3****:206. doi: 10.3389/fmicb.2012.00206*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 26 July 2012
doi: 10.3389/fmicb.2012.00264

Evolutionary analysis of functional divergence among chemokine receptors, decoy receptors, and viral receptors

Hiromi Daiyasu[1]*, Wataru Nemoto[2] and Hiroyuki Toh[3]

[1] *Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan*
[2] *Division of Life Science and Engineering, School of Science and Engineering, Tokyo Denki University, Saitama, Japan*
[3] *Computational Biology Research Center, Advanced Industrial Science and Technology, Tokyo, Japan*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Hironori Sato, National Institute of Infectious Diseases, Japan
Mikita Suyama, Kyushu University, Japan

****Correspondence:***
Hiromi Daiyasu, Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, 1-5, Yamadaoka, Suita, Osaka 565-0871, Japan.
e-mail: daiyasu@ist.osaka-u.ac.jp

Chemokine receptors (CKRs) function in the inflammatory response and in vertebrate homeostasis. Decoy and viral receptors are two types of CKR homologs with modified functions from those of the typical CKRs. The decoy receptors are able to bind ligands without signaling. On the other hand, the viral receptors show constitutive signaling without ligands. We examined the sites related to the functional difference. At first, the decoy and viral receptors were each classified into five groups, based on the molecular phylogenetic analysis. A multiple amino acid sequence alignment between each group and the CKRs was then constructed. The difference in the amino acid composition between the group and the CKRs was evaluated as the Kullback–Leibler (KL) information value at each alignment site. The KL information value is considered to reflect the difference in the functional constraints at the site. The sites with the top 5% of KL information values were selected and mapped on the structure of a CKR. The comparisons with decoy receptor groups revealed that the detected sites were biased on the intracellular side. In contrast, the sites detected from the comparisons with viral receptor groups were found on both the extracellular and intracellular sides. More sites were found in the ligand binding pocket in the analyses of the viral receptor groups, as compared to the decoy receptor groups. Some of the detected sites were located in the GPCR motifs. For example, the DRY motif of the decoy receptors was often degraded, although the motif of the viral receptors was basically conserved. The observations for the viral receptor groups suggested that the constraints in the pocket region are loose and that the sites on the intracellular side are different from those for the decoy receptors, which may be related to the constitutive signaling activity of the viral receptors.

Keywords: chemokine receptors, decoy receptors, viral receptors, GPCR, molecular evolution

INTRODUCTION

The members of the chemokine (CK) family play important roles in regulating cell migration against inflammation, immune surveillance, and oncogenesis in vertebrates (Zlotnik and Yoshie, 2000). The CKs are classified into four subfamilies: CC, CXC, CX3C, and XC, based on the cysteine positions in their motifs (Zlotnik and Yoshie, 2000). CKs exert their activities through binding to their corresponding receptors. Presently, more than 40 CKs and 18 chemokine receptors (CKRs) have been identified in the human genome: 10 CCRs, six CXCRs, one XCR, and one CX3CR (Nomiyama et al., 2011). The CKR homologs are widely distributed among the vertebrate genomes. For example, homologs have even been identified from sea lampreys, which are one of the most primitive vertebrates (Nomiyama et al., 2011). The amino acid sequence identities among the CKRs and the homologs range from 25 to 80%, and the CKRs constitute a subfamily in the class A G protein-coupled receptors (GPCRs). The CKRs have broad ligand specificities (Nomiyama et al., 2011), and each receptor is able to interact with several CKs, and *vice versa*. This binding promiscuity makes it difficult to develop drugs to pinpoint the specific function of each CKR. Among these receptors, only the structure of CXCR4 has been solved by X-ray crystallography (Wu et al., 2010). Like the other GPCRs, this structure is characterized by the seven transmembrane (TM) helices, although T4 lysozyme was inserted within the intracellular loop (ICL) 3 between the TM helices 5 and 6, to stabilize the crystal. The extracellular cavity of CXCR4 is reportedly larger and wide open, as compared to those of other GPCR structures (Wu et al., 2010).

In addition to the traditional CKRs, five non-signaling CKR homologs have been identified in vertebrate genomes: CCRL1 (also known as CCX-CKR), CCRL2, CCBP2 (D6), CXCR7, and DARC (Duffy antigen receptor; Graham, 2009; Leick et al., 2010; Naumann et al., 2010). They are called "decoy" or "silent" receptors, because they are able to bind to several CKs without ligand-induced signaling. Most of them are constitutively internalized with or without ligands, and only the receptors are recycled to the cell membrane. Their functions are considered to regulate inflammatory responses by controlling the volume of free extracellular CKs, through internalization and degradation (Bonecchi et al., 2010). Like the traditional CKRs, these decoy receptors show a

broad CK-binding spectrum. CCRL1 interacts with several homeostatic CC-type CKs (Comerford et al., 2006), whereas CCBP2 and DARC interact with inflammatory CKs (Graham, 2009). CXCR7 interacts with the dual-functional CXC-type CKs (Naumann et al., 2010) without activating G proteins (Thelen and Thelen, 2008). CCRL2 is known to be a multifunctional receptor (Yoshimura and Oppenheim, 2011). Like other decoy receptors, it regulates the amount of free CKs. At the same time, it functions as a receptor for an adipokine called chemerin, although the ligand binding does not induce signaling and the receptor is not internalized even after ligand engagement. DARC is the most distantly related to the CKRs among the five decoy receptors, and was originally identified as a malarial parasite receptor (Bonecchi et al., 2010). The receptor also binds to the CC- and CXC-type inflammatory CKs.

Chemokine receptors homologs have been detected in double stranded DNA viruses, such as herpesvirus and poxvirus. These viruses are considered to have gained these proteins by horizontal gene transfer during the course of evolution (Slinger et al., 2011). The viral receptors are constitutively active without ligands, although some of them can bind to CKs. We studied five groups of viral proteins as described below. E1 is derived only from equid herpesvirus 2 of γ-herpesvirinae, which interacts with CCL11 (Camarda et al., 1999). ORF74 is derived from several γ-herpesviruses, and interacts with a broad range of CXC-type CKs (Maussang et al., 2009). The β-herpesviruses also have several CKR homologs. Among them, UL33 is encoded by the genomes of various vertebrate viruses, although its ligands have not been identified (Gruijthuijsen et al., 2002). On the other hand, the US27, US28, and vGPCRs, which share high sequence similarity, have only been identified in the primate β-herpesviruses (Sahagun-Ruiz et al., 2004). US28 is characterized as a receptor for CC-type ligands (Maussang et al., 2009). Several poxviruses, such as capripox virus, deerpox virus, and yatapox virus, also encode CKR homologs in their genomes. The receptors of poxviruses not only share high amino acid sequence similarity to CCR8, but also the CCR8-like CK-binding profile; that is, high affinity to CCL1 (Najarro et al., 2006). These viral receptors are considered to contribute to the escape from and/or the perturbation of the host immune system, and are involved in inflammatory diseases and cancer (Slinger et al., 2011), although the mechanisms of these receptors in viral pathogenesis remain poorly understood.

The CKRs and their homologs have been classified into three functionally different types, from the viewpoints of ligand binding and signaling. The traditional CKRs bind their ligands, which induce signal transduction. The decoy receptors also bind ligands, although the process does not induce signal transduction. In contrast, the viral receptors exhibit signaling activity without ligand binding. The decoy receptors and the viral receptors are considered to have functionally differentiated after their divergence from the traditional CKRs, by gene duplication or horizontal gene transfer. Therefore, the functional differentiation of these three types is expected to have changed the functional constraints at the amino acid sites responsible for the ligand binding and/or signaling. If the sites involved in the functional differentiation can be identified, then the information about the sites would be helpful to understand the mechanisms for the signaling associated with ligand-induced conformational changes. Several different methods have been developed to detect the amino acid sites involved in the functional differentiation of homologous proteins from a multiple sequence alignment, and they are roughly classified into two types. One of them examines the difference in the evolutionary rate at each alignment site among the proteins with different functions (Gu, 1999; Simon et al., 2002), while the other compares the amino acid compositions at each alignment site among the proteins with different functions (Landgraf et al., 1999; Hannenhalli and Russell, 2000). We applied the latter method, the comparison of amino acid compositions, to identify the sites involved in the functional differentiation of CKR homologs. The difference in the amino acid composition at each alignment site between two groups (CKRs and decoy receptors, or CKRs and viral receptors) was calculated as the Kulback–Leibler (KL) information value (Hannenhalli and Russell, 2000; Ichihara et al., 2004). The sites with large KL information values were selected as the candidates for the functional differentiation. The amino acid residues corresponding to the selected sites were mapped on the tertiary structure of CXCR4. The comparison of the CKRs and decoy receptors revealed that the sites with large KL information values were concentrated on the cytosolic side of the CKR structure, with statistical significance. In contrast, there was no such bias in the distribution of the sites with large KL values between the CKRs and viral receptors. Based on the detected sites and the distribution of the corresponding residues on the tertiary structure, the underlying mechanisms for the functional divergence of the CKR homologs will be discussed.

MATERIALS AND METHODS

AMINO ACID SEQUENCE DATA

The amino acid sequences of the CKRs and their homologs, including decoy receptors and viral receptors, were collected by searching the non-redundant protein sequence database at the NCBI site[1] with BLAST version 2.2.25 (Altschul et al., 1997). The amino acid sequence of human CXCR4 (GI number of NCBI: 1705894) was used as the query for the BLAST search. The sequence similarity search was also performed against the Ensembl[2] and elephant shark genome project[3] genome databases. When several amino acid sequences were almost identical, one of them was selected as the representative. The sequences used in this study are shown in **Table 1**.

AMINO ACID SEQUENCE ALIGNMENT AND PHYLOGENETIC ANALYSIS

A multiple amino acid sequence alignment was produced with the alignment software MAFFT, version 6.857 (Katoh et al., 2002; Katoh and Toh, 2008). At first, 444 traditional CKRs were aligned. This result was manually refined, based on information about the secondary structures. Subsequently, 178 sequences consisting of decoy and viral receptors were added to the CKR alignment one by one, using the profile option of Clustal W (version 1.83; Thompson et al., 1994). Based on the alignment, an unrooted molecular phylogenetic tree was constructed by the neighbor-joining (NJ) method (Saitou and Nei, 1987). The genetic distance between

[1] http://www.ncbi.nlm.nih.gov/BLAST/
[2] http://www.ensembl.org/index.html
[3] http://esharkgenome.imcb.a-star.edu.sg/

Table 1 | The sequences of the CKRs, decoy receptors, and viral receptors.

CKRs									
CCR1									
416802	114586498	332215794	297206879	3023506	85718627	118150798	48675909	283837817	194221405
1705891	10120494	281343586	209863082	84370370	126341640				
CCR2 AND CCR5									
1705896	116243032	2851566	2494974	110278904	213391512	48675899	148234591	301754037	75073875
75073881	75072034	75073877	33521616	9502108	33521612	5712983	75073171	5713007	75075056
75074166	75073886	38605083	75074950	3023510	48428812	48427940	75073874	75069418	75074956
3913250	75073879	6831507	75073880	6831506	75069417	6831508	6831510	75073878	6831511
38604970	3023504	6831509	3023503	38604969	75073883	75073884	75073876	75073882	75074954
75074955	75074952	5713069	5713068	75070083	154813802	13431410	114586511	297712573	332266801
296225031	1168965	291393559	301754035	57101676	147901663	303227941	48675907	10719941	2506483
145226674	145423899	126341644	126341394	149632073	149632071	154813804	326922093	224045497	113951665
224045499	327282149	327282151	327282147	148238158					
CCR3									
1705892	149632069	126341642	1705893	6831505	281343587	55976357	209863084	48675903	303227943
57163985	149728986	205830369	296225029	3023507	62510458	3023509	332215796	297671507	55620263
6831512	149632067								
CCR4									
1705894	297671782	332215473	109052678	296228310	194221518	62899791	301767336	154152187	1705895
26449155	225571128	291399774	290649642	126341582	149455250	327282179	326922159	118086158	224045511
CCR6									
2851567	332825448	332245368	297679621	74136427	296483830	301766648	73945797	194227505	291415344
8134362	61557091	126311276	149637480	166159172	326915616	224047748	327262258	301612736	AAVX01068499.1
153791315	213512406								
CCR7									
1352335	1352336	41054914	296202786	332847660	297701272	332258459	187475071	197210544	75070300
48374059	149724475	301779133	73965967	291406000	126308140	224086466	326934127	311771569	327275717
148222097	301626915	148922928	301616384	AAVX01326265.1	AAVX01024218.1				
CCR8									
1707884	71896604	326922147	296399392	224045515	3334152	27721715	10719948	114586090	297671676
332215595	296228403	149728750	57103782	301785880	303227947	291393287	126341586	224045501	124249288
AAVX01061874.1									
CCR9									
114152781	301623067	169145191	209155804	159155092	113951675	149632061	27229230	8134364	109041099
297671522	114586481	296225018	48675913	194221411	115311322	148356263	291393553	73985992	301754023
126341634	224045505	ENSACAP00000019440	AAVX01140752.1						
CCR10									
62298314	156104886	297701070	109115520	332260917	296201470	291406157	94536880	303227949	113205696
194216876	73965655	281344547	157819219	12643802	126307934	327275281			
CXCR1 AND CXCR2									
108936015	2494963	2494962	110825972	110825970	110825971	124357	157063152	2494966	194211303
194043812	149711459	57111007	6685568	301755776	301755774	296205556	23305862	297264881	1352454
125987816	2494967	2494968	547719	290542297	1352455	547718	126337864	126337862	290650152
2494965	81913011	326922912	78482916	71896165	327260352	327260354	148223850	149531934	292617830
3298340	185134540	47220980	118344614	189523763	3298358	47220226	AAVX01477245.1		

(Continued)

Table 1 | Continued

CXCR3									
2829400	222537776	332265855	297710303	149758513	311276475	75072906	75070299	281337759	291407679
75070286	76363509	76364160	185133155	213513980	47218519	169154030	58272233	58272235	301618339
49118568	327289267	169154032	185133520						
CXCR4									
46577576	3913205	3023448	114152796	75074809	3023451	75073173	46577575	75072692	128999
75072471	197253269	3023449	301784615	149730555	114149257	2494971	2494970	149637056	327260636
224056102	45382915	126326273	82241554	123884047	82249002	6318165	17223091	319099413	63102334
3551197	47215024	185133162	27802639						
CXCR5									
416718	311264026	291412974	73955058	301788472	291173052	416719	461630	297690401	332837885
332208422	126326932	326933405	71894759	301606664	326676225	ENSACAP00000016849	AAVX01026304.1		
CXCR6									
3121816	296225022	81917290	149018110	163915588	301607738	48675917	71153257	3121823	38503255
291173054	73985805	301754025	3121822	38503164	10719922	ENSMODP00000032629	ENSACAP00000013878	ENSOANP00000010838	
CX3CR1									
1351394	297671678	332215591	226342927	109041508	296228401	281352825	73990285	122136266	149729043
8134357	548703	238055160	126341584	149495131	224045513	296399391	326922149	50732904	327282177
XCR1									
1170008	114586489	332215787	297671509	109041073	296225024	194221407	303227953	48675911	73985988
301754029	291393557	12585214	157822209	126341638	149632065	326922097	113951667	224045503	327282173
292629502	66911140	326679306	301607740	291190313	225706150	47215603	AAVX01263959.1		
Decoy receptors									
CCRL1									
14285406	55621142	297671993	109049361	296228075	194221598	147898731	301781760	73990094	544463
291399807	68565247	109483837	301616697	148228890	317419986	292627507	47208340	148725584	ENSMODP00000032599
ENSOANP00000010818	ENSGALP00000019109	ENSACAP00000000733							
CCRL2									
114586515	108885280	297671505	75075026	296225035	48675905	115496362	194221403	73985969	281343590
108885281	157824077	291393561	126341646	149632075					
CCBP2									
20455469	114586376	297671602	296224947	291393242	149729016	57103810	301780460	194040849	62752046
14547935	14547939	126335978	224046888	50732143					
CXCR7									
115502380	55619711	297669797	109101586	296205948	149711234	132206	301789855	47117863	10720245
311273312	148356261	291414068	126314602	149633404	134085621	224054077	327260743	71896089	148223972
158254308	47226985	221307557	AAVX01259911.1						
DARC									
67476970	27734275	297663060	27734274	296229319	291397675	293341477	27734283	149755929	160332326
311254049	74006341	301783805	126307326	327287460					
Viral receptors									
E1									
124738385	124738389	124738361	9628003	124738365	124738377	124738369	124738381	124738383	124738379
124738375	124738373	124738371	124738391	124738393	124738399	124738395	124738401		

(Continued)

Table 1 | Continued

ORF74									
139472805	4154096	46519489	18653888	9626030	30348580	9631265	9629596	262285115	321496625
124738284	124738278	9628076	124738274						
UL33									
10998155	1717998	1717999	222354475	52139219	254770904	284518933	254771070	290564391	242345651
59803016	20026636	28412128	14251021	124248174	190886816	14916726	9845324	213159183	
βHV									
59800434	20026758	229270263	51556673	137159	20026757	229270249	51556669	229270262	51556671
229270250	33694234	229270261	23194507	229270231	51556668				
pox									
38229303	12085128	157939769	38229175	146746499	12084990	62637392	211956287	586239	211956433
62637539	226437997	226438057	2495049	226438017	13876663	226437981	226437989	226437993	

The sequences obtained from the NCBI database are indicated by the GI numbers. The sequences with names starting with AAVX were obtained from the elephant shark genome project database, while the names stating with ENS indicate sequences obtained from the Ensembl database.

every pair of aligned sequences was calculated as a maximum likelihood estimate (Felsenstein, 1996), under the JTT model (Jones et al., 1992) for the amino acid substitutions. The sites including gaps in the alignment were excluded from the calculation. The statistical significance of the NJ tree topology was evaluated by a bootstrap analysis (Felsenstein, 1985) with 1,000 iterative tree reconstructions. Two software packages, PHYLIP 3.5c (Felsenstein, 1993) and MOLPHY 2.3b3 (Adachi and Hasegawa, 1996), were used for the phylogenetic analyses. A cluster of decoy or viral receptors with a bootstrap probability greater than 80% was adopted as a group of receptors with different functions from the traditional CKRs.

CALCULATION OF THE KULLBACK–LEIBLER INFORMATION VALUE

The multiple alignment thus obtained was reconstructed into 10 alignments, each consisting of two groups, the traditional CKRs and one of the decoy receptor groups or viral receptor groups. We then calculated the amino acid compositions of the two groups at each alignment site, according to the multiple alignment. We used the method adopted in PSI-BLAST (Altschul et al., 1997) to estimate the site-specific amino acid composition. The weighting method of Henikoff and Henikoff (1994) was used for the residue count. The weight for the pseudocount β, was set to 0.1. For the calculation of the pseudocount, λ_u, a parameter for ungapped BLAST, was calculated for each alignment by the Newton–Raphson method (Ewens and Grant, 2001). When more than half of the sequences had gaps at an alignment site, the calculation of the site-specific amino acid composition and the following investigation were skipped. Next, the difference in the amino acid composition between two groups at each alignment site was calculated as the KL information value. The KL information value is defined as follows:

$$\sum_{i=1}^{20} p(i) \log \frac{p(i)}{q(i)} \tag{1}$$

where p and q are the site-specific amino acid residue compositions for the two groups, which are estimated by the method described above. The parameter i indicates that the summation is obtained over 20 amino acid residues. KL information does not satisfy one of the distance axioms, symmetry. To satisfy this condition, the KL information was modified as follows:

$$\sum_{i=1}^{20} p(i) \log \frac{p(i)}{q(i)} + \sum_{i=1}^{20} q(i) \log \frac{q(i)}{p(i)} \tag{2}$$

Formula 2 was used to predict the sites subjected to different functional constraints between the two groups. In this study, the functional constraint at a site of a protein sequence is defined as the extent of intolerance to mutation at the site, due to a reduction of the protein function by the mutation. This is a special case of the cumulative relative entropy developed by Hannenhalli and Russell (2000), which is applicable to an alignment consisting of multiple groups. When the KL information value of an alignment site was located in the top 5% of the distribution of KL information values for all of the sites, the site was regarded as a site under different constraints between the groups (Ichihara et al., 2004).

Among them, the sites that fell in the gap region of CXCR4 in the alignment were neglected, because the subsequent analyses were performed based on mapping onto the CXCR4 structure.

STATISTICAL EVALUATION FOR BIAS IN THE SPATIAL DISTRIBUTION OF THE SITES UNDER DIFFERENT CONSTRAINTS

We examined the statistical significance for the bias in the positions of the selected sites by the KL information values on the reference CXCR4 structure (PDB ID: 3ODU), using the following procedure. At first, we calculated the geometric center of the three extracellular loops (ECLs) and the N-terminal region, and that of the three ICLs. The coordinates of the Cα atoms were used for the calculation. The C-terminal region (residues 303–328) was not used in the calculation of the geometric center of the intracellular side, since this region extended into the cytosolic region. The chimeric lysozyme region was also neglected from the calculation. A unit vector on the axis connecting the two geometric centers, which originated from the midpoint between the geometric centers toward the geometric center of the extracellular side, was calculated. The inner product between the unit vector and a vector from the midpoint to the Cα atom of every residue, except for those of the chimeric lysozyme region, was then calculated. The inner product score indicated the projected position of the residue on the axis (see **Figure 1**). The positive or negative score corresponded to the extracellular or cytosolic location of a residue, respectively, relative to the geometric center. The distribution of the inner product scores for the residues selected by the KL information values was compared with those of the remaining residues by the two-sided *t*-test. The null hypothesis was the same for all of the tests: the average of the residues corresponding to the sites with large KL information values is the same as that of the remaining residues. For the statistical test, the function in the statistical computing software R, "*t*-test," was used for the evaluation.

RESIDUE INDICATION

The sites of each group selected by the KL information values are indicated on the corresponding sites of CXCR4 in this study. When the site has the number based on Ballesteros–Weinstein nomenclature (Ballesteros and Weinstein, 1995), the figure is also shown in the superscript. In this notation, the first digit indicates the number of the TM helix, and the following digit is the position counted from the most conserved site in each TM, to which the number 50 is assigned.

RESULTS

THE PHYLOGENETIC ANALYSIS

The multiple alignment of 622 sequences were constructed, which is downloadable from the URL: http://seala.cbrc.jp/~toh/suppl.html. The alignment of the representative sequences is shown in **Figure 2**. Based on the alignment, the phylogenetic tree of the CKRs and the decoy and viral receptors was constructed (**Figure 3**). Several clusters with high bootstrap probability (>80%) were identified in the tree, which included five decoy receptor groups and five viral receptor groups. The decoy receptor groups are referred to as CCRL1, CCRL2, CCBP2, CXCR7, and DARC, according to the constituent receptors. The numbers in each group were 23, 15, 15, 24, and 15, respectively. On the other hand, the viral receptor groups are referred to as E1, ORF74, UL33, βHV, and pox. The first three groups were named according to the constituent receptors. The βHV group consists US28, US27, and vGPCRs. Pox is a group of receptors derived from poxviruses. The numbers in each viral group were 18, 14, 19, 16, and 19. The evolutionary relationships between the CKRs and the decoy and viral receptors shown in the figure were roughly similar to those reported previously (Rosenkilde et al., 2001; Zlotnik et al., 2006). Murphy et al. (2000) suggest that the evolutionary rates of the CKRs are faster than those of the other GPCRs, because of the immune functions of CKRs. The long branch lengths suggested that the evolutionary rates of the receptors belonging to CCRL2, DARC, ORF74, UL33, βHV, and pox may be higher than those of the traditional CKRs, although we refrained from further examination of evolutionary rate accelerations in this study. In the subsequent analyses, each group of the decoy and viral receptors thus obtained was compared with the group of the traditional CKRs.

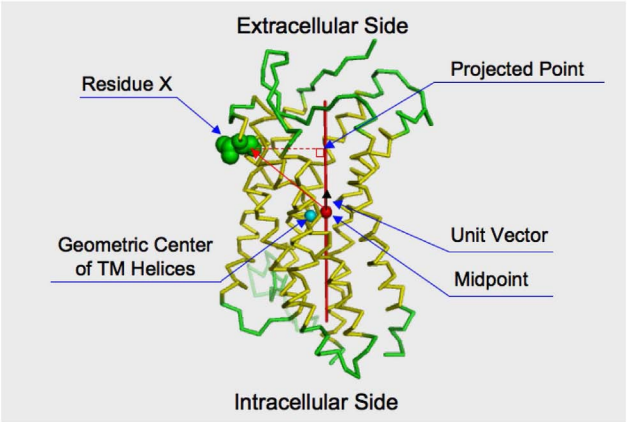


FIGURE 1 | Projection of a residue on the axis connecting the intracellular and extracellular sides of the receptor. The structure of CXCR4 is shown by the ribbon model. The membrane spanning helices indicated by the structural element page for CXCR4 in GPCRDB (http://www.gpcr.org/7tm/) are colored yellow. The sphere colored cyan indicates the geometric center of the alpha carbons of the membrane spanning helices. The red axis connects the geometric center of extracellular loops and the N-terminal loop and that of the intracellular loops. The midpoint of the axis is indicated by a filled sphere colored red. The distance between the cyan and red spheres is close (3.26 Å). That is, the midpoint is considered to roughly reflect the geometric center of the transmembrane helices. How to take the orthogonal projection of an amino acid residue to the axis is shown by using Residue X. Consider a vector from the midpoint to the Cα atom of the residue. By taking an inner product between the vector and a unit vector, which runs along the axis and is originated from the midpoint. The projected point is obtained by taking the inner product.

DETECTION OF SITES WITH LARGE KL INFORMATION VALUES

The differences in the amino acid composition at each alignment site were examined between the traditional CKRs and each group of decoy and viral receptors. The sites with large KL information values in the top 5% are summarized in **Table 2**. The residues corresponding to such sites were mapped on the structure of CXCR4 (**Figure 4**). As shown in **Table 2**, about 11 ~ 14

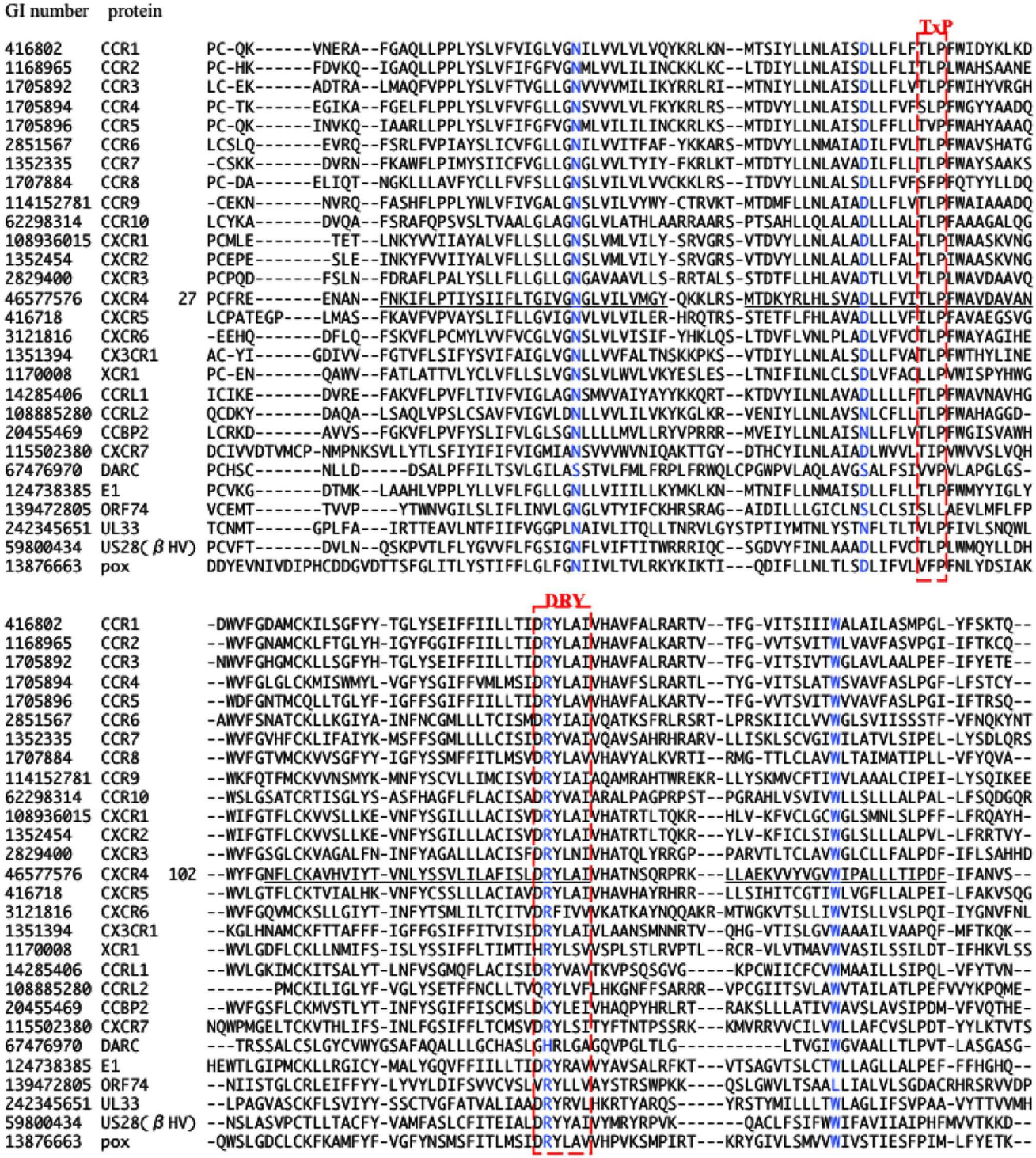


FIGURE 2 | Multiple amino acid sequence alignment of CKRs, decoy, and viral receptors. The GI number and protein name of the representative protein from each group are shown at the left side of the aligned amino acid sequence. In the case of CXCR4, the residue number of the first residue of the aligned sequence is shown after the protein name, and the TM regions described in the GPCRDB (http://www.gpcr.org/7tm/) are indicated by underlines. The corresponding sites for x.50 of Ballesteros–Weinstein nomenclature are colored blue. Four motifs, TxP, DRY, CWxP, and NPxx$Y_{5.6}$F, are enclosed by red line.

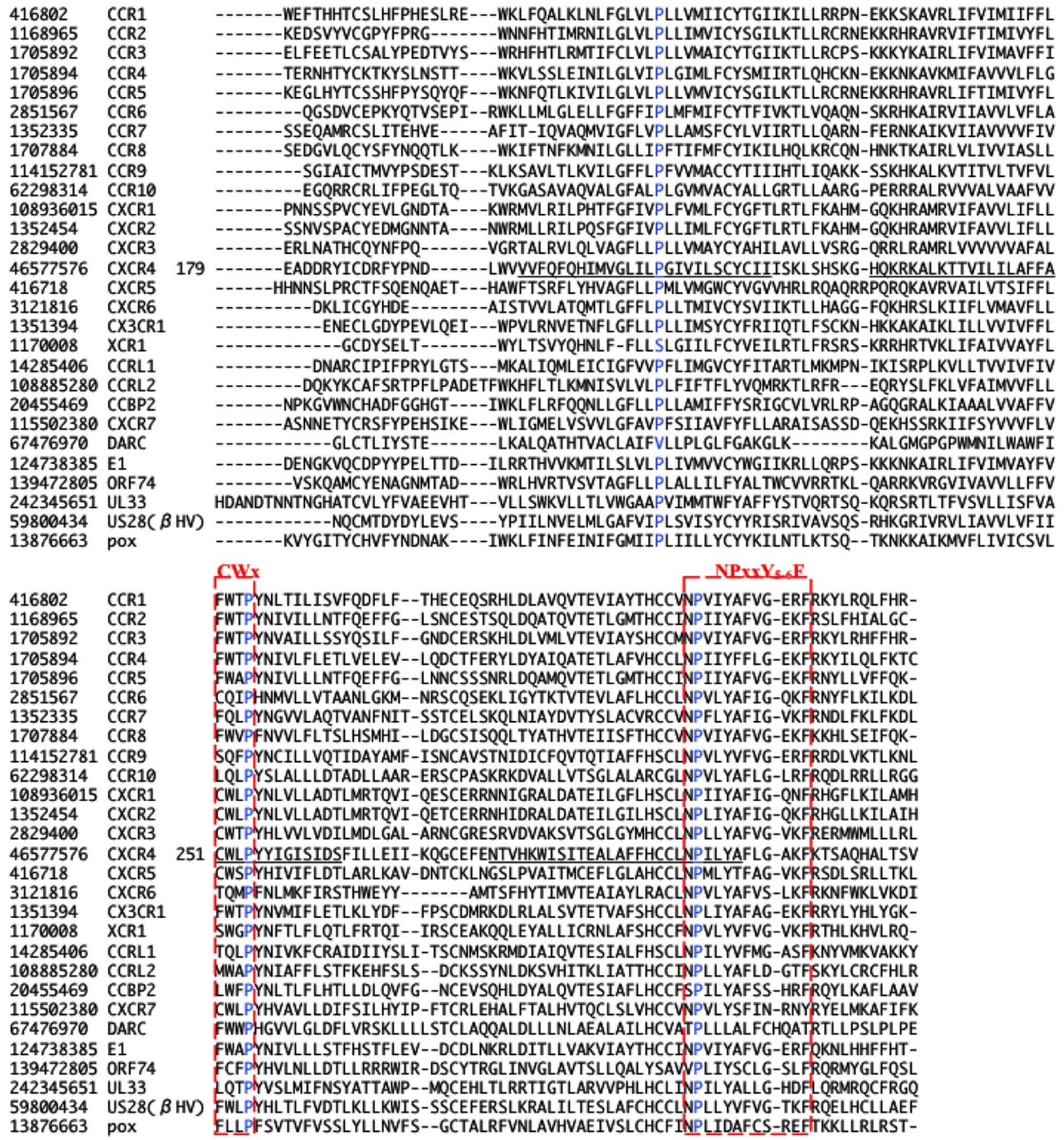


FIGURE 2 | Continued.

sites were selected from each group with the comparison of CKRs, and they included the sites in the sequences for GPCRs or the CKR-specific motif. Several sites that have been experimentally identified to be important for ligand binding or signaling were also selected. In addition, many uncharacterized sites were detected.

The DRY (Asp-Arg-Tyr) motif of the GPCRs is conserved as the sequence DRYLAIV in the traditional CKRs, from the end of TM3 to ICL2 (Graham, 2009). The motif is related to signal transduction, through interactions with G proteins. The conserved R134$^{3.50}$ is involved in the interchanges between the inactive and active conformations of GPCRs. In the inactive conformation, this Arg

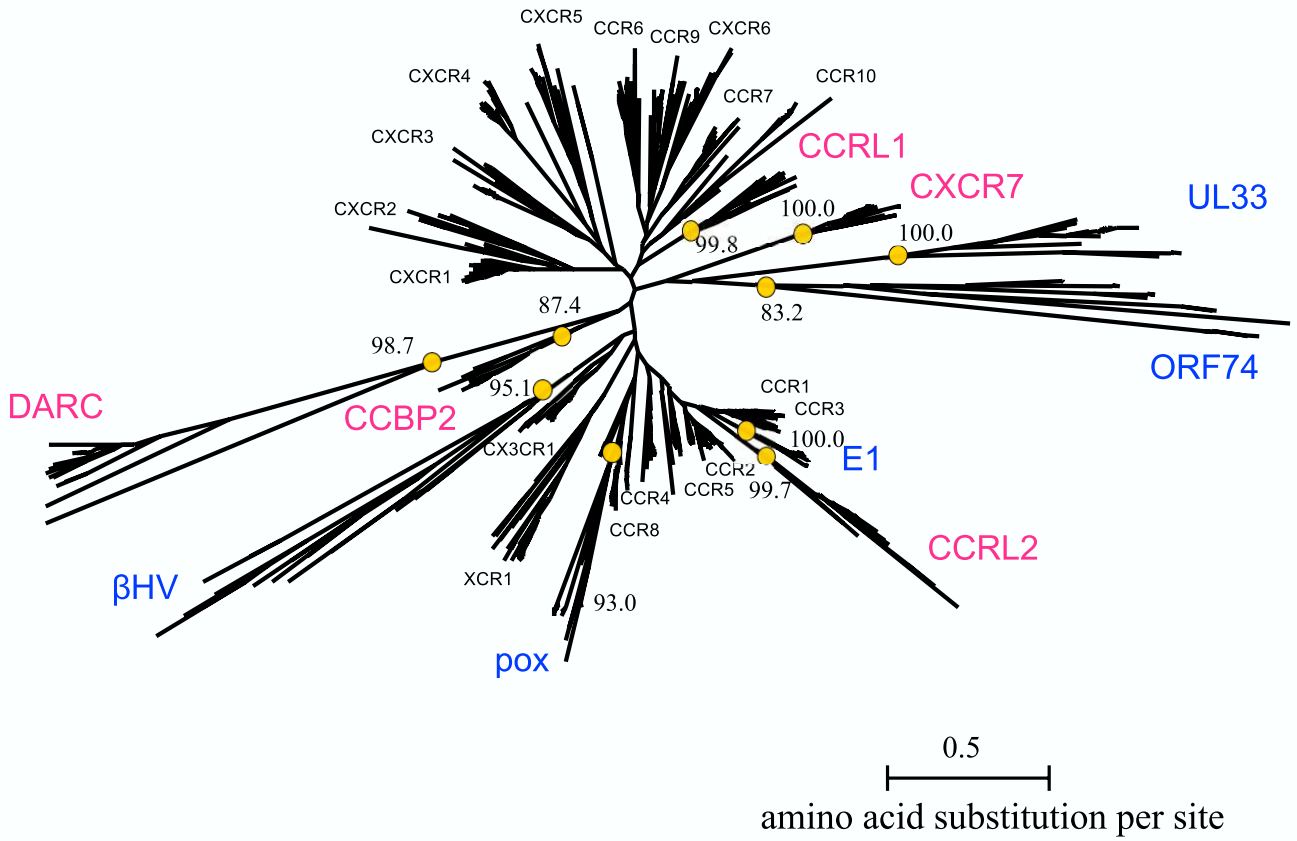


FIGURE 3 | Phylogenetic tree of CKRs and their homologs. The tree for the 622 sequences listed in **Table 1** is shown. The names of the CKRs (black), the decoy receptor groups (magenta), and the viral receptor groups (blue) are indicated near the receptor clusters. The bootstrap probabilities of the decoy and viral receptor groups are shown at the nodes corresponding to the common ancestors of the groups, which are indicated by circles.

interacts with its neighboring residue, D133$^{3.49}$, but in the active conformation, the residue interacts with Y219$^{5.58}$ (Holst et al., 2010). The sites in the DRY region of the DRYLAIV sequence were only detected from the analyses with the decoy receptor groups, CCRL2, CCBP2, and DARC. In addition, Y219$^{5.58}$ was also detected from the analysis with the DARC group. On the other hand, the sites in the LAIV$^{3.52\sim3.55}$ region of the DRYLAIV sequence were detected from the examinations with the decoy and viral receptor groups, CCBP2, UL33, and βHV. The CWxP motif is located in the middle of TM6. This W252$^{6.48}$ is believed to function as a micro-switch in the receptor activation mechanism, and P254$^{6.50}$ creates a kink in this helix, around which TM7 performs its rigid body movements during activation (Nygaard et al., 2009). The corresponding sites of this motif were detected from the analyses of two decoy receptor groups, CCRL1 and CCRL2, but not from those of any viral receptor group. The fifth site of the NPxxY$_{5\text{-}6}$F motif in TM7, Y302$^{7.53}$ functions in the interchange of an inactive rotamer conformation (Nygaard et al., 2009). The sites of this motif were detected from the investigations with every decoy receptor group and two viral receptor groups, ORF74 and UL33. The TxP motif of TM2 is known as a specific motif of the traditional CKRs. It is known that the TxP motif in TM2 is specific for traditional CKRs (Govaerts et al., 2001). The third site of the TxP motif, P92$^{2.58}$, bends the helix, which determines the intra-helical location that is involved in the receptor activation. The sites of the motif were detected from the analyses of two viral receptor groups, the ORF74 and pox groups, but not from the assessment with any decoy receptor group. In addition, several sites corresponding to highly conserved positions in GPCRs, which are denoted as x.50 by the Ballesteros–Weinstein nomenclature, such as N56$^{1.50}$ and D84$^{2.50}$, were detected from analyses of several groups (see **Table 2**). **Table 2** also shows the other important residues experimentally identified as having binding or signaling functions.

We examined which sites were commonly selected from the comparisons. No site was shared in all of the comparisons. Furthermore, there was no site commonly detected from the analyses with all of the decoy receptor groups or all of the viral receptor groups. However, several sites were detected from the different comparisons. For example, the sites corresponding to D74$^{2.40}$, D84$^{2.50}$, R134$^{2.50}$, A141$^{3.57}$, T142$^{3.58}$, S144$^{3.60}$, C218$^{5.57}$, K230$^{6.26}$, T241$^{6.37}$, C251$^{6.47}$, G306$^{8.47}$, and K308$^{8.49}$ were detected from at least two assessments with decoy receptor groups. Most of these sites are located in ICL2, 3, and the C-terminal region. Among these sites, D84$^{2.50}$, A141$^{3.57}$, C218$^{5.57}$, and K308$^{8.49}$ were also detected from at least one analysis with the viral receptor group. W94$^{2.60}$, W102 (ECL1), L136$^{3.52}$, H140$^{3.56}$, G207$^{5.46}$, L208$^{5.47}$, and K308$^{8.49}$ were detected from at least two analyses of the viral receptor groups.

Table 2 | Selected sites with large KL information values.

Residue (CXCR4)	Position	Region	B and W	Remarks	KL value	Frequency (%)				Reference
A. (Decoy receptors)						**CKRs group**				
CCRL1										
H203	Extra	TM5	5.42	Pocket	7.03					(L) Scholten et al. (2012)
						E	93.3			
C251	Extra	TM6	6.47	CWxP	6.71	C	39.7	F	31.7	(S) Nygaard et al. (2009)
						T	93.5			
A307	Intra	C	8.48	$NPxxY_{5\text{-}6}F$	7.29	E	33.8	V	32.0	
						A	49.5			
K308	Intra	C	8.49	$NPxxY_{5\text{-}6}F$	8.07	K	73.0			(L) Scholten et al. (2012)
						S	86.7			
Y121	Intra	TM3	3.37	Pocket	8.50	Y	67.0			
						V	40.3	S	34.3	
L125	Intra	TM3	3.41		7.19	L	55.6	F	38.0	
						Q	54.7			
T142	Intra	ICL2	3.58		6.67	T	33.2	V	32.2	
						P	59.7			
S144	Intra	ICL2	3.60		6.80	A	47.9			
						Q	36.5			
K230	Intra	ICL3	6.26		7.74	R	35.4			
						N	50.6			
G231	Intra	ICL3	6.27		8.26	N	39.9			
						I	43.4	W	31.8	
R235	Intra	ICL3	6.31		7.29	H	39.0			
						S	54.7			
T241*	Intra	ICL3	6.37		8.78	I	67.5			
						L	92.8	W	30.3	
I261	Extra	TM6	6.57		8.35	L	47.1			
						C	48.8			
L317	Intra	C	8.58		7.44	L	44.9			
						A	89.6			
CCRL2										
T73*	Intra	ICL1	2.39		9.69	T	91.9			(L) Scholten et al. (2012)
						E	38.7	G	34.7	
D84	Intra	TM2	2.50		8.20	D	95.4			(S) Rosenkilde et al. (2008)
						N	92.9			(S) Nygaard et al. (2010)
D133	Intra	ICL2	3.49	DRY	8.08	D	88.1			(L) Scholten et al. (2012)
						Q	84.4			
Y190	Extra	ECL2	–	Pocket	9.14	Y	47.5			(S) Zhou et al. (2001)
						R	47.1	K	33.4	
A237*	Intra	ICL3	6.33		9.72	A	79.1			(L) Scholten et al. (2012)
						L	83.5			
C251	Extra	TM6	6.47	CWxP	9.27	C	39.9	F	31.5	(S) Nygaard et al. (2009)
						M	72.8			
F292	Extra	TM7	7.43		10.56	F	46.3			(L) Scholten et al. (2012)
						T	61.6			(L) Choi et al. (2005)
G306	Intra	C	8.47	$NPxxY_{5\text{-}6}F$	11.94	G	96.0			
						D	93.1			
K308	Intra	C	8.49	$NPxxY_{5\text{-}6}F$	8.96	K	73.8			(L) Scholten et al. (2012)
						T	36.3			
E31	Extra	N	–	Pocket	9.72					
						Y	78.0			

(Continued)

Table 2 | Continued

Residue (CXCR4)	Position	Region	B and W	Remarks	KL value	Frequency (%)				Reference
							CKRs group			
A141	Intra	ICL2	3.57		8.63	A	82.2			
I215	Intra	TM5	5.54		8.31	M	72.6			
						F	69.4			
I223	Intra	ICL3	5.62		8.16	I	40.0	L	31.8	
						R	66.8			
CCBP2										
R134	Intra	ICL2	3.50	DRY	10.75	R	99.0			(S) Deupi and Standfuss (2011)
						K	65.9			(S) Holst et al. (2010)
A137	Intra	ICL2	3.53	DRY	9.22	A	77.1			
						E	53.4			
G306	Intra	C	8.47	NPxxY$_{5\text{-}6}$F	9.35	G	96.2			
						S	53.7			
V59	Intra	TM1	1.53		7.61	V	94.3			
						L	81.3			
T142	Intra	ICL2	3.58		10.12	T	33.2	V	31.9	
						Q	57.0			
S144	Intra	ICL2	3.60		7.54	A	48.0			
						H	33.5			
A152	Intra	ICL2	4.41		7.65					
						K	60.6			
Y157	Intra	TM4	4.46		7.47	C	49.1	S	36.1	
S224	Intra	ICL3	5.63		7.68					
						C	80.7			
K230	Intra	ICL3	6.26		7.89	R	36.1	Q	34.4	
						L	67.8			
Q233	Intra	ICL3	6.29		11.12	K	36.1			
						G	91.9			
CXCR7										
D74	Intra	ICL1	2.40		11.53	D	72.0			(L) Scholten et al. (2012)
						H	90.8			
F87	Extra	TM2	2.53		10.25	F	75.4			(L) Tian et al. (2005)
						V	84.5			
G306	Intra	C	8.47	NPxxY$_{5\text{-}6}$F	11.52	G	96.2			
						N	66.8			
K38	Extra	N	1.32	Pocket	10.59					
						Y	86.7			
G55	Intra	TM1	1.49		11.17	G	97.8			
						A	91.1			
M63	Intra	TM1	1.57		10.87	L	49.7			
						N	90.6			
M72*	Intra	ICL1	2.38		10.57	M	41.0			
						E	54.2	D	37.3	
L86	Extra	TM2	2.52		9.89	L	79.1			
						C	55.5	W	38.4	
A141	Intra	ICL2	3.57		11.00	A	81.9			
						F	67.4			
C218*	Intra	ICL3	5.57		10.28	C	94.0	R	41.7	
						F	79.9			

(Continued)

Table 2 | Continued

Residue (CXCR4)	Position	Region	B and W	Remarks	KL value	Frequency (%)				Reference
							CKRs group			
K236	Intra	ICL3	6.32		10.01	K	56.4			
						S	54.7			
L238*	Intra	ICL3	6.34		10.15	V	32.9	I	31.4	
						R	60.5			
L244	Intra	TM6	6.40		11.42	V	54.0			
						Y	92.6			
K271	Extra	ECL3	–		10.02					
						F	79.8			
DARC										
N56	Intra	TM1	1.50		11.51	N	98.8			(S) Rosenkilde et al. (2008)
						S	78.2			(S) Nygaard et al. (2010)
D74	Intra	ICL1	2.40		11.91	D	72.3			(L) Scholten et al. (2012)
						R	44.9	W	30.1	
D84	Intra	TM2	2.50		11.61	D	95.5			(S) Rosenkilde et al. (2008)
						S	73.8			(S) Nygaard et al. (2010)
Y116	Extra	TM3	3.32	Pocket	11.54	Y	57.9			(L) Scholten et al. (2012)
						W	63.8			(L) Surgand et al. (2006)
R134	Intra	ICL2	3.50	DRY	12.36	R	98.9			(S) Deupi and Standfuss (2011)
						G	61.7			(S) Holst et al. (2010)
Y135	Intra	ICL2	3.51	DRY	13.42	Y	92.6			
						P	83.1			
Y219*	Intra	ICL3	5.58		12.72	Y	96.2			(S) Holst et al. (2010)
	G	52.1								
Y302	Intra	TM7	7.53	$NPxxY_{5\text{-}6}F$	11.86	Y	95.7			(L) Scholten et al. (2012)
						L	75.4			(S) Rosenkilde et al. (2008)
F309	Intra	C	8.50	$NPxxY_{5\text{-}6}F$	12.49	F	98.7			(S) Rosenkilde et al. (2008)
						A	54.7			
V214	Intra	TM5	5.53		11.34	V	52.4			
						P	93.5			
C218*	Intra	ICL3	5.57		12.38	C	94.1			
						L	63.6			
T241*	Intra	ICL3	6.37		13.28	I	67.3			
						W	75.7			
L246	Intra	TM6	6.42		12.77					
						W	96.0			
C296	Intra	TM7	7.47		11.56	C	88.7			
						V	64.4			
B. (Viral receptors)										
E1										
Y116	Extra	TM3	3.32	Pocket	7.62	Y	56.1			(L) Scholten et al. (2012)
						C	88.6			(L) Surgand et al. (2006)
Q66	Intra	ICL1	1.60		6.34					
						M	83.5			
A95	Extra	TM2	2.61		8.18	A	66.0			
						M	63.8			
V99	Extra	TM2	2.65		6.66	A	34.8			
						G	84.8			
N106	Extra	ECL1	3.22		6.48					
						I	71.1			

(Continued)

Table 2 | Continued

Residue (CXCR4)	Position	Region	B and W	Remarks	KL value	Frequency (%)				Reference
						CKRs group				
S123	Intra	TM3	3.39	Pocket	8.02	G	49.6	S	36.4	
						Q	54.8			
G207	Extra	TM5	5.46	Pocket	7.77	G	90.7			
						S	84.3			
C220	Intra	ICL3	5.59		6.62					
						Y	47.3	W	41.2	
G231	Intra	ICL3	6.27		9.44	N	39.6			
						P	84.4			
ORF74										
D84	Intra	TM2	2.50		11.25	D	95.6			(S) Rosenkilde et al. (2008)
						S	63.3			(S) Nygaard et al. (2010)
P92	Extra	TM2	2.58	TxP	11.63	P	98.4			(L) Govaerts et al. (2001)
						L	59.7			(S) Wu et al. (2010)
W94	Extra	TM2	2.60	Pocket	9.67	W	74.7			(L) Scholten et al. (2012)
										(S) Rosenkilde et al. (2010)
V112	Extra	TM3	3.28	Pocket	9.31	V	37.3			(L) Scholten et al. (2012)
						E	38.4			
W161	Intra	TM4	4.50			W	99.3			(C) Ballesteros and Weinstein (1995)
						F	32.9			
N298	Intra	TM7	7.49	NPxxY$_{5\text{-}6}$F	11.57	N	94.9			(S) Rosenkilde et al. (2008)
						V	38.2			(S) Nygaard et al. (2010)
F304	Intra	C	8.45	NPxxY$_{5\text{-}6}$F	9.17	F	94.6			
						L	57.4			
A307	Intra	C	8.48	NPxxY$_{5\text{-}6}$F	10.63	E	33.4	V	32.2	
						S	94.5			
K308	Intra	C	8.49	NPxxY$_{5\text{-}6}$F		K	73.8			(L) Scholten et al. (2012)
						L	48.0			
Y76 *	Intra	ICL1	2.42		8.84	Y	60.6	F	32.7	
						L	81.9			
A83	Intra	m	2.49		10.43	A	55.3	S	40.1	
						N	62.9			
H140 *	Intra	ICL2	3.56		9.90	H	49.5			
						F	50.0			
A237 *	Intra	ICL3	6.33		9.31	A	79.4			
						V	65.4	I	30.2	
UL33										
L120	Extra	TM3	3.36	Pocket	13.59	F	70.9			(L) Surgand et al. (2006)
						C	96.4			
L136	Intra	ICL2	3.52	DRY	11.63	L	66.3			
						R	74.7			
V139	Intra	ICL2	3.55	DRY	12.66	V	90.5			
						H	74.4			
L208	Extra	TM5	5.47	Pocket	13.48	F	70.1			(S) Holst et al. (2010)
						G	95.2			
A291	Extra	TM7	7.42		11.64	A	60.0	G	31.0	(L) Scholten et al. (2012)
						P	95.1			
K308	Intra	C	8.49	NPxxY$_{5\text{-}6}$F	11.86	K	73.3			(L) Scholten et al. (2012)
						D	74.3			
G55	Intra	TM1	1.49		12.83	G	97.9			
						L	46.4	M	37.0	

(Continued)

Table 2 | Continued

Residue (CXCR4)	Position	Region	B and W	Remarks	KL value	Frequency (%)				Reference
							CKRs group			
W102	Extra	ECL1		Pocket	12.73	W	96.3			
						L	34.3			
A141	Intra	ICL2	3.57		11.49	A	82.0			
						R	83.8			
G207	Extra	TM5	5.46	Pocket	15.89	G	90.4			
						W	96.7			
C218 *	Intra	ICL3	5.57		11.64	C	93.6			
						F	96.2			
I222 *	Intra	ICL3	5.61		11.47	I	75.4			
						F	96.2			
Y256	Extra	TM6	6.52	Pocket	11.30	N	77.3			
						V	48.6			
C296	Intra	TM7	7.47		12.72	C	88.6			
						L	59.0			
βHV										
T73 *	Intra	ICL1	2.39		9.04	T	92.5			(L) Scholten et al. (2012)
						S	50.1			
L136	Intra	ICL2	3.52	DRY	10.76	L	66.1			
						S	32.8			
D171	Extra	TM4	4.60	Pocket	8.92					(L) Tian et al. (2005)
						Y	47.7			
Y190	Extra	ECL2	–	Pocket	11.26	Y	47.3			(S) Zhou et al. (2001)
						N	70.6			
C274	Extra	ECL3	–		12.37	C	96.2			(C) Wu et al. (2010)
W102	Extra	ECL1	–	Pocket	11.59	W	96.5			
F104	Extra	ECL1	–		10.98	F	81.4			
						S	31.9			
K110	Extra	ECL1	3.26		10.19	K	85.5			
						I	44.8			
N119	Extra	TM3	3.35		8.46	N	48.1	G	33.1	
						P	37.0			
H140 *	Intra	ICL2	3.56		8.66	H	49.0			
						W	38.6			
W283	Extra	TM7	7.34		9.65	A	76.5			
						F	37.0			
pox										
C28	Extra	N	–		9.87	C	90.9			(C) Wu et al. (2010)
						Y	37.6			
P42	Extra	TM1	1.36		7.91	P	70.4			(L) Scholten et al. (2012)
						I	62.2			
T90	Extra	TM2	2.56	TxP	6.79	T	68.6			(S) Govaerts et al. (2001)
										(S) Alvarez Arias et al. (2003)
W94	Extra	TM2	2.60	Pocket	10.13	W	74.6			(L) Scholten et al. (2012)
						I	39.4			(S) Rosenkilde et al. (2010)
L208	Extra	TM5	5.47	Pocket	8.06	F	70.8			(S) Holst et al. (2010)
						M	67.6			
F248	Intra	TM6	6.44	Pocket	9.82	F	83.3			(S) Deupi and Standfuss (2011)
						S	49.9	T	33.8	(L) Surgand et al. (2006)
P27	Extra	N	–		7.38	P	56.2			
						D	46.0	E	30.8	

(Continued)

Table 2 | Continued

Residue (CXCR4)	Position	Region	B and W	Remarks	KL value	Frequency (%)				Reference
						CKRs group				
E31	Extra	N	–	Pocket	11.70					
						Y	95.2			
F36	Extra	N	1.30		7.00	F	64.3			
						V	35.1			
L61	Intra	TM1	1.55		6.93	L	41.2			
						T	44.2			
I215	Intra	TM5	5.54		7.28	M	73.4			
						L	65.3			
I221 *	Intra	ICL3	5.60		6.16					
						K	85.4			
S227	Intra	ICL3	–		6.53	L	45.4			
						K	85.4			
E277	Extra	ECL3	7.28		6.16	S	33.3			
						L	33.8	F	30.7	

(A) The list of the sites detected from the comparisons with five decoy receptor groups. (B) The list of the sites detected from the comparisons with five viral receptors. Each row corresponds to a site with a large KL information value. The first column indicates the residue type and the residue number of CXCR4, to which the selected site corresponded. "∗" Indicates a site located within 5 Å from the DRY motif in the CXCR4 structure. The second column indicates whether the site is located on the extracellular or intracellular side. The location was determined for the t-test (see Materials and Methods). The third column shows the position of the site in the primary structure of a GPCR (N-terminus, TM, ICL, ECL, and C-terminus). The fourth column indicates the site by the Ballesteros–Weinstein nomenclature. The fifth column provides remarks about the site, such as experiments and motifs. The "pocket" in this column was calculated at the CASTp site (Liang et al., 1998). A blank entry in the fifth column means that the site has not been characterized yet. The sixth column indicates the KL value of each site. The seventh column indicates the frequencies of the residues at each site. The upper half of the column indicates the frequencies for CKRs, and the lower half indicates the frequencies for the group under comparison. Only the residues with frequencies greater than 30% are shown. The eighth column indicates whether the site is involved in ligand binding (L), signaling (S), or conservation (C). The literature for experimental evidence or observations of the characteristics is also shown in the column, although the experiments were not always performed with the receptors under consideration, but with the homologs in the CKR family.

None of them, except for K308$^{8.49}$, was detected from the analyses of any decoy receptor group.

STATISTICAL TEST FOR THE SPATIAL BIAS OF THE SITES WITH LARGE KL INFORMATION VALUES

As shown in **Figure 4**, the distribution of the sites selected from the analyses with the decoy receptor groups seemed to be biased toward the cytosolic side of the CKR structure. In contrast, there did not seem to be any trends in the distribution of the sites obtained from the analyses with the viral receptor groups. To quantitatively examine the observations, the residues corresponding to the selected sites and the remaining residues were projected on the axis connecting the center of gravity of the ECLs including the N-terminal region, and that of the ICLs (see **Figure 5**). Based on the projection on the axis, *t*-tests were performed as described in the Section "Materials and Methods."

The results of the *t*-tests are summarized in **Table 3**. As shown in this table, the null hypothesis was rejected in three cases of the analyses with decoy receptor groups, CCRL1, CCBP2, and DARC, under the significance level of 5%. To examine the bias further, the one-sided *t*-test was applied to the observations about the decoy receptor groups. The null hypothesis was the same as that of the two-sided test, but the alternative hypothesis was that the average of the residue with the large KL value is smaller than that of the remaining residues. We found that the null hypothesis was rejected for four cases with decoy receptor groups, CCRL1, CCBP2, DARC, and CXCR7 (data not shown). That is, the distribution of the residues corresponding to the sites with large KL information values of the decoy receptor groups, except for CCRL2, was biased toward the intracellular side of the receptor. The two-sided *t*-test was also applied to the analyses of the viral receptor groups. In all cases, the null hypothesis was not rejected. This result suggested that the residues selected by the KL information values of the viral receptors were distributed on both the extracellular and intracellular sides.

DISCUSSION

DECOY RECEPTORS

The difference in the amino acid composition at an alignment site between two receptor groups, as evaluated by the KL information value, was considered to reflect the difference in the functional constraints at the site between the groups. As described above, decoy receptors are able to bind to CKs, but do not induce signaling. The sites detected by the KL information value would reflect the functional difference. Actually, the sites included in several motifs, such as DRY, CWxP, and NPxxY$_{5\text{-}6}$F, which are involved in signaling, were detected. The bias in the locations of the detected sites on the intracellular side was statistically significant by the two-sided or one-sided *t*-test in four out of five decoy receptor groups. Especially, all of the sites detected from the analysis of CCBP2

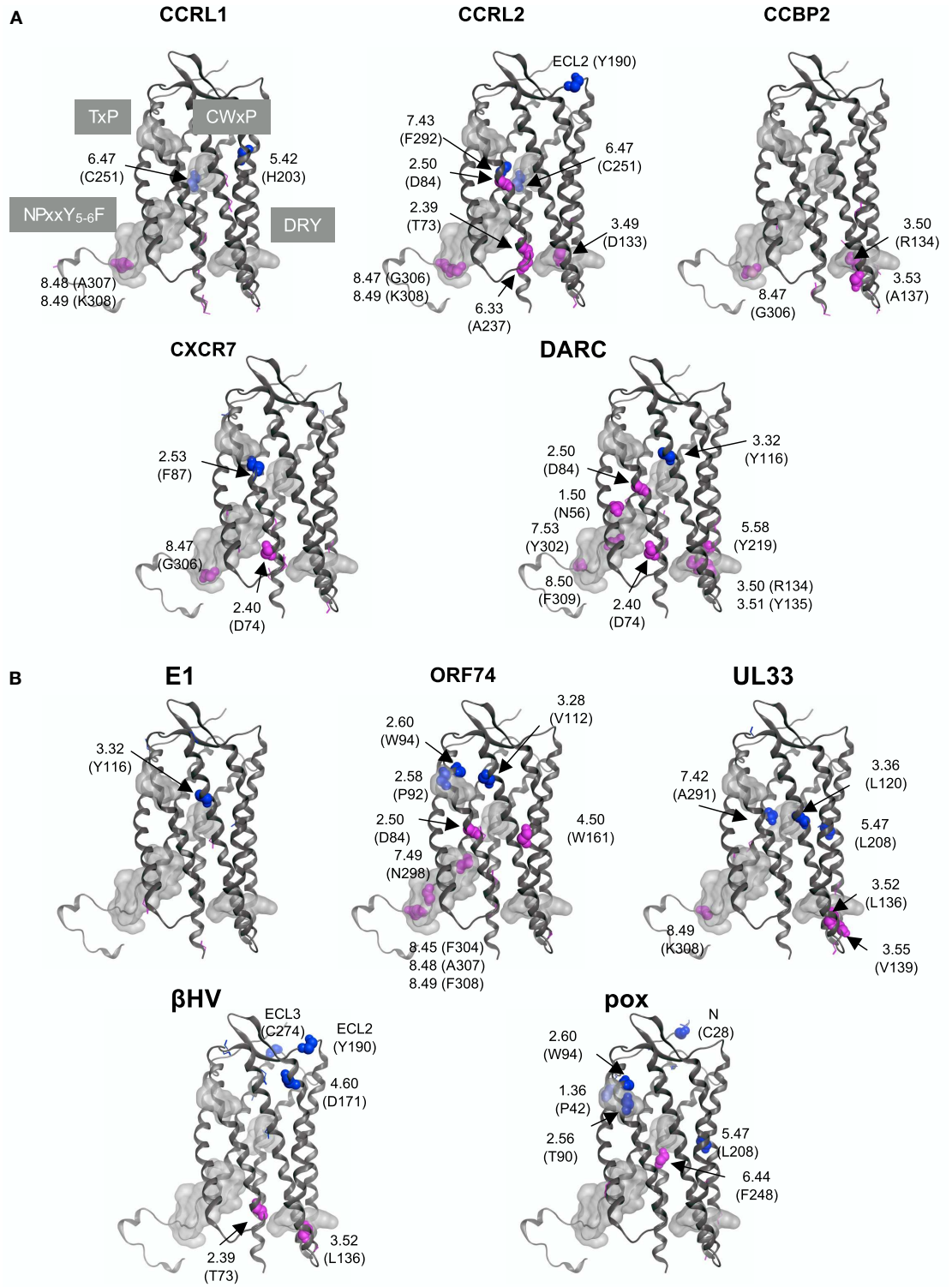


FIGURE 4 | Mapping of the sites with large KL information values on the CXCR4 structure. The sites detected from the comparisons with **(A)** five decoy receptor groups and with **(B)** five viral receptor groups are mapped on the main chain structure of CXCR4. The residues corresponding to the detected sites with information about function and/or motif are depicted by space filling models, and are indicated according to the Ballesteros–Weinstein nomenclature. The corresponding amino acid residue types and numbers of CXCR4 are also shown in parentheses. On the other hand, the sites without any information are indicated by line models. The four motif regions are indicated by gray surface models. The residues that mapped on the extracellular side are colored blue, and those that mapped on the intracellular side are colored red.

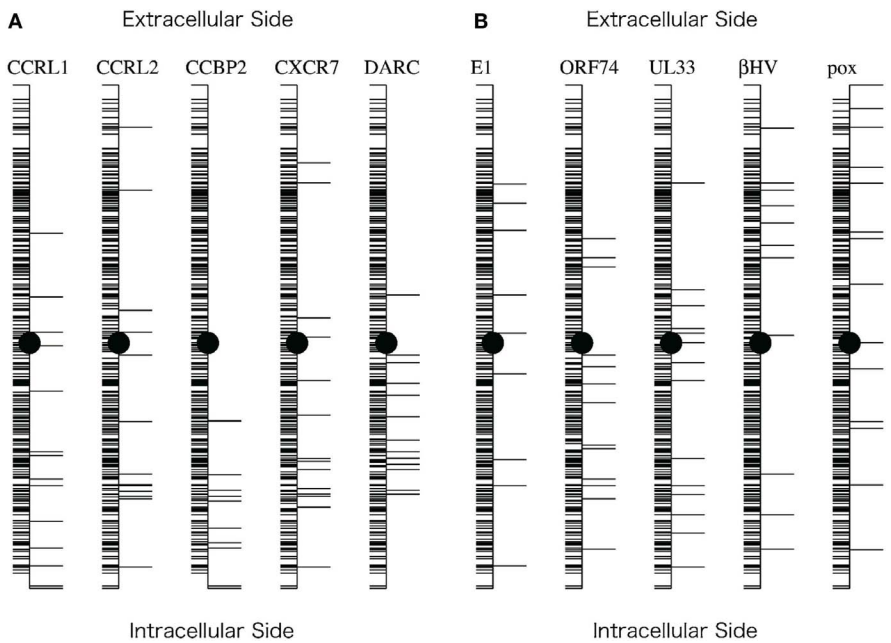


FIGURE 5 | Projection of the residues on the axis connecting the intracellular and extracellular sides of the receptor. (A) Projections corresponding to the five comparisons with the decoy receptor groups. **(B)** Projections corresponding to the five comparisons with the viral receptor groups. The axis generation method and the projection method are described in the Section "Materials and Methods." In each figure, the vertical line indicates the axis. Horizontal lines on the right side of the axis indicate the projected positions of the residues corresponding to the sites selected based on the KL information values. Horizontal lines on the left side indicate the projected positions of the remaining residues. The filled circles indicate the midpoint of the axis.

were located on the intracellular side. The test with the CCRL2 group was the only one that did not suggest a statistically significant bias in the distribution of the detected sites. As described above, CCRL2 is also able to bind to chemerin (Yoshimura and Oppenheim, 2011). The adaptation to the new ligand may have introduced the change in the functional constraints on the extracellular side, which may be the reason why the null hypothesis was not rejected. This observation suggested that the functional divergence of CCRL2 was induced under different selective pressure, as compared to the other decoy receptors after gene duplication. CCRL2 forms a gene cluster together with the genes for CCR1, 2, 3, and 5 in several mammalian genomes (Nomiyama et al., 2011). The close relationship of CCRL2 to these CKRs and its distant relationship to the other decoy receptors in the phylogenetic tree (**Figure 3**) were consistent with the conservation of the gene orders in the genomes, although the bootstrap probabilities for the relationships were not always high. The evolutionary relationship and the conserved gene order, together with the acquisition of binding activity to a new ligand, suggested a unique evolutionary position of CCRL2 relative to the other decoy receptors.

The lack of signal transduction activity in the decoy receptors is attributed to the degeneration of the DRY motif (Comerford et al., 2007). Our study suggested that the degenerations of other motifs and functional residues may also be related to functional changes. For example, two decoy groups, CCRL1 and CXCR7, contained the typical DRY motif. However, the sites in other motifs that are related to the conformational change associated with the active-inactive switch had large KL information values in these decoy receptors. This observation suggested that the constraints for the residue conservation at the sites in the traditional CKRs are looser in the two decoy receptor groups (see **Table 2**). In addition to the motif sites, the highly conserved sites in the TM regions of GPCRs, including the traditional CKRs (x.50 in the Ballesteros–Weinstein nomenclature), had large KL information values in the analyses with several decoy receptor groups. The use of different amino acid residues at such sites may lead to functional and/or structural changes. Several sites with uncharacterized functional relationships also showed large KL information values. Most of them were

Table 3 | Results of *t*-tests.

DECOY RECEPTORS	
CCRL1	3.87×10^{-3}
CCRL2	0.142
CCBP2	3.37×10^{-7}
CXCR7	0.066
DARC	1.51×10^{-4}
VIRAL RECEPTORS	
E1	0.981
ORF74	0.080
UL33	0.098
βHV	0.308
pox	0.144

The p-value for each two-sided t-test is shown. The details of the tests are described in the Section "Materials and Methods."

found in ICLs 2 and 3. As these loops are considered to interact with G proteins, the sites detected on the loops may be involved in the loss of the signaling function of the decoy receptors.

VIRAL RECEPTORS

We anticipated that the sites detected from the analyses with the viral receptor groups would be found on the extracellular side, since viral receptors exhibit signaling activity without ligand binding. As described above, however, the sites with the large KL information values were found not only on the extracellular side, but also on the intracellular side. We examined the detected sites from the different viewpoint. CASTp[4] (Liang et al., 1998) is a program to identify pocket regions in a given tertiary structure. When we applied CASTp to the coordinates of CXCR4, the pocket region corresponding to the ligand binding cavities of GPCRs was reported with the highest score. The residues consisting of the pocket region were mainly projected on the extracellular side of the axis (see **Figure 1**), although some residues were projected on the intracellular side. The numbers of detected sites located in the pocket regions of the five decoy receptor groups were 2, 2, 0, 1, and 1, whereas 3, 2, 5, 3, and 4 sites were located in the pocket regions of the five viral receptor groups (see **Table 2**). The number of sites was transformed into the ratio to the total number of detected sites for each receptor group. The one-sided t-test showed that the difference in the ratios between the decoy and viral receptor groups was statistically significant (p-value = 0.003864). That is, more sites were detected in the pocket region in the viral receptor groups, as compared to the decoy receptor groups. As shown in **Table 2**, in addition, about half of the sites in the pocket region have been characterized as being involved in ligand recognition. These sites are often occupied by conserved, bulky amino acid residues in CKRs. The result suggested that the functional constraints at the ligand binding region are different between the viral receptors and the traditional CKRs, as we first expected.

The sites in the DRY motif were not detected in any of the viral receptor groups. This motif was basically conserved in the viral receptors, except for the ORF74 group. A previous study reported that ORF74 performs signal transduction, despite the fact that the DRY motif is changed to DTW (Rosenkilde et al., 2005). They also showed that the introduction of the DRY sequence into ORF74 induces functional reduction. In our study, the sequences collected as the ORF74 group showed variations in this region. For example, equid herpesvirus 2 has DTW, whereas the rodent and primate herpesviruses have xRC or xRY. Each variation includes the residues identical to those of the original DRY motif, which may have reduced the KL information value and led to the failure in the detection of the sites. Instead, the sites in the TxP and $NPxxY_{5\text{-}6}F$ motifs and the sites spatially surrounding the DRY motif were detected from the analysis of the ORF74 group (see **Table 2**). The amino acid replacements in the two motifs, which are considered to be involved in the conformational change, and those of the residues near the DRY motif may have contributed to the maintenance of the signaling activity of ORF74, despite the deviation from the typical DRY motif. In contrast, no sites in any motif were detected from the comparison with the E1 group. The E1 receptor reportedly lacks constitutive signaling activity (Rosenkilde et al., 2008). The conservation of the motifs suggested the difference in the signaling functions between the E1 group and other viral receptor groups.

We had not expected to detect the sites on the intracellular side from the comparisons with the viral receptor groups, since these receptors exhibit signaling activity without ligand binding. However, quite a few sites with large KL information values were also found on the intracellular side. As described above, the overlap of the selected sites between the decoy receptors and the viral receptors was small. The difference in the selected sites on the intracellular side between the viral receptor groups and the decoy receptor groups may be basically related to the difference in the activities of the receptor groups. That is, the sites of the viral receptor groups under the constraint to maintain the signaling without ligand binding may be different from the sites of the decoy receptor groups, where the functional constraints may have been weakened due to the loss of the signaling activity.

CONCLUSION

We have identified the alignment sites (and the corresponding amino acid residues) that may be responsible for the functional changes from CKRs to decoy receptors or viral receptors. The distributions of the identified residues on the tertiary structure seemed to reflect the functional differences. This prediction could be examined by an experimental study, such as amino acid replacement, or a computational study with molecular dynamic simulations. Such studies could provide deep insights into the mechanism of GPCR signaling through conformational changes. The experimental and computational confirmations of our prediction remain as future endeavors.

ACKNOWLEDGMENTS

Hiromi Daiyasu was supported by a Grant-in-Aid for Scientific Research and MEXT SPIRE Supercomputational Life Science. Hiroyuki Toh was supported by the Target Tanpaku program.

[4]http://sts.bioengr.uic.edu/castp/

REFERENCES

Adachi, J., and Hasegawa, M. (1996). *MOLPHY (Programs for Molecular Phylogenetics), Version 2.3b3.* Tokyo: Institute of Statistical Mathematics.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Alvarez Arias, D., Navenot, J. M., Zhang, W. B., Broach, J., and Peiper, S. C. (2003). Constitutive activation of CCR5 and CCR2 induced by conformational changes in the conserved TXP motif in transmembrane helix 2. *J. Biol. Chem.* 278, 36513–36521.

Ballesteros, J. A., and Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* 25, 366–428.

Bonecchi, R., Savino, B., Borroni, E. M., Mantovani, A., and Locati, M. (2010). Chemokine decoy receptors: structure-function and biological properties. *Curr. Top. Microbiol. Immunol.* 341, 15–36.

Camarda, G., Spinetti, G., Bernardini, G., Mair, C., Davis-Poynter, N., Capogrossi, M. C., and Napolitano, M. (1999). The equine herpesvirus 2 E1 open reading frame encodes a functional chemokine receptor. *J. Virol.* 73, 9843–9848.

Choi, W. T., Tian, S., Dong, C. Z., Kumar, S., Liu, D., Madani, N., An, J., Sodroski, J. G., and Huang, Z. (2005). Unique ligand binding sites on CXCR4 probed by a chemical biology approach: implications for the design of selective human immunodeficiency virus type 1 inhibitors. *J. Virol.* 79, 15398–15404.

Comerford, I., Litchfield, W., Harata-Lee, Y., Nibbs, R. J., and McColl, S. R. (2007). Regulation of chemotactic networks by "atypical" receptors. *Bioessays* 29, 237–247.

Comerford, I., Milasta, S., Morrow, V., Milligan, G., and Nibbs, R. (2006). The chemokine receptor CCX-CKR mediates effective scavenging of CCL19 in vitro. *Eur. J. Immunol.* 36, 1904–1916.

Deupi, X., and Standfuss, J. (2011). Structural insights into agonist-induced activation of G-protein-coupled receptors. *Curr. Opin. Struct. Biol.* 21, 541–551.

Ewens, W. J., and Grant, G. R. (2001). *Statistical Methods in Bioinformatics: An Introduction (Statistics for Biology and Health)*. New York: Springer.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package), Version 3.5c.* Seattle: University of Washington.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth. Enzymol.* 266, 418–427.

Govaerts, C., Blanpain, C., Deupi, X., Ballet, S., Ballesteros, J. A., Wodak, S. J., Vassart, G., Pardo, L., and Parmentier, M. (2001). The TXP motif in the second transmembrane helix of CCR5. A structural determinant of chemokine-induced activation. *J. Biol. Chem.* 276, 13217–13225.

Graham, G. J. (2009). D6 and the atypical chemokine receptor family: novel regulators of immune and inflammatory processes. *Eur. J. Immunol.* 39, 342–351.

Gruijthuijsen, Y. K., Casarosa, P., Kaptein, S. J., Broers, J. L., Leurs, R., Bruggeman, C. A., Smit, M. J., and Vink, C. (2002). The rat cytomegalovirus R33-encoded G protein-coupled receptor signals in a constitutive fashion. *J. Virol.* 76, 1328–1338.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664–1674.

Hannenhalli, S. S., and Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* 303, 61–76.

Henikoff, S., and Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* 243, 574–578.

Holst, B., Nygaard, R., Valentin-Hansen, L., Bach, A., Engelstoft, M. S., Petersen, P. S., Frimurer, T. M., and Schwartz, T. W. (2010). A conserved aromatic lock for the tryptophan rotameric switch in TM-VI of seven-transmembrane receptors. *J. Biol. Chem.* 285, 3973–3985.

Ichihara, H., Daiyasu, H., and Toh, H. (2004). How does a topological inversion change the evolutionary constraints on membrane proteins? *Protein Eng. Des. Sel.* 17, 235–244.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* 9, 286–298.

Landgraf, R., Fisher, D., and Eisenberg, D. (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* 12, 943–951.

Leick, M., Catusse, J., Follo, M., Nibbs, R. J., Hartmann, T. N., Veelken, H., and Burger, M. (2010). CCL19 is a specific ligand of the constitutively recycling atypical human chemokine receptor CRAM-B. *Immunology* 129, 536–546.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science* 7, 1884–1897.

Maussang, D., Vischer, H. F., Leurs, R., and Smit, M. J. (2009). Herpesvirus-encoded G protein-coupled receptors as modulators of cellular function. *Mol. Pharmacol.* 76, 692–701.

Murphy, P. M., Baggiolini, M., Charo, I. F., Hébert, C. A., Horuk, R., Matsushima, K., Miller, L. H., Oppenheim, J. J., and Power, C. A. (2000). International Union of Pharmacology. XXII. Nomenclature for chemokine receptors. *Pharmacol. Rev.* 52, 145–176.

Najarro, P., Gubser, C., Hollinshead, M., Fox, J., Pease, J., and Smith, G. L. (2006). Yaba-like disease virus chemokine receptor 7L, a CCR8 orthologue. *J. Gen. Virol.* 87(Pt 4), 809–816.

Naumann, U., Cameroni, E., Pruenster, M., Mahabaleshwar, H., Raz, E., Zerwes, H. G., Rot, A., and Thelen, M. (2010). CXCR7 functions as a scavenger for CXCL12 and CXCL11. *PLoS ONE* 5, e9175. doi:10.1371/journal.pone.0009175

Nomiyama, H., Osada, N., and Yoshie, O. (2011). A family tree of vertebrate chemokine receptors for a unified nomenclature. *Dev. Comp. Immunol.* 35, 705–715.

Nygaard, R., Frimurer, T. M., Holst, B., Rosenkilde, M. M., and Schwartz, T. W. (2009). Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol. Sci.* 30, 249–259.

Nygaard, R., Valentin-Hansen, L., Mokrosinski, J., Frimurer, T. M., and Schwartz, T. W. (2010). Conserved water-mediated hydrogen bond network between TM-I, -II, -VI, and -VII in 7TM receptor activation. *J. Biol. Chem.* 285, 19625–19636.

Rosenkilde, M. M., Benned-Jensen, T., Frimurer, T. M., and Schwartz, T. W. (2010). The minor binding pocket: a major player in 7TM receptor activation. *Trends Pharmacol. Sci.* 31, 567–574.

Rosenkilde, M. M., Kledal, T. N., and Schwartz, T. W. (2005). High constitutive activity of a virus-encoded seven transmembrane receptor in the absence of the conserved DRY motif (Asp-Arg-Tyr) in transmembrane helix 3. *Mol. Pharmacol.* 68, 11–19.

Rosenkilde, M. M., Smit, M. J., and Waldhoer, M. (2008). Structure, function and physiological consequences of virally encoded chemokine seven transmembrane receptors. *Br. J. Pharmacol.* 153(Suppl. 1), S154–S166.

Rosenkilde, M. M., Waldhoer, M., Lüttichau, H. R., and Schwartz, T. W. (2001). Virally encoded 7TM receptors. *Oncogene* 20, 1582–1593.

Sahagun-Ruiz, A., Sierra-Honigmann, A. M., Krause, P., and Murphy, P. M. (2004). Simian cytomegalovirus encodes five rapidly evolving chemokine receptor homologues. *Virus Genes* 28, 71–83.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Scholten, D. J., Canals, M., Maussang, D., Roumen, L., Smit, M., Wijtmans, M., de Graaf, C., Vischer, H., and Leurs, R. (2012). Pharmacological modulation of chemokine receptor function. *Br. J. Pharmacol.* 165, 1617–1643.

Simon, A. L., Stone, E. A., and Sidow, A. (2002). Inference of functional regions in proteins by quantification of evolutionary constraints. *Proc. Natl. Acad. Sci. U.S.A.* 99, 2912–2917.

Slinger, E., Langemeijer, E., Siderius, M., Vischer, H. F., and Smit, M. J. (2011). Herpesvirus-encoded GPCRs rewire cellular signaling. *Mol. Cell. Endocrinol.* 331, 179–184.

Surgand, J. S., Rodrigo, J, Kellenberger, E., and Rognan, D. (2006). A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* 62, 509–538.

Thelen, M., and Thelen, S. (2008). CXCR7, CXCR4 and CXCL12: an eccentric trio? *J. Neuroimmunol.* 198, 9–13.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680.

Tian, S., Choi, W. T., Liu, D., Pesavento, J., Wang, Y., An, J., Sodroski, J. G., and Huang, Z. (2005). Distinct functional sites for human immunodeficiency virus type 1 and stromal cell-derived factor 1alpha on CXCR4 transmembrane helical domains. *J. Virol.* 79, 12667–12673.

Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., Brooun, A., Wells, P., Bi, F. C., Hamel, D. J., Kuhn, P., Handel, T. M., Cherezov, V., and Stevens, R. C. (2010). Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* 330, 1066–1071.

Yoshimura, T., and Oppenheim, J. J. (2011). Chemokine-like receptor 1 (CMKLR1) and chemokine (C-C motif) receptor-like 2 (CCRL2); two multifunctional receptors with unusual properties. *Exp. Cell Res.* 317, 674–684.

Zhou, N., Luo, Z., Luo, J., Liu, D., Hall, J. W., Pomerantz, R. J., and Huang, Z. (2001). Structural and functional characterization of human CXCR4 as a chemokine receptor and HIV-1 co-receptor by mutagenesis and molecular modeling studies. *J. Biol. Chem.* 276, 42826–42833.

Zlotnik, A., and Yoshie, O. (2000). Chemokines: a new classification system and their role in immunity. *Immunity* 12, 121–127.

Zlotnik, A., Yoshie, O., and Nomiyama, H. (2006). The chemokine and chemokine receptor superfamilies and their molecular evolution. *Genome Biol.* 7, 243.

Conflict of Interest Statement: 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.

Received: 06 June 2012; paper pending published: 17 June 2012; accepted: 05 July 2012; published online: 26 July 2012.

Citation: Daiyasu H, Nemoto W and Toh H (2012) Evolutionary analysis of functional divergence among chemokine receptors, decoy receptors, and viral receptors. Front. Microbio. ***3****:264. doi: 10.3389/fmicb.2012.00264*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 July 2012
doi: 10.3389/fmicb.2012.00254

An assembly model of Rift Valley fever virus

Mirabela Rusu[1†], Richard Bonneau[2,3], Michael R. Holbrook[4,5], Stanley J. Watowich[6], Stefan Birmanns[1], Willy Wriggers[7†] and Alexander N. Freiberg[4]*

[1] School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA
[2] Center for Genomics and Systems Biology, Biology Department, New York University, New York, NY, USA
[3] Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[4] Department of Pathology, Institute for Human Infections and Immunity, University of Texas Medical Branch, Galveston, TX, USA
[5] National Institute of Allergy and Infectious Diseases-Integrated Research Facility, Frederick, MD, USA
[6] Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA
[7] Department of Physiology and Biophysics, Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, NY, USA

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Dale L. Barnard, Utah State University, USA
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan

***Correspondence:**
Alexander N. Freiberg, Department of Pathology, University of Texas Medical Branch, 301 University Boulevard, Galveston, TX 77555-0609, USA.
e-mail: anfreibe@utmb.edu

†Present address:
Mirabela Rusu, Biomedical Engineering Department, Rutgers State University of New Jersey, Piscataway, NJ, USA;
Willy Wriggers, D. E. Shaw Research, New York, NY, USA.

Rift Valley fever virus (RVFV) is a bunyavirus endemic to Africa and the Arabian Peninsula that infects humans and livestock. The virus encodes two glycoproteins, Gn and Gc, which represent the major structural antigens and are responsible for host cell receptor binding and fusion. Both glycoproteins are organized on the virus surface as cylindrical hollow spikes that cluster into distinct capsomers with the overall assembly exhibiting an icosahedral symmetry. Currently, no experimental three-dimensional structure for any entire bunyavirus glycoprotein is available. Using fold recognition, we generated molecular models for both RVFV glycoproteins and found significant structural matches between the RVFV Gn protein and the influenza virus hemagglutinin protein and a separate match between RVFV Gc protein and Sindbis virus envelope protein E1. Using these models, the potential interaction and arrangement of both glycoproteins in the RVFV particle was analyzed, by modeling their placement within the cryo-electron microscopy density map of RVFV. We identified four possible arrangements of the glycoproteins in the virion envelope. Each assembly model proposes that the ectodomain of Gn forms the majority of the protruding capsomer and that Gc is involved in formation of the capsomer base. Furthermore, Gc is suggested to facilitate intercapsomer connections. The proposed arrangement of the two glycoproteins on the RVFV surface is similar to that described for the alphavirus E1-E2 proteins. Our models will provide guidance to better understand the assembly process of phleboviruses and such structural studies can also contribute to the design of targeted antivirals.

Keywords: bunyavirus assembly, protein structure prediction, hybrid modeling, multi-body refinement, multi-resolution registration

INTRODUCTION

Rift Valley fever virus (RVFV) is a member of the family *Bunyaviridae* (genus *Phlebovirus*), transmitted primarily by mosquitoes and is endemic throughout much of Africa and, in recent years in the Arabian Peninsula. The virus causes outbreaks in a wide range of vertebrate hosts, with humans and livestock being the most affected. Infection of livestock can result in economically disastrous abortion storms and high mortality among young animals. In humans, the virus causes a variety of pathologic effects with less than 1% of infections thought to result in fatal hemorrhagic fever or encephalitis (MMWR, 2007). However, during the outbreak in Kenya, from November 2006 to January 2007, the fatality rate in humans reached nearly 30% (MMWR, 2007). RVFV is considered a high consequence emerging infectious disease threat and is also of concern as a bioterrorism agent. RVFV is classified as Category A select agent by CDC and USDA. Currently, there are no commercially available vaccines or therapeutics.

RVFV is a typical enveloped bunyavirus and has a trisegmented, negative-sense RNA genome, and most likely enters the host cells via receptor-mediated endocytosis, which requires an acid-activated membrane fusion step (Lozach et al., 2010, 2011). The two glycoproteins, Gn and Gc, are expressed as a precursor polypeptide, which is then co-translationally cleaved prior to maturation of the envelope glycoproteins (Collett et al., 1985; Wasmoen et al., 1988). For transport from the endoplasmic reticulum to the Golgi apparatus, both newly synthesized glycoproteins are required (Gerrard and Nichol, 2002). Within the virion, the surface glycoproteins are anchored in the envelope membrane as type-I integral membrane proteins and are responsible for receptor recognition and binding, and entry into target cells through fusion between viral and cellular membranes. In contrast to most other negative-stranded RNA viruses, bunyaviruses lack a matrix protein and the cytoplasmic tails of Gn and Gc likely interact directly with the ribonucleoprotein complex inside the virus particle (Overby et al., 2007; Piper et al., 2011). Gn and Gc form oligomers and are organized on the virus surface as cylindrical hollow spikes that cluster into distinct capsomers. The virus surface is covered with 122 capsomers arranged on an icosahedral lattice with a triangulation number of 12 (Freiberg et al., 2008; Huiskonen et al., 2009;

Sherman et al., 2009). Computational studies have predicted RVFV Gc to be a class II viral fusion protein (Garry and Garry, 2004). Owing to their importance in the process of virion maturation, receptor binding, and fusion with the host cell, both glycoproteins form attractive targets for the design of antiviral drugs blocking the receptor binding and/or fusion processes.

Structural data for bunyavirus glycoproteins are available for the hantavirus and Crimean–Congo hemorrhagic fever virus Gn cytoplasmic tails (Estrada et al., 2009, 2011; Estrada and De Guzman, 2011). However, no crystallographic data are available for any bunyavirus glycoprotein ectodomain. Bioinformatic investigation and molecular homology modeling of the bunyavirus Gc proteins of the five different genera revealed that they share a limited number of similar sequences with each other and that they have sequence similarity with the alphavirus E1 protein, suggesting that bunyavirus Gc proteins could be class II viral fusion proteins (Garry and Garry, 2004; Tischler et al., 2005; Plassmeyer et al., 2007; Hepojoki et al., 2010). Further, experiments with members from other bunyavirus genera supported the major role Gc plays during fusion with the host cell membrane and entry (Plassmeyer et al., 2005, 2007; Shi et al., 2009). Three-dimensional (3D) molecular model structures for Gc have been described for members of different genera, such as La Crosse virus (*Orthobunyavirus*), Sandfly fever virus (*Phlebovirus*), Andes and Tula viruses (Hantaviruses), and have been used successfully to study the functionality of fusion peptides and the interaction and oligomerization of glycoproteins (Garry and Garry, 2004; Tischler et al., 2005; Hepojoki et al., 2010; Soldan et al., 2010). Most of these studies targeted the Gc protein; much less information is available for the Gn protein. It has been suggested that the phlebovirus Gn plays a role in receptor binding and that it might have structural similarity to the alphavirus E2 protein (Garry and Garry, 2004).

To better understand the assembly of bunyaviruses and the functional interaction between Gn and Gc glycoproteins, we sought to generate 3D structure models for RVFV Gn and Gc monomers using bioinformatic approaches. Specifically, homology models were created following established virus protein prediction strategies (Garry and Garry, 2004, 2008, 2009; Tischler et al., 2005; Lee et al., 2009; Hepojoki et al., 2010). Subsequently, we used these model structures to evaluate possible positions within the existing cryo-electron microscopy (cryoEM) density map of RVFV virions to predict protein–protein interaction interfaces and to propose an assembly model for RVFV. We suggest that RVFV Gn and Gc are arranged topologically within the virus particle, with some similarity to the E1 and E2 proteins of alphaviruses. Our model indicates that RVFV Gn could be involved in receptor binding and covers the fusion loop of Gc at neutral pH, while Gc is proposed to play a major role during the membrane fusion step.

MATERIALS AND METHODS

PROTEIN SEQUENCE ANALYSIS

For sequence and structural analysis, the RVFV vaccine strain MP-12 glycoprotein encoding nucleotide sequence (GenBank DQ380208) was used. The secondary structure of RVFV Gn and Gc, respectively, were examined using Jpred3[1] (Cole et al., 2008).

[1]http://www.compbio.dundee.ac.uk/www-jpred/

Table 1 | Prediction of location of transmembrane domains.

	RVFV Gn[a]	RVFV Gc[a]
EXPASY	429–449 [21][b]	470–490 [21]
HMMTOP	429–451 [23]	470–494 [25]
	517–535 [19][c]	
SOSUI	433–455 [23]	469–491 [23]
	515–536 [22][c]	
TMHMM	432–454 [23]	469–491 [23]
Average	429–455 [27]	469–494 [26]
	515–536 [22][c]	

[a] *RVFV MP-12 glycoprotein length [SwissProt #P21401] Gn: 536 aa; Gn: 507 aa.*

[b] *Numbers indicate the length of the transmembrane domains.*

[c] *Second TMD in Gn corresponds to signal peptide.*

To define the location of the glycoprotein transmembrane domains (TMD) (**Table 1**), as well as cytoplasmic tail domains (CTD), the EXPASY[2], HMMTOP[3], SOSUI[4], and TMHMM[5] servers were used (Hirokawa et al., 1998; Tusnady and Simon, 1998; Krogh et al., 2001). We used the NetN Glyc 1.0 Server[6] to predict the locations of N-glycosylation sites.

PROTEIN STRUCTURE PREDICTION

Initial backbone models were generated using the fold recognition Meta Server[7] (Kajan and Rychlewski, 2007), which used alignments from the FFAS_03 program[8] to the two templates (Jaroszewski et al., 2005). These models agreed with alignments found using other fold recognition methods, increasing our confidence in these fold predictions. Side chains were added and models were refined using Modeller[9] (Eswar et al., 2006). The atomic model of the Gn glycoprotein was generated based on the 1918 influenza H1 hemagglutinin protein (PDB ID: 1RD8, Stevens et al., 2004), specifically the HA1 chain, for which a 14.15% sequence identity was observed. Similarly, the atomic model of the Gc glycoprotein was built based on the Semliki Forest virus (SFV) structural E1 protein fitted into the Sindbis virus cryoEM map (PDB ID: 1LD4, Zhang et al., 2002) from an observed sequence identity of 13.83%. In addition to these structures sub-optimal FFAS_03 alignments and derived models were also evaluated in the context of the cryoEM density including alignments of RVFV Gc to PDB structures of the Chikungunya E1-E2 envelope glycoprotein complex fitted into the SFV cryoEM map (PDB ID: 2XFC; PDB ID: 1RER, Gibbons et al., 2004; Li et al., 2010), dengue virus E protein (PDB ID: 1P58, Zhang et al., 2003), integrin binding fragment of human fibrillin-1 (PDB ID: 1UZJ, Lee et al., 2004) and alignment of Gn to the EAP45/ESCRT GLUE domain (PDB ID: 2HTH–chain A, Alam et al., 2006). All of the proteins identified

[2]http://expasy.org/
[3]http://www.enzim.hu/hmmtop/
[4]http://bp.nuap.nagoya-u.ac.jp/sosui/
[5]http://www.cbs.dtu.dk/services/TMHMM/
[6]http://www.cbs.dtu.dk/services/NetNGlyc/
[7]http://meta.bioinfo.pl/submit_wizard.pl
[8]http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl
[9]http://salilab.org/modeller/

as similar to Gc are class II fusion proteins, and due to the similar sequence identity between the different homologs, an alternative atomic model of the RVFV Gc protein was built based on the structure of the Chikungunya virus E1 protein fitted into the cryoEM reconstruction of SFV (PDB ID: 2XFC; chain A, Voss et al., 2010). Since the overall shape of Gc is conserved between the models based on the Sindbis virus and SFV E1 protein, we did not further evaluate their positioning in the RVFV cryoEM map.

FITTING OF GLYCOPROTEIN STRUCTURES INTO THE cryoEM DENSITY

The 3D models of the RVFV Gn and Gc glycoproteins were fitted into the RVFV vaccine strain MP-12 cryoEM map (Sherman et al., 2009). The organization of the two glycoproteins within the RVFV envelope was identified following a hybrid approach that combined an interactive exploration of the exhaustive search outcome with a multi-body refinement procedure. The multi-body refinement is described in detail in Birmanns et al. (2011) and a summary is provided here. The multi-step approach (**Figure 2**; see Results) was applied to generate an atomic model for the triangular face of the RVFV. First, an exhaustive search using the tool colores from the package Situs[10] (Wriggers, 2010) was applied to explore possible placement for each of the Gc and Gn glycoproteins. The molecular modeling software Sculptor[11] (Birmanns et al., 2011) was used for the interactive exploration of the exhaustive search results to select placements that are in agreement with computed Gn/Gc ratio within each capsomer type (Huiskonen et al., 2009; Sherman et al., 2009) and that show reduced steric clashing. Several such docking locations were identified for both Gc and Gn, and multiple models were iteratively refined by searching for the architecture that best described the density of the asymmetric unit.

RESULTS

FOLD RECOGNITION OF THE RVFV GLYCOPROTEINS

Both RVFV glycoproteins, Gn and Gc, are known to be type-I integral transmembrane proteins. Before obtaining fold recognition and molecular model predictions of the two RVFV glycoproteins, the primary amino acid sequences of the entire Gn and Gc were analyzed for predicted TMD, ecto- and endo-domains (CTD), glycosylation sites, and consensus secondary structure prediction elements (**Figure 1A**). Gn is predicted to displays a mixture of α-helical, β-strands, and random coil secondary structural elements (**Figure 1B**). The N-terminus has a slightly higher content of β-strands, while the C-terminus is rich in α-helical elements located in the regions predicted for the TMD and CTD. Rift Valley fever virus Gc exhibits predominantly β-strands, a very low content of α-helices and a high content of random coiling (**Figure 1C**). Most of the α-helical elements are found in the regions predicted for the transmembrane and short CTD, as already described for the Gn protein. Garry and Garry (2004) suggested that the Gc glycoproteins of bunyaviruses are class II viral fusion proteins. Class II fusion proteins, such as the envelope glycoprotein E of tick-borne encephalitis virus and the E1 protein of Sindbis virus, are composed mostly of antiparallel β-sheets, similar to the secondary structure prediction for RVFV Gc.

[10]http://situs.biomachina.org/
[11]http://sculptor.biomachina.org/

MODEL BUILDING AND STRUCTURAL DESCRIPTION OF THE RVFV Gn AND Gc GLYCOPROTEINS

To model the 3D structures of Gn and Gc, and to verify that RVFV Gc adopts a class II fusion protein fold, we initially focused on the near full-length RVFV Gn (530 aa in length) and Gc (507 aa in length) protein sequences (**Figure 1A**). However, molecular models could only be generated for the two glycoprotein ectodomains, so the TMDs and CTDs were removed from further analysis. Throughout the manuscript, the terms RVFV Gn and Gc are used to describe the ectodomain for each glycoprotein and not the entire glycoprotein itself.

The fold recognition revealed that the best matching profile for RVFV Gn resulted in a hit which had structural similarity to the Influenza 1918 human H1 hemagglutinin, specifically the receptor binding domain HA1 (**Figure 1B**). The molecular model generated for RVFV Gc was obtained based on the Sindbis virus and Chikungunya E1 proteins (**Figure 1C**). This result was expected, since bioinformatic analysis had already predicted that the bunyavirus Gc protein has sequence similarity with the alphavirus E1 protein, suggesting that bunyavirus Gc proteins are class II viral fusion proteins (Garry and Garry, 2004). Furthermore, all of the proteins identified as similar to Gc are class II fusion proteins (see Materials and Methods). As shown in **Figure 1C**, the modeled structure for RVFV Gc resembles the overall fold of a class II fusion protein (Kielian, 2006; Kielian and Rey, 2006).

The Gn and Gc model was evaluated in terms of stereochemical and geometric parameters such as bond lengths, bond angles, torsion angles, and packing environment and was found to satisfy all stereochemical criteria (assessed by VADAR statistics software package; Willard et al., 2003). For the 3D models, the (Φ, Ψ) values calculated for each amino acid residue of the individual model structures were within the allowed region of the Ramachandran plot (Ramachandran and Sasisekharan, 1968; data not shown).

The Gc protein consists of three domains, with predominantly β-strand content, which is in accordance with the amino acid sequence analysis (data not shown). The nomenclature of these three domains has been defined by analogy with the alphavirus E1 protein, domain I (central domain), domain II and domain III. Domain II contains two predicted glycosylation sites at positions N794 and N829 and also bears the predicted fusion loop of RVFV Gc that potentially inserts into the target host membrane during the pH-dependent virus fusion step (Garry and Garry, 2004). The location of the fusion loop is highlighted in purple in **Figure 1C**. Domain III, separated from the first two domains by a short stretch, forms an Ig-like β-barrel structure and contains two glycosylation sites at positions N1035 and N1077. On-going studies in our laboratory found that removal of the glycosylation sites in Gc has a negative effect on virus assembly and maturation (ANF, unpublished results). In contrast, the predicted 3D model for the ectodomain of RVFV Gn represents an elongated structure with a globular head domain (**Figure 1B**). The membrane-distal domain consists of a globular head, which displays a mixture of β-strands, and slightly less α-helical and random coil content. A stem-like region connects the globular domain with the TMD, which is not displayed in the 3D structure. The head domain also contains the glycosylation site at position N285.

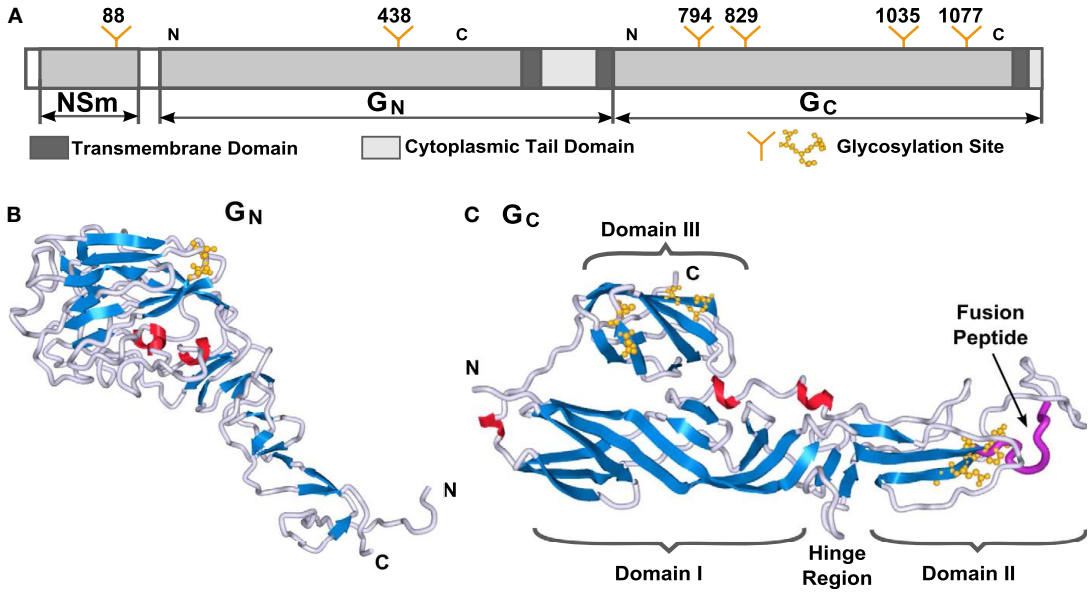


FIGURE 1 | Three-dimensional structure models of RVFV Gn and Gc proteins. (A) Schematic representation of the RVFV M-segment polyprotein. Transmembrane and cytoplasmic tail domains are highlighted in dark gray or white bars, respectively. N-Glycosylation sites are indicated with the position of the respective Asn residue. The regions of the two glycoproteins used for molecular modeling are indicated with N and C. 3D molecular models for RVFV **(B)** Gn and **(C)** Gc are shown. Secondary structures are highlighted in blue for β-strands, red for α-helices, and gray for turns. The predicted location of the fusion peptide within Gc is represented in purple. The domain nomenclature in modeled Gc were used in adoption to the alphavirus E1 protein. The molecular graphics in this paper were generated with Sculptor (Birmanns et al., 2011) and Chimera (Pettersen et al., 2004).

The predicted N-glycosylation sites were in agreement with the findings from Kakach et al., 1989; yellow spheres in **Figures 1B,C**). All glycosylation sites on Gn and Gc are fully surface accessible, which supports our model structures.

GLYCOPROTEIN MODELING IN THE RVFV PARTICLE

Recently, we determined the 3D structure of the RVFV vaccine strain MP-12 by single-particle cryoEM at 27 Å resolution (Sherman et al., 2009). The reconstruction shows the $T = 12$ icosahedral envelope of the virion, depicting different types of capsomers (Freiberg et al., 2008; Sherman et al., 2009). Using the two model structures of Gn and Gc, we sought to identify their organization within capsomers by means of cross-correlation and built a model for the entire glycoprotein layer of the virion.

The glycoprotein layer is composed of capsomers showing different symmetry order (Freiberg et al., 2008; Huiskonen et al., 2009; Sherman et al., 2009). Pentons are located around the fivefold symmetry axis while hexons organize around the threefold, quasi threefold, and twofold axes. Although an icosahedral symmetry is imposed when reconstructing the cryoEM map of the virus, the hexons show different symmetry orders and can be averaged to increase the level of detail of the volumetric data. Such practice is common in modeling structures at low resolution, where averaging is applied to increase the signal-to-noise ratio of the data. First, the three different types of hexons were extracted, aligned, and then an averaged volume from the 11 copies was computed (rotations included). This averaged hexon, displaying a sixfold symmetry, was used to construct an average density for the asymmetric unit and the corresponding triangular face. The cryoEM density of the averaged face was utilized as target volume inside the envelope for the global docking of the Gc and Gn glycoproteins, respectively, inside the envelope. An exploration of all possible translations and rotations (9° step size) was performed for each glycoprotein with the colores tool of the Situs package (Wriggers, 2010). This exhaustive search allowed the estimation of the optimal cross-correlation coefficient, providing the list of top scoring placements. Colores also provided the optimal score and corresponding rotation for each voxel in the cryoEM map. This 3D scoring landscape was further investigated using interactive peak search, as described below. Due to the resolution of the cryoEM map, the top scoring placements provided by the exhaustive search were identified in the high-density regions of the map. Such arrangement of glycoproteins generated an atomic model with major steric clashes and prevented the assembly of the capsomers according to the Gn/Gc ratios estimated by Sherman et al. (2009). Therefore, we further investigated the results of the exhaustive search using interactive exploration techniques (Heyd and Birmanns, 2009) provided by the molecular modeling software Sculptor (Birmanns et al., 2011). This approach permitted us to augment the selection of cross-correlation peaks with expert knowledge such as the Gn/Gc ratio inside the capsomers. Multiple docking locations were thus selected for each type of glycoprotein resulting in several Gn/Gc pairs considered for further modeling steps. Each Gn/Gc pair was subjected to the procedure described in **Figure 2**.

First, the interactively selected placements were employed to create an initial model of the hexon located at the threefold axis. This atomic model, composed of 6xGc and 6xGn units, was

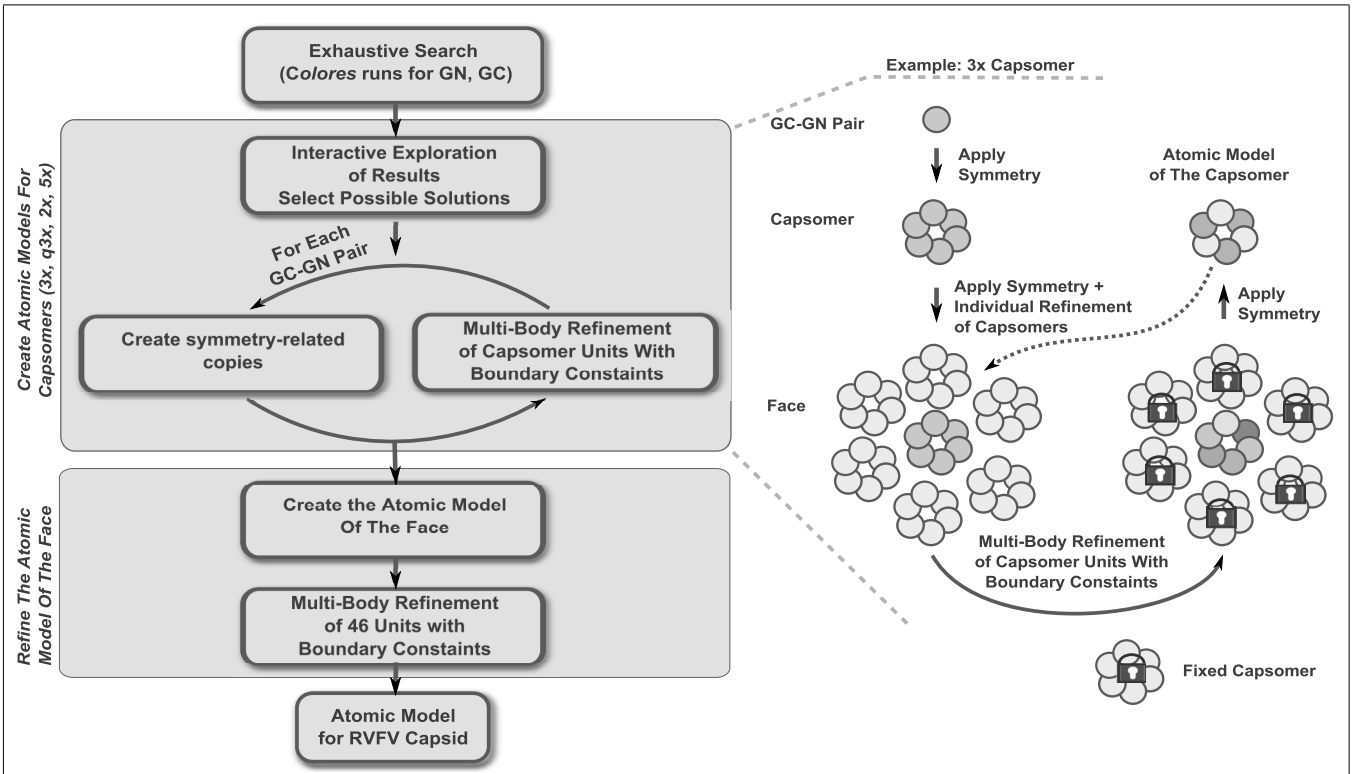


FIGURE 2 | Schematic representation of the modeling steps undertaken to create the atomic model of the RVFV envelope. A detailed description of the individual steps is found in the text.

also placed in the neighboring capsomers. We proceeded with a multi-body Powell refinement analysis of the raw volume of these fragments (as described in Birmanns et al., 2011), while at the same time applying boundary constraints. Such a local optimization simultaneously refines the translation and rotation of each glycoprotein in the capsomer by maximizing the cross-correlation coefficient. As multiple fragments are considered at the same time, the refinement prevents the glycoproteins from overlapping or from causing major steric clashes. The technique permits the introduction of boundary constraints in the form of atomic models describing the neighboring capsomers. Such constraints were not well defined in the first steps of the modeling and therefore the individual glycoproteins building the neighboring capsomers were also considered in the multi-body refinement. As the different types of capsomers were identified, the neighboring capsomers became available and were utilized as constraints in the refinement. No symmetry was technically considered during the refinement, yet the units effectively adopted the symmetry exhibited by the capsomer volume. For example, a threefold symmetry became apparent when refining the B capsomers which are organized around the threefold symmetry axis. The multi-body refinement was iterated several times until the placement of the glycoproteins was stable. As an atomic model was generated for each type of capsomer, a final multi-body refinement was undertaken to create the asymmetric unit. Forty-six units, 23 Gc and 23 Gn glycoproteins, were simultaneously refined while constraining the 15 neighboring capsomers.

INTRA- AND INTER-CAPSOMER PLACEMENT OF RVFV Gn AND Gc

We applied the described procedure (**Figure 2**) to 11xGn/Gc pairs obtained by combining the interactively selected Gc and Gn glycoproteins. Some of these pairs were discarded during the modeling as it become apparent that they prevented the generation of models with good stereochemical quality and appropriate Gn/Gc ratios. At the end of the procedure, four models were produced with cross-correlation coefficients above 0.783 (**Figures A1–A4** in Appendix). The top scoring model had a correlation of 0.798 and is shown in **Figures 3** and **4**. This model had an estimated volume of approx. 1,300,000 Å^3 for the hexon and approximately 1,100,000 Å^3 for the pentons, in agreement with our previous calculations (Sherman et al., 2009).

Although the resolution of the 3D map of RVFV was limited, we were able to derive an assembly model through docking of the molecular Gn and Gc models using an iterative refinement and neighboring constraints (**Figures 3A–C**). In total, four possible arrangements of the glycoproteins in the virion envelope were identified and the predicted arrangement of the two glycoproteins leads to both, homo- and hetero-dimeric contacts between Gn and Gc (**Figures A1–A4** in Appendix).

While our approach generated four possible models for the virion envelope, the organization of the glycoproteins is conserved between these models (**Figures A1–A4** in Appendix). The Gc glycoprotein forms the icosahedral scaffold and remains consistent in the four models. It can be ascribed to the density identified as the viral "skirt" around the base of each capsomer.

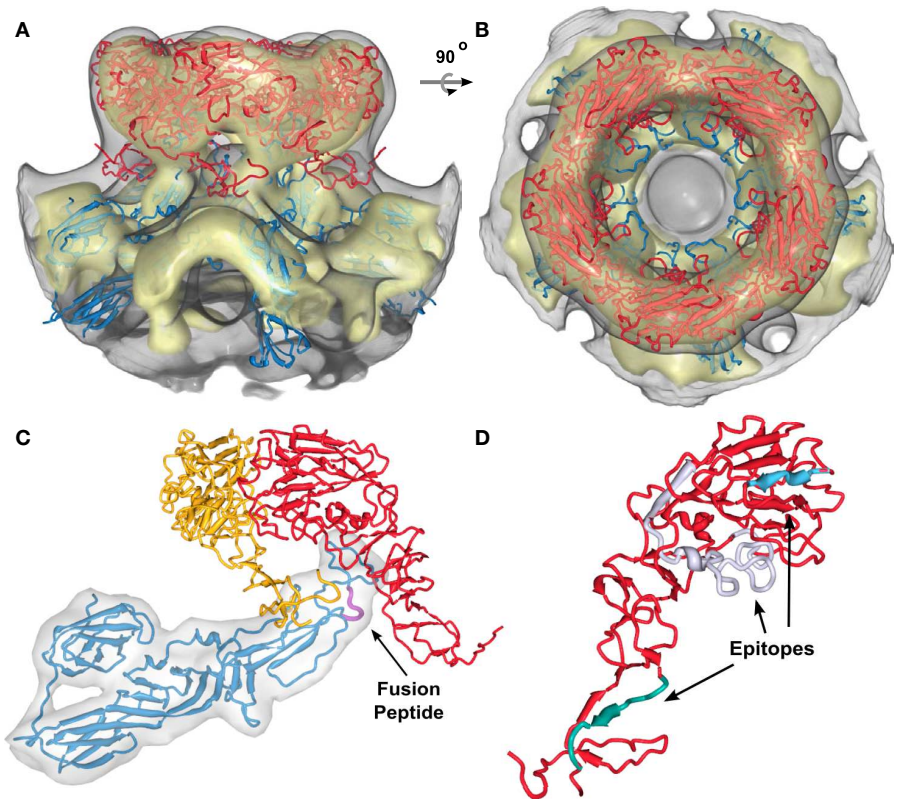


FIGURE 3 | Positioning of the Gn and Gc molecular models into the RVFV cryoEM reconstruction for the top scoring model. (A,B) Show the glycoprotein arrangement within a penton extracted from the cryoEM density. The cryoEM density is represented as a gray transparent capsomer and the glycoprotein monomer models are indicated in red (Gn) and blue (Gc). Gn could only be positioned in the outer caldera of the capsomer and Gc in the skirt region of the capsomer. Two different viewing angles are shown (side-view, and top-view). **(C)** One structural unit (Gn-Gc heterodimer) and an adjacent Gn monomer have been extracted from the docking results shown in **(A)**. Within the basic structural unit, the head domain of the Gn model (red and yellow) covers domain II of Gc. The predicted location of the fusion peptide shown in domain II of Gc is highlighted in magenta and indicated by the black arrow. **(D)** Epitopes for three monoclonal antibodies recognizing Gn (Keegan and Collett, 1986) are highlighted. These epitopes are corresponding to the monoclonal antibodies 4-32-8D (gray), 4-D4 (blue), and 3C-10 (green).

On the other hand, the Gn glycoprotein is placed in the protruding envelope yet has different angles relative to the scaffold. A close investigation of possible placements of Gn allowed us to group our four models into two main classes, in which Gn has a mirrored orientation with roughly ±45° relative to the scaffold (**Figures A1–A4** in Appendix). The two possible placements are a result of the overall fold of the Gn glycoprotein as derived from homology modeling. The large globular domain of Gn drives the glycoprotein in the protruding capsomer, however the C- and N-terminus form a stalk region of reduced dimension, that provides insufficient constraints for the registration and thus the two different orientations. Moreover, the Gn model is incomplete at the C-terminus due to the lack in similarity with known protein structures (which prevented a homology based modeling of the region). Current on-going research in the laboratory is focused on providing experimental data to differentiate between the two potential orientations of Gn reported on the sequence similarity between the Gn proteins from two bunyavirus genera, namely hantaviruses and tospoviruses, with the Sindbis virus E2 protein. However, no significant sequence similarity was detected between the phlebovirus Gn and alphavirus E2 proteins. This might explain why comparison of the structural model for RVFV Gn with the recently solved alphavirus E2 protein structure (Li et al., 2010; Voss et al., 2010) did not reveal any structural similarity. The location of the two RVFV glycoproteins suggested in our model is plausible as Gn fits into the outer density of the capsomers and the model is consistent with the available biological data on RVFV. Keegan and Collett (1986) localized distinct antigenic determinants on the Gn glycoprotein and we chose three of these mapped epitopes and highlighted them in our molecular model for Gn (**Figure 3D**). Two of these epitopes, which are recognized by neutralizing monoclonal antibodies, are surface exposed (highlighted in blue and gray in **Figure 3D**). The epitope recognized by a non-neutralizing and non-protective antibody is located within the predicted stem region of Gn (highlighted in green in **Figure 3D**). In our model for Gn, this region interacts with domain II of Gc and also covers the fusion loop (highlighted in **Figure 3C**). The placement of Gc within the RVFV particle has similarities to that of the alphavirus E2 arrangement (Roussel et al., 2006). Domain II of E2 is the main interacting domain with E1, E2 has a position within the spike with a slight upward orientation on the virion surface

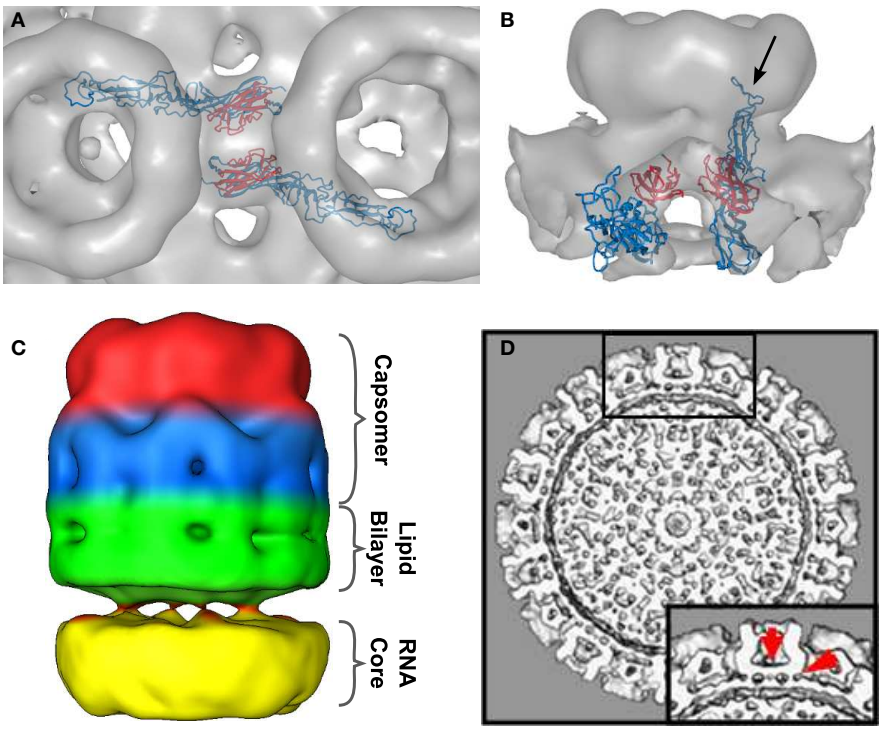


FIGURE 4 | Intercapsomer connections for the top scoring model. **(A)** Top-view of two neighboring capsomers (gray cryoEM density) with two Gc monomers shown in blue. The domain III's (red) are very well positioned within the ridges connecting adjacent capsomers. The fusion peptide is directed to the capsomer center. **(B)** Side-view of one capsomer along the tunnel located beneath the connecting ridges. Two Gcs are shown and their proposed position within the cryoEM density. The black arrow indicates the location of the fusion peptide within domain II. The domain IIIs are highlighted in red to indicate their placement within the ridges. **(C)** CryoEM density of one extracted penton at a very low threshold (0.54). The outer region of the capsomer is indicated in red (representing mainly Gn molecules), the capsomer base in blue (representing mainly Gc molecules), the lipid envelope in green, and the density corresponding to the RNP core is shown in yellow. Densities spanning the gap between the lipid bilayer and the RNP core are representing the glycoprotein cytoplasmic tails. **(D)** Surface-shaded representation of the central section of the RVFV cryoEM map viewed along the fivefold orientation. The sections show glycoprotein protrusions on virus surface, lipid bilayer, and RNP core. In the lower right corner a blow-up of the boxed area is shown. Red arrows point to clearly defined densities spanning the lipid bilayer. These densities represent glycoprotein transmembrane domains and are located either on the outer edge of the capsomer or directly beneath the connecting channels.

and also forms the skirt of the spike (Li et al., 2010; Voss et al., 2010).

In the cryoEM reconstructions of RVFV, a strong density bridging neighboring capsomers has been described (Freiberg et al., 2008; Huiskonen et al., 2009; Sherman et al., 2009). These ridges are located halfway between the rim of the capsomer and the lipid bilayer of the virion. Inside these ridges a channel approximately 18 Å in diameter runs between adjacent capsomers and interconnects the inner cavities of the neighboring capsomers. In our model of the glycoprotein arrangement, Gc can be placed into the dense region of these ridges (**Figure 4A**). Specifically, the domain III of two Gc molecules from adjacent capsomers filled the density (highlighted in red in **Figures 4A,B**). In the side-view of the structure, one can clearly see how domain III forms the tunnel-like structure (**Figure 4B**). Further, the position of the fusion peptide oriented to the capsomer center is displayed (arrow in **Figure 4B**).

A similar model for the RVFV envelope was also obtained when building the Gc glycoprotein structure based on that of the Chikungunya virus E1 protein (Voss et al., 2010; data not shown). Again, Gn forms the protrusion spikes of the capsomers, while Gc is the main component of the icosahedral scaffold. Similarly, the domain III of Gc is the main component of the ridges between the capsomers. However, in this model the stem-like region of Gn is partially involved in the formation of the ridges as well (data not shown). Unlike in the previous model, in this model the fusion peptide located within Gc, points more outward from the capsomer but is still covered by the Gn glycoprotein.

DISCUSSION

The family *Bunyaviridae*, the largest RNA virus family with more than 350 named isolates, is organized into five genera based upon genetic and antigenic differences (Elliott, 2009). While many studies have focused on molecular aspects of transcription, replication, pathogenesis, and vaccine development, little is known about the structural organization and physical interactions of bunyavirus glycoproteins within the virion. Recently, cryoEM structures have been solved for the phleboviruses RVFV (Freiberg et al., 2008; Huiskonen et al., 2009; Sherman et al., 2009) and Uukuniemi virus (Overby et al., 2008), and the hantaviruses Tula (Huiskonen et al., 2010) and Hantaan viruses (Battisti et al., 2011). These

structures did not only increase our basic knowledge regarding the assembly of the member viruses of this important virus family but also revealed that the bunyavirus glycoproteins can occur in multiple arrangements. While phlebovirus glycoproteins are arranged on the virion surface in $T = 12$ icosahedral symmetry, the hantavirus glycoproteins are arranged in a grid-like pattern. It is possible that the size of the glycoprotein molecules and the number of their TMD are factors contributing to the different arrangement of the glycoproteins on the surface of the member viruses of the various genera. However, due to the lack of an experimentally proven structure for any entire bunyavirus glycoprotein, we applied fold recognition structure prediction to generate 3D structural models for the RVFV Gn and Gc ectodomain monomers. The glycoprotein structures have been further analyzed in combination with the RVFV cryoEM structure previously solved by our group and others. Identifying the organization of the glycoproteins in the cryoEM envelope was achieved by using a modeling framework involving global and constrained local search. This framework was developed for RVFV, yet it may be applied to other multi-component assemblies.

HYPOTHETICAL ASSEMBLY MODEL FOR RIFT VALLEY FEVER VIRUS

The two RVFV glycoproteins, Gn and Gc, are organized in 122 distinct capsomers on the virion surface, extending ∼96 Å above the lipid envelope. Our docking framework (**Figure 2**) allowed the identification of four potential arrangements of the glycoproteins Gn and Gc within the virion envelope (**Figures A1–A4** in Appendix). These models are mainly intended to represent a starting point for future research in analyzing the overall architecture of the phlebovirus envelope, as well as the virion assembly and fusion process. While we are aware of the fact that the described interactions between Gn and Gc homology models cannot be used to draw detailed conclusions at the molecular level, we can make the statement that Gn-Gc heterodimers form the basic structural unit in the capsomers in each of our four models. We hypothesize that hexons and pentons are comprised of six and five Gn-Gc heterodimers, respectively, with Gn being more solvent exposed and forming the capsomer spike and the Gc protein lying partially underneath, closer to the lipid membrane and forming the capsomer base. This arrangement is likely, since neutralizing monoclonal antibodies against both Gn and Gc have been described (Besselaar and Blackburn, 1991). In addition to interactions between Gn and Gc within each heterodimer, there are also interactions between neighboring structural units. A Gc molecule from one heterodimer contacts the stalk region of an adjacent Gn molecule, which is part of the neighboring heterodimer (**Figure 3C**). A recent study has shown that hantavirus glycoproteins form complex intra- and inter-molecular disulfide bonds between Gn and Gc, which contributes to the assembly and stability of the virus particle (Hepojoki et al., 2010). The RVFV Gn and Gc ectodomains used for our molecular modeling have 23 and 20 cysteines, respectively, and it is possible that similar inter- and intra-molecular disulfide bonds are present as well.

For our generated molecular models, we found significant structural matches between the RVFV Gn and the receptor binding domain of the Influenza virus hemagglutinin protein, and a separate match between the RVFV Gc protein and the alphavirus E1 protein. Since earlier bioinformatic investigation of the bunyavirus Gc protein has already predicted it to be a class II viral fusion protein (Garry and Garry, 2004), our findings for RVFV Gc were expected.

The alphavirus spike complex consists of a trimer of heterodimers $[(E1\text{-}E2)_3]$ and is mediated by interactions between E2 and E1 TMDs (Lescar et al., 2001; Pletnev et al., 2001). Even though we did not include the glycoprotein TMD and CTD in our fold predictions, it is possible that the Gn and Gc proteins interact with each other via their transmembrane regions and that the glycoproteins interact with the ribonucleoprotein complex via their Gn/Gc cytoplasmic tails. The interaction of the TMDs may represent an additional determinant in the heterodimer assembly. This hypothesis is strengthened by our description of protein densities spanning the space between the RNP core and the lipid bilayer within the RVFV particle (Sherman et al., 2009; **Figure 4C**). A recently published study by Piper et al. (2011) described the requirement of the RVFV Gn protein for genome packaging and showed that the Gn cytoplasmic tail is necessary for this process. In our RVFV cryoEM reconstruction we noticed the presence of densities spanning the virus envelope at the positions of capsomers (Sherman et al., 2009). These densities most likely represent the Gn and Gc TMDs and seem to be situated directly at the center of the ridges between neighboring capsomers and at the outer edges of the capsomers (red arrows in **Figure 4D**).

In contrast to many other lipid enveloped RNA viruses, bunyaviruses do not contain a matrix protein that has the function of linking and stabilizing the nucleocapsid and viral envelope proteins. Based on our model, we suggest that a highly organized arrangement of the Gn and Gc glycoprotein ectodomains is responsible for overall virion stability and that the capsomer–capsomer interactions play a central role in defining the icosahedral virion symmetry.

Multiple monoclonal antibodies against RVFV Gn and Gc have been described (Besselaar and Blackburn, 1991, 1994) and the epitopes on the ectodomain of Gn have been mapped (Keegan and Collett, 1986). In our model, the epitopes for the monoclonal antibodies 4-D4 and 4-32-8D, which have neutralizing and protective functions, are localized and surface-exposed in the globular head domain of Gn (**Figure 3D**). This domain caps Gc domain II and fusion loop and it may be that the neutralizing effect of these two antibodies is explained by either preventing receptor binding or potential rearrangement of Gn post-receptor attachment and, hence, inhibition of fusion, since the fusion loop will not be exposed to the host membrane. The epitope recognized by another monoclonal antibody, 3C-10, which has been described as non-neutralizing and non-protective in the mouse model has been localized in the stalk region of the Gn model (**Figure 3D**). In our model, this region can be found to be localized close to the ridges, connecting adjacent capsomers (**Figure 3A**). It is possible that this epitope is not freely accessible in the native conformation within the virion. The Gc domain III forming the capsomer connections may represent a steric block preventing antibody binding.

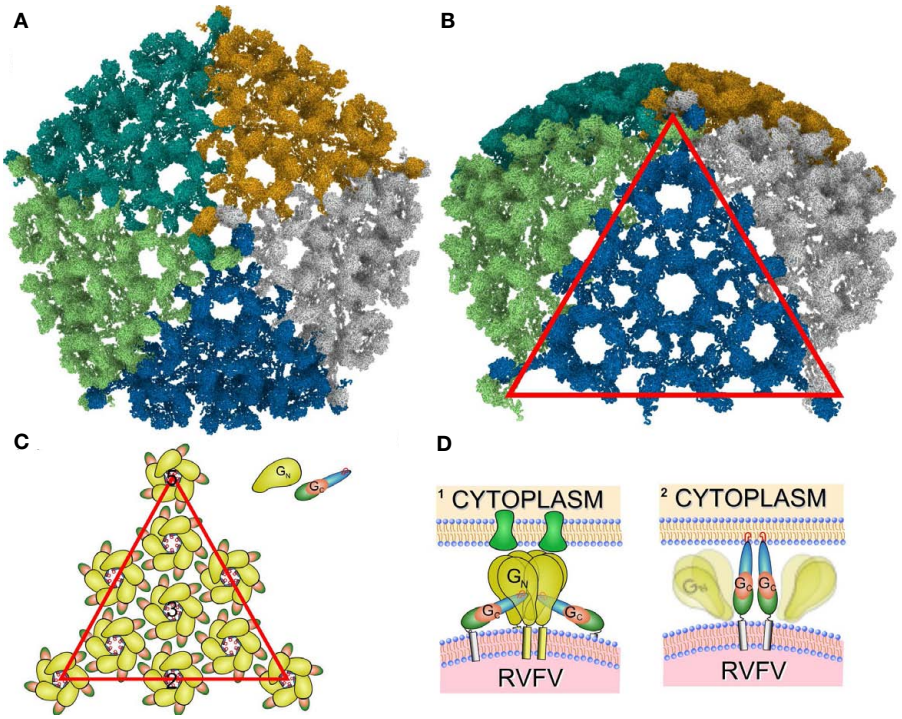


FIGURE 5 | Overview of the RVFV glycoprotein shell. (A) The proposed $T = 12$ icosahedral protein layer formed by Gn and Gc. Individual subunits are color coded. **(B)** Tilted representation as shown in **(A)**. The red triangle represents one triangular face. **(C)** Schematic representation of the Gn and Gc contacts. Drawn is one of the 20 triangular faces of the icosahedrons enclosing the RVFV particle and the distribution of the Gn and Gc glycoproteins [corresponding to red triangle in **(B)**]. Black numbers denote icosahedral two-, three-, and fivefold symmetry axes. Gn monomers are represented as bulb-like structures in yellow, and Gc monomers as a tube-like structure. The individual domains are represented in red (domain I), blue (domain II), and green (domain III). The fusion peptides are indicated as red circles, and are pointing to the capsomer center. **(D)** Hypothetical model of the RVFV – host cell interaction. The RVFV glycoproteins Gn and Gc are represented according to our model and show similarities to the alphavirus E1 and E2 proteins. (1) Gn is depicted as the receptor binding protein and binds to the host cell receptor (green). (2) After receptor binding the uptake of the RVFV particle is initiated and an acidification step of the endocytic vesicle triggers the dissociation of Gn and Gc. This results in the formation of potentially Gc trimers (in accordance with current models for class II fusion proteins) and insertion of the fusion peptides into the host cell membrane.

In conclusion, structural models have been developed for the RVFV glycoproteins, Gn and Gc. The structural aspects of these protein models allowed us to generate four putative assembly models indicating how Gn and Gc may interact within and between capsomers. The top scoring model (as indicated by the highest cross-correlation coefficient) for the icosahedral shell of RVFV is presented in **Figure 5**. Our model has certain similarity to the described assembly model of alphaviruses, in terms of the fact that in bunyavirus surface proteins the receptor binding and membrane fusion activities most likely reside in two different glycoproteins (similar to the E1 and E2 glycoproteins in alphaviruses). However, while the alphavirus spike is formed by trimers of E1/E2 heterodimers, RVFV Gn/Gc heterodimers are organized in pentameric and hexameric capsomers. In flaviviruses, the E protein is responsible for both receptor binding and fusion. Further, the fusion peptide of the RVFV Gc protein sticks up and is oriented against Gn, similar to the findings for the alphavirus E1 and E2 proteins, whereas in the flaviviruses the fusion peptides are held down and are oriented against the interface of the E protein domain I and III.

The presented arrangement of Gn and Gc and description of their interactions may play an important role in glycoprotein folding and maturation, capsomer and virus assembly, virus fusion, and neutralization of infection. On-going site-directed mutagenesis experiments using a reverse-genetics system (Ikegami et al., 2005) are currently being used to evaluate the proposed glycoprotein interactions. The new information reported in this study, will not only impact our understanding of the assembly of phleboviruses and other bunyaviruses, but may also be exploited in furthering our understanding of the complex antigenic interactions of the many member viruses of the family *Bunyaviridae*. Such structural studies are hoped also to contribute to the design of effective antivirals.

ACKNOWLEDGMENTS

We thank Drs. Alan Barrett and Fred Murphy for helpful comments and discussions. This work was supported by a training fellowship from the W. M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Virus Imaging (Alexander N. Freiberg), and in part by a grant from the National Institutes of Health (R01GM62968, Willy Wriggers).

REFERENCES

Alam, S. L., Langelier, C., Whitby, F. G., Koirala, S., Robinson, H., Hill, C. P., and Sundquist, W. I. (2006). Structural basis for ubiquitin recognition by the human ESCRT-II EAP45 GLUE domain. *Nat. Struct. Mol. Biol.* 13, 1029–1030.

Battisti, A. J., Chu, Y. K., Chipman, P. R., Kaufmann, B., Jonsson, C. B., and Rossmann, M. G. (2011). Structural studies of Hantaan virus. *J. Virol.* 85, 835–841.

Besselaar, T. G., and Blackburn, N. K. (1991). Topological mapping of antigenic sites on the Rift Valley fever virus envelope glycoproteins using monoclonal antibodies. *Arch. Virol.* 121, 111–124.

Besselaar, T. G., and Blackburn, N. K. (1994). The effect of neutralizing monoclonal antibodies on early events in Rift Valley fever virus infectivity. *Res. Virol.* 145, 13–19.

Birmanns, S., Rusu, M., and Wriggers, W. (2011). Using Sculptor and Situs for simultaneous assembly of atomic components into low-resolution shapes. *J. Struct. Biol.* 173, 428–435.

Cole, C., Barber, J. D., and Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* 36, W197–W201.

Collett, M. S., Purchio, A. F., Keegan, K., Frazier, S., Hays, W., Anderson, D. K., Parker, M. D., Schmaljohn, C., Schmidt, J., and Dalrymple, J. M. (1985). Complete nucleotide sequence of the M RNA segment of Rift Valley fever virus. *Virology* 144, 228–245.

Elliott, R. M. (2009). Bunyaviruses and climate change. *Clin. Microbiol. Infect.* 15, 510–517.

Estrada, D. F., Boudreaux, D. M., Zhong, D., St Jeor, S. C., and De Guzman, R. N. (2009). The hantavirus glycoprotein G1 tail contains dual CCHC-type classical zinc fingers. *J. Biol. Chem.* 284, 8654–8660.

Estrada, D. F., Conner, M., Jeor, S. C., and Guzman, R. N. (2011). The structure of the hantavirus zinc finger domain is conserved and represents the only natively folded region of the Gn cytoplasmic tail. *Front. Microbiol.* 2:251. doi:10.3389/fmicb.2011.00251

Estrada, D. F., and De Guzman, R. N. (2011). Structural characterization of the Crimean-Congo hemorrhagic fever virus Gn tail provides insight into virus assembly. *J. Biol. Chem.* 286, 21678–21686.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* Chap. 5, Unit 5.6.

Freiberg, A. N., Sherman, M. B., Morais, M. C., Holbrook, M. R., and Watowich, S. J. (2008). Three-dimensional organization of Rift Valley fever virus revealed by cryoelectron tomography. *J. Virol.* 82, 10341–10348.

Garry, C. E., and Garry, R. F. (2004). Proteomics computational analyses suggest that the carboxyl terminal glycoproteins of Bunyaviruses are class II viral fusion protein (beta-penetrenes). *Theor. Biol. Med. Model.* 1, 10.

Garry, C. E., and Garry, R. F. (2008). Proteomics computational analyses suggest that baculovirus GP64 superfamily proteins are class III penetrenes. *Virol. J.* 5, 28.

Garry, C. E., and Garry, R. F. (2009). Proteomics computational analyses suggest that the bornavirus glycoprotein is a class III viral fusion protein (gamma penetrene). *Virol. J.* 6, 145.

Gerrard, S. R., and Nichol, S. T. (2002). Characterization of the golgi retention motif of Rift Valley fever virus G(N) glycoprotein. *J. Virol.* 76, 12200–12210.

Gibbons, D. L., Vaney, M. C., Roussel, A., Vigouroux, A., Reilly, B., Lepault, J., Kielian, M., and Rey, F. A. (2004). Conformational change and protein–protein interactions of the fusion protein of Semliki Forest virus. *Nature* 427, 320–325.

Hepojoki, J., Strandin, T., Vaheri, A., and Lankinen, H. (2010). Interactions and oligomerization of hantavirus glycoproteins. *J. Virol.* 84, 227–242.

Heyd, J., and Birmanns, S. (2009). Immersive structural biology: a new approach to hybrid modeling of macromolecular assemblies. *Virtual Real.* 13, 245–255.

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14, 378–379.

Huiskonen, J. T., Hepojoki, J., Laurinmaki, P., Vaheri, A., Lankinen, H., Butcher, S. J., and Grunewald, K. (2010). Electron cryotomography of Tula hantavirus suggests a unique assembly paradigm for enveloped viruses. *J. Virol.* 84, 4889–4897.

Huiskonen, J. T., Overby, A. K., Weber, F., and Grunewald, K. (2009). Electron cryo-microscopy and single-particle averaging of Rift Valley fever virus: evidence for GN-GC glycoprotein heterodimers. *J. Virol.* 83, 3762–3769.

Ikegami, T., Won, S., Peters, C. J., and Makino, S. (2005). Rift Valley fever virus NSs mRNA is transcribed from an incoming anti-viral-sense S RNA segment. *J. Virol.* 79, 12106–12111.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., and Godzik, A. (2005). FFAS03: a server for profile-profile sequence alignments. *Nucleic Acids Res.* 33, W284–W288.

Kajan, L., and Rychlewski, L. (2007). Evaluation of 3D-Jury on CASP7 models. *BMC Bioinformatics* 8, 304. doi:10.1186/1471-2105-8-304

Kakach, L. T., Suzich, J. A., and Collett, M. S. (1989). Rift Valley fever virus M segment: phlebovirus expression strategy and protein glycosylation. *Virology* 170, 505–510.

Keegan, K., and Collett, M. S. (1986). Use of bacterial expression cloning to define the amino acid sequences of antigenic determinants on the G2 glycoprotein of Rift Valley fever virus. *J. Virol.* 58, 263–270.

Kielian, M. (2006). Class II virus membrane fusion proteins. *Virology* 344, 38–47.

Kielian, M., and Rey, F. A. (2006). Virus membrane-fusion proteins: more than one way to make a hairpin. *Nat. Rev. Microbiol.* 4, 67–76.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580.

Lee, M. S., Lebeda, F. J., and Olson, M. A. (2009). Fold prediction of VP24 protein of Ebola and Marburg viruses using de novo fragment assembly. *J. Struct. Biol.* 167, 136–144.

Lee, S. S., Knott, V., Jovanovic, J., Harlos, K., Grimes, J. M., Choulier, L., Mardon, H. J., Stuart, D. I., and Handford, P. A. (2004). Structure of the integrin binding fragment from fibrillin-1 gives new insights into microfibril organization. *Structure* 12, 717–729.

Lescar, J., Roussel, A., Wien, M. W., Navaza, J., Fuller, S. D., Wengler, G., and Rey, F. A. (2001). The fusion glycoprotein shell of Semliki Forest virus: an icosahedral assembly primed for fusogenic activation at endosomal pH. *Cell* 105, 137–148.

Li, L., Jose, J., Xiang, Y., Kuhn, R. J., and Rossmann, M. G. (2010). Structural changes of envelope proteins during alphavirus fusion. *Nature* 468, 705–708.

Lozach, P. Y., Kuhbacher, A., Meier, R., Mancini, R., Bitto, D., Bouloy, M., and Helenius, A. (2011). DC-SIGN as a receptor for phleboviruses. *Cell Host Microbe* 10, 75–88.

Lozach, P. Y., Mancini, R., Bitto, D., Meier, R., Oestereich, L., Overby, A. K., Pettersson, R. F., and Helenius, A. (2010). Entry of bunyaviruses into mammalian cells. *Cell Host Microbe* 7, 488–499.

MMWR. (2007). Rift Valley fever outbreak – Kenya, November 2006-January 2007. *MMWR Morb. Mortal. Wkly. Rep.* 56, 73–76.

Overby, A. K., Pettersson, R. F., Grunewald, K., and Huiskonen, J. T. (2008). Insights into bunyavirus architecture from electron cryotomography of Uukuniemi virus. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2375–2379.

Overby, A. K., Popov, V. L., Pettersson, R. F., and Neve, E. P. (2007). The cytoplasmic tails of Uukuniemi virus (Bunyaviridae) G(N) and G(C) glycoproteins are important for intracellular targeting and the budding of virus-like particles. *J. Virol.* 81, 11381–11391.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612.

Piper, M. E., Sorenson, D. R., and Gerrard, S. R. (2011). Efficient cellular release of Rift Valley fever virus requires genomic RNA. *PLoS ONE* 6, e18070. doi:10.1371/journal.pone.0018070

Plassmeyer, M. L., Soldan, S. S., Stachelek, K. M., Martin-Garcia, J., and Gonzalez-Scarano, F. (2005). California serogroup Gc (G1) glycoprotein is the principal determinant of pH-dependent cell fusion and entry. *Virology* 338, 121–132.

Plassmeyer, M. L., Soldan, S. S., Stachelek, K. M., Roth, S. M., Martin-Garcia, J., and Gonzalez-Scarano, F. (2007). Mutagenesis of the La Crosse virus glycoprotein supports a role for Gc (1066–1087) as the fusion peptide. *Virology* 358, 273–282.

Pletnev, S. V., Zhang, W., Mukhopadhyay, S., Fisher, B. R., Hernandez, R., Brown, D. T., Baker, T. S., Rossmann, M. G., and Kuhn, R. J. (2001). Locations of carbohydrate sites on alphavirus glycoproteins show that E1 forms an icosahedral scaffold. *Cell* 105, 127–136.

Ramachandran, G. N., and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem.* 23, 283–438.

Roussel, A., Lescar, J., Vaney, M. C., Wengler, G., and Rey, F. A. (2006). Structure and interactions at the viral surface of the envelope protein E1 of Semliki forest virus. *Structure* 14, 75–86.

Sherman, M. B., Freiberg, A. N., Holbrook, M. R., and Watowich, S. J. (2009). Single-particle cryo-electron microscopy of Rift Valley fever virus. *Virology* 387, 11–15.

Shi, X., Goli, J., Clark, G., Brauburger, K., and Elliott, R. M. (2009). Functional analysis of the Bunyamwera orthobunyavirus Gc glycoprotein. *J. Gen. Virol.* 90, 2483–2492.

Soldan, S. S., Hollidge, B. S., Wagner, V., Weber, F., and Gonzalez-Scarano, F. (2010). La Crosse virus (LACV) Gc fusion peptide mutants have impaired growth and fusion phenotypes, but remain neurotoxic. *Virology* 404, 139–147.

Stevens, J., Corper, A. L., Basler, C. F., Taubenberger, J. K., Palese, P., and Wilson, I. A. (2004). Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus. *Science* 303, 1866–1870.

Tischler, N. D., Gonzalez, A., Perez-Acle, T., Rosemblatt, M., and Valenzuela, P. D. (2005). Hantavirus Gc glycoprotein: evidence for a class II fusion protein. *J. Gen. Virol.* 86, 2937–2947.

Tusnady, G. E., and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.* 283, 489–506.

Voss, J. E., Vaney, M. C., Duquerroy, S., Vonrhein, C., Girard-Blanc, C., Crublet, E., Thompson, A., Bricogne, G., and Rey, F. A. (2010). Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature* 468, 709–712.

Wasmoen, T. L., Kakach, L. T., and Collett, M. S. (1988). Rift Valley fever virus M segment: cellular localization of M segment-encoded proteins. *Virology* 166, 275–280.

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., and Wishart, D. S. (2003). VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* 31, 3316–3319.

Wriggers, W. (2010). Using Situs for the integration of multi-resolution structures. *Biophys. Rev.* 2, 21–27.

Zhang, W., Chipman, P. R., Corver, J., Johnson, P. R., Zhang, Y., Mukhopadhyay, S., Baker, T. S., Strauss, J. H., Rossmann, M. G., and Kuhn, R. J. (2003). Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. *Nat. Struct. Biol.* 10, 907–912.

Zhang, W., Mukhopadhyay, S., Pletnev, S. V., Baker, T. S., Kuhn, R. J., and Rossmann, M. G. (2002). Placement of the structural proteins in Sindbis virus. *J. Virol.* 76, 11645–11658.

Conflict of Interest Statement: 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.

Received: 25 April 2012; accepted: 29 June 2012; published online: 19 July 2012.

Citation: Rusu M, Bonneau R, Holbrook MR, Watowich SJ, Birmanns S, Wriggers W and Freiberg AN (2012) An assembly model of Rift Valley fever virus. Front. Microbio. ***3****:254. doi: 10.3389/fmicb.2012.00254*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

APPENDIX

CAPSOMER ARCHITECTURE

Between the four models, the overall architecture is preserved, with RVFV Gc forming the scaffold of the capsomers (blue and green ribbon representation in **Figures A1–A4**) and RVFV Gn being localized in the protruding envelope (red and yellow ribbon representation in **Figures A1–A4**). A close investigation of the different models revealed that the angle of the Gn monomers relative to the scaffold is different between the four models. The cross correlation coefficient was estimated for each model relative to the entire envelope. In order to construct the model of the entire envelope, 60 copies of the asymmetric unit were placed according to the icosahedral symmetry. The cross correlation coefficients were 0.798, 0.790, 0.785, and 0.783, for the first, second, third, and fourth top-scoring model.

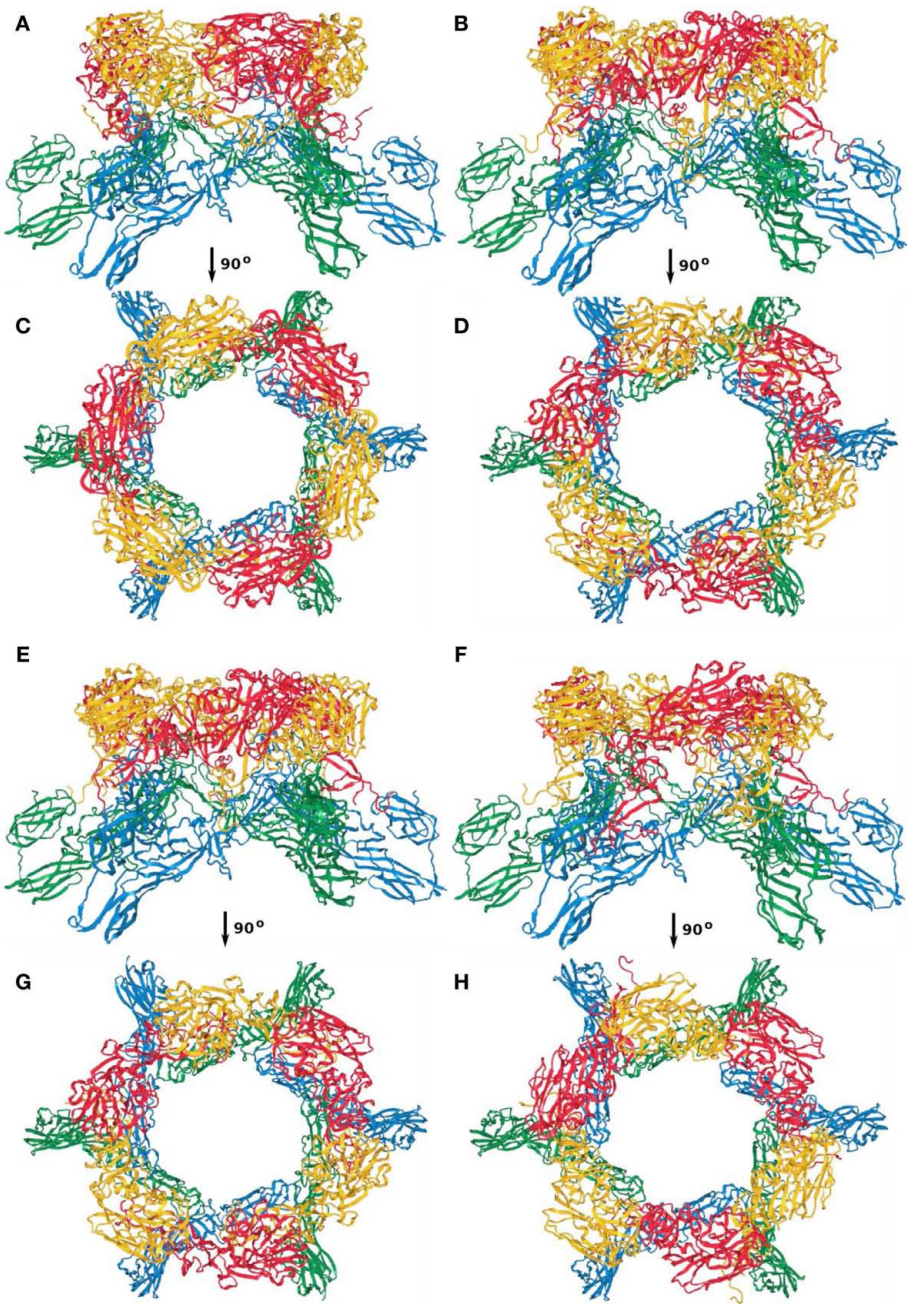


FIGURE A1 | Capsomer twofold axis; (A,C) first; (B,D) second; (E,G) third; (F,H) fourth top-scoring model.

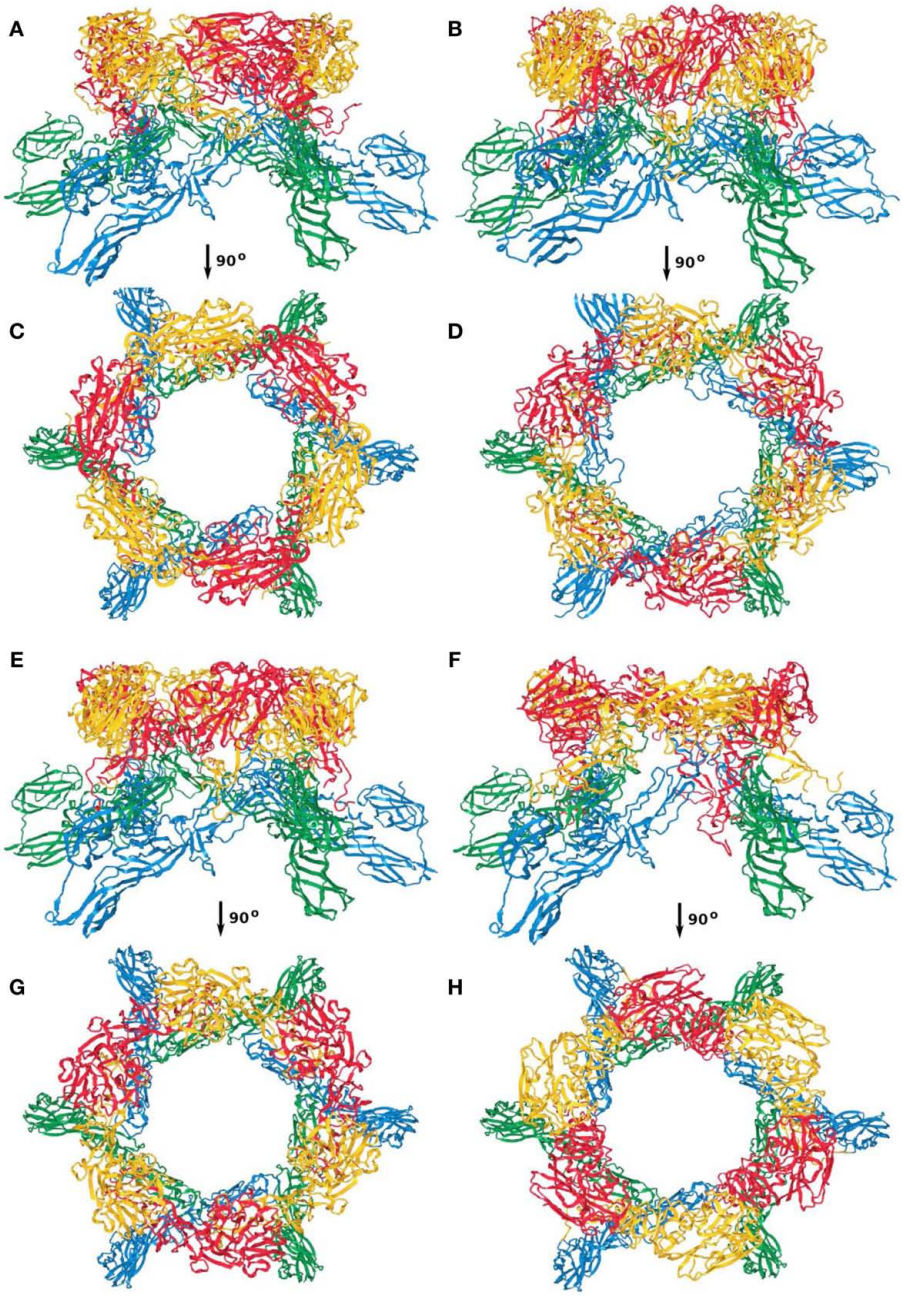


FIGURE A2 | Capsomer threefold axis; (A,C) first; (B,D) second; (E,G) third; (F,H) fourth top-scoring model.

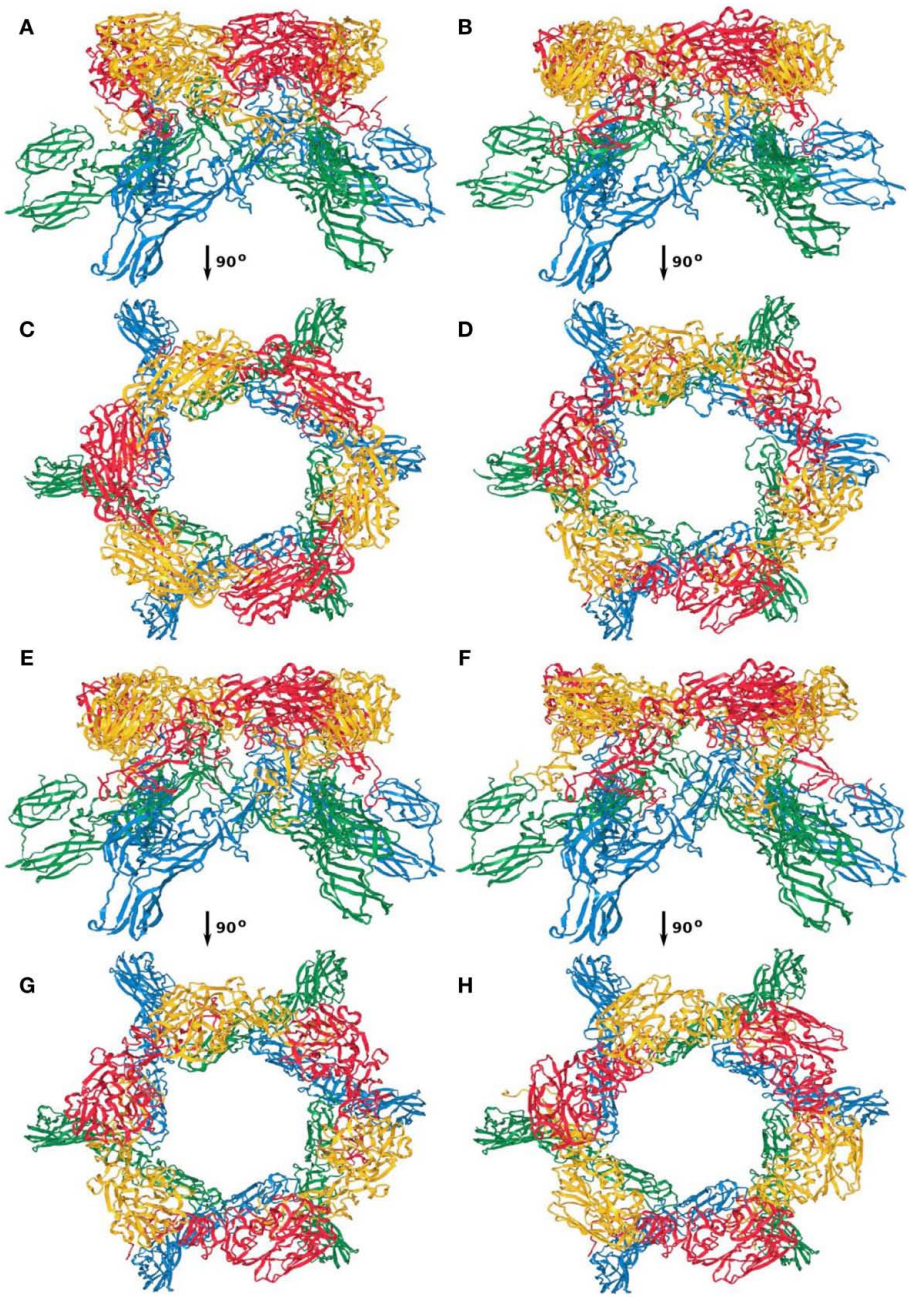


FIGURE A3 | Capsomer quasi-threefold axis; (A,C) first; (B,D) second; (E,G) third; (F,H) fourth top-scoring model.

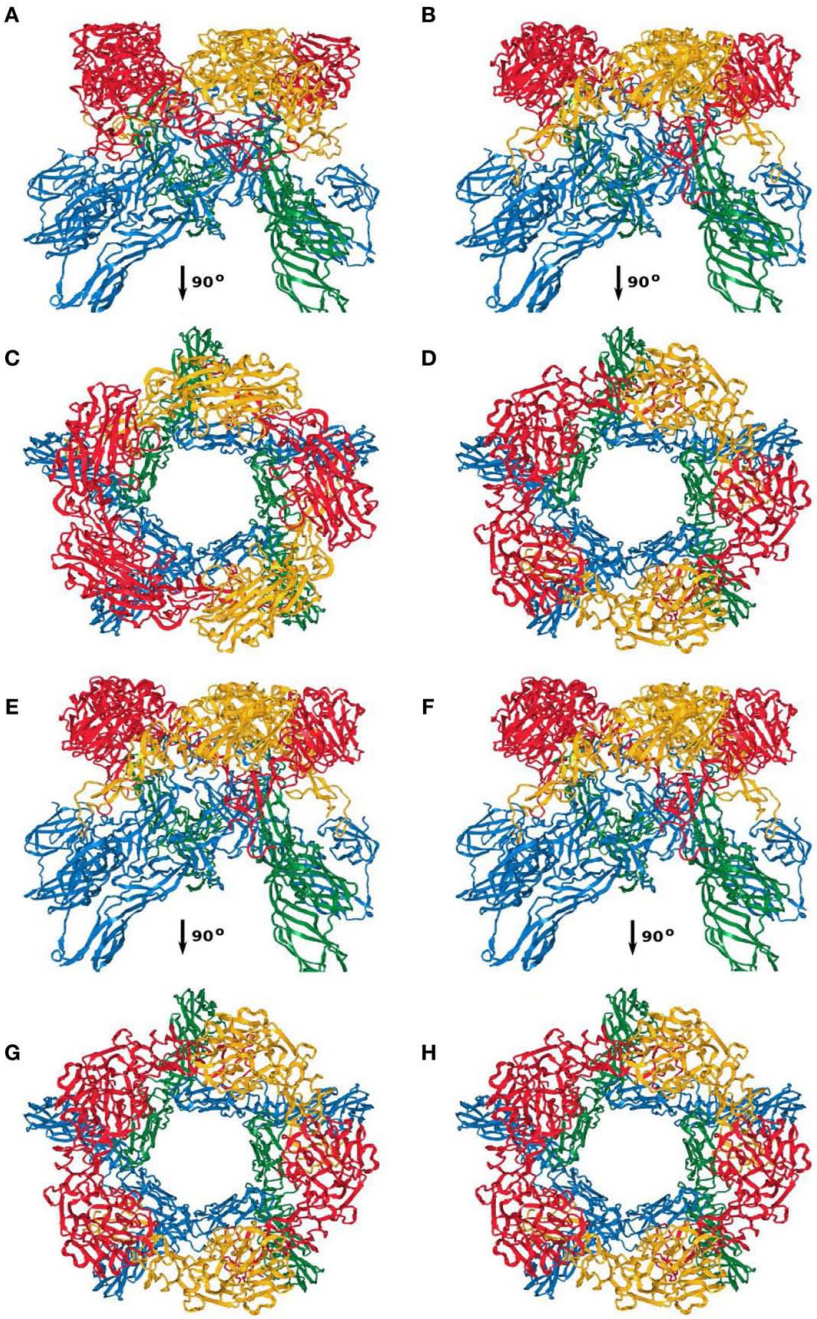


FIGURE A4 | Capsomer fivefold axis; (A,C) first; (B,D) second; (E,G) third; (F,H) fourth top-scoring model.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 05 September 2012
doi: 10.3389/fmicb.2012.00312

Structural basis for specific recognition of substrates by sapovirus protease

Masaru Yokoyama[1‡], Tomoichiro Oka[2,3‡], Hirotatsu Kojima[4], Tetsuo Nagano[4], Takayoshi Okabe[4], Kazuhiko Katayama[2], Takaji Wakita[2], Tadahito Kanda[1†] and Hironori Sato[1]*

[1] *Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo, Japan*
[2] *Department of Virology II, National Institute of Infectious Diseases, Tokyo, Japan*
[3] *Food Animal Health Research Program, Ohio Agricultural Research and Development Center, Department of Veterinary Preventive Medicine, The Ohio State University, Wooster, OH, USA*
[4] *Open Innovation Center for Drug Discovery, The University of Tokyo, Japan*

Edited by:
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan

Reviewed by:
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan
Ming Tan, Cincinnati Children's Hospital Medical Center, USA

***Correspondence:**
Hironori Sato, Pathogen Genomics Center, National Institute of Infectious Diseases, 4-7-1 Gakuen, MusashiMurayama-shi, Tokyo 208-0011, Japan.
e-mail: hirosato@nih.go.jp

†***Present address:***
Tadahito Kanda, The Center of Research Network for Infectious Diseases, RIKEN, Tokyo, Japan.

‡*Masaru Yokoyama and Tomoichiro Oka have contributed equally to this work.*

Sapovirus (SaV) protease catalyzes cleavage of the peptide bonds at six sites of a viral polyprotein for the viral replication and maturation. However, the mechanisms by which the protease recognizes the distinct sequences of the six cleavage sites remain poorly understood. Here we examined this issue by computational and experimental approaches. A structural modeling and docking study disclosed two small clefts on the SaV protease cavity that allow the stable and functional binding of substrates to the catalytic cavity via aromatic stacking and electrostatic interactions. An information entropy study and a site-directed mutagenesis study consistently suggested variability of the two clefts under functional constraints. Using this information, we identified three chemical compounds that had structural and spatial features resembling those of the substrate amino acid residues bound to the two clefts and that exhibited an inhibitory effect on SaV protease *in vitro*. These results suggest that the two clefts provide structural base points to realize the functional binding of various substrates.

Keywords: sapovirus protease, substrate recognition, P1 and P4 amino acid residues, 3-D models, amino acid diversity, mutagenesis, 3-D pharmacophore, inhibitor screening

INTRODUCTION

Sapovirus (SaV) is a non-enveloped RNA virus that belongs to the family Caliciviridae and causes gastroenteritis in humans and swine (Chiba et al., 1979, 2000; Guo et al., 1999; Hansman et al., 2007). The SaV genome is a single-stranded RNA that encodes two or three open reading frames (ORFs; Liu et al., 1995; Noel et al., 1997; Numata et al., 1997; Guo et al., 1999; Robinson et al., 2002). The ORF1 encodes six non-structural proteins (NS1, NS2, NS3, NS4, NS5, and NS6-7) and a structural protein, the capsid protein (VP1; Oka et al., 2006, 2009). The NS6-7 protein contains the chymotrypsin-like protease domain (the 3C-like protease; Oka et al., 2005a,b, 2007; Robel et al., 2008) and the RNA-dependent RNA polymerase domain (the 3-D-like polymerase; Fullerton et al., 2007; Bull et al., 2011). The ORF1 precursor protein is post-translationally cleaved at six sites by the 3C-like protease (Oka et al., 2005b, 2006).

The SaV 3C-like protease domain comprises 146 amino acid residues (Oka et al., 2007). This enzyme cleaves the peptide bonds of specific dipeptides, such as the glutamic acid/glycine (E/G), glutamine/glycine (Q/G), and glutamic acid/alanine (E/A; Oka et al., 2006). However, these dipeptide motifs exist in the non-cleaved sites of the ORF1 polyprotein, indicating that additional amino acid residues are required for the specific recognition of substrates (Oka et al., 2006). In this regard, calicivirus proteases have a large cavity that can accommodate substrate peptides with several amino acid in lengths (Nakamura et al., 2005; Zeitler et al., 2006; Oka et al., 2007). It is conceivable that these substrate amino acid residues around the cleavage sites, termed the P4, P3, P2, P1, P1′, P2′, P3′, and P4′ sites, are all involved to some extent, either directly or indirectly, in the recognition and cleavage by protease. However, there must be a division of roles: previous studies on the calicivirus proteases consistently suggest more extensive involvement of the substrate amino acid residues upstream of the peptide bond of the cleavage sites in the cleavage by proteases (Wirblich et al., 1995; Sosnovtsev et al., 1998; Hardy et al., 2002; Belliot et al., 2003; Scheffler et al., 2007; Robel et al., 2008). In the case of SaV, the substrate P1 and P4 amino acid residues in particular are physicochemically more conserved among different SaV strains (Oka et al., 2009) and more sensitive to the substitutions (Robel et al., 2008; Oka et al., 2009). Therefore, these amino acid residues may provide the specific contact sites with SaV protease. However, due to the lack of structural

information on SaV protease and its substrates, such interaction remains unclear.

Recent advances in the hardware and software for biomolecular simulation and bioinformatics have rapidly improved the precision and performance of these techniques. We have applied some of these techniques, in combination with experimental methods, to understand the structural and evolutionary basis of the virological phenomena (Oka et al., 2007, 2009; Motomura et al., 2008, 2010; Naganawa et al., 2008; Shirakawa et al., 2008; Yokoyama et al., 2010, 2012; Ode et al., 2011; Sakuragi et al., 2012). In this study, by combining methods of homology modeling, the automated ligand docking, Shannon entropy analysis, site-directed mutagenesis, and *in silico* screening of SaV inhibitors, we studied the structural basis for the substrate recognition by SaV protease.

MATERIALS AND METHODS

STRUCTURAL MODELING OF SaV PROTEASE DOCKED TO THE SUBSTRATE OCTAPEPTIDES

We first constructed a ligand-free protease domain model of the SaV Mc10 strain (Oka et al., 2005b; GenBank accession number: AY237420) by the homology modeling technique (Sanchez et al., 2000; Baker and Sali, 2001) as described previously (Oka et al., 2007). The modeling was performed using tools available in the Molecular Operating Environment (MOE; Chemical Computing Group, Inc., Montreal, QC, Canada). As the modeling template, we used the high-resolution crystal structure of norovirus 3C-like protease at a resolution of 1.50 Å [Protein Data Bank (PDB) code: 2FYQ; Zeitler et al., 2006] because, like SaV, the norovirus belongs to the family Caliciviridae, and thus the protease sequence shows a higher identity to the SaV protease sequences (about 25% identity) than to the other available 3C-like protease sequences of viruses. We applied the multiple sequence alignment approach (Baker and Sali, 2001) using the reported 3C-like proteases to minimize misalignments of the target and template sequences, as described previously (Oka et al., 2007; Shirakawa et al., 2008). The sequences used for the alignment included those of the rhinovirus 3C-like protease (PDB code: 1CQQ; Matthews et al., 1999), the poliovirus 3C-like protease (PDB code: 1L1N; Mosimann et al., 1997), and the hepatitis A virus 3C-like protease (PDB code: 1QA7; Bergmann et al., 1999). The alignment was done with the alignment tool MOE-Align, and homology modeling was done with the tool MOE-Homology in MOE. We optimized the 3-D model thermodynamically via energy minimization using the MOE and an AMBER99 force field (Ponder and Case, 2003). We further refined the physically unacceptable local structure of the models based on a Ramachandran plot evaluation using MOE. The 3-D models of the six octapeptides corresponding to the six cleavage sites of the ORF1 precursor protein of the SaV Mc10 strain (NS1/NS2, NS2/NS3, NS3/NS4, NS4/NS5, NS5/NS6-7, and NS6-7/VP1) were constructed using the Molecular Builder module in MOE. Subsequently, the thermodynamically and physically optimized protease models were used to construct protease-substrate complex models. Individual octapeptide models were docked with the optimized SaV protease domain model described above, using the automated ligand docking program ASEDock2005 (Goto et al., 2008) operated in MOE as described previously (Yokoyama et al., 2010). Default setting in ASEDock2005 was applied for the search of the candidate docking structures, and the structures with the best docking score expressed by the arbitrary docking energy (U_{dock}) in ASEDock2005 (Kataoka and Goto, 2008) were selected for the analysis of the protease-substrate interaction sites.

ANALYSIS OF AMINO ACID DIVERSITY WITH INFORMATION ENTROPY

The amino acid diversity at individual sites of the SaV protease domain was analyzed with Shannon entropy scores as described previously (Sander and Schneider, 1991; Mirny and Shakhnovich, 1999; Oka et al., 2009). The amino acid sequences of the protease domain of various human SaV strains from different geographic regions in the world were obtained from GenBank (the number of sequences is 19; accession numbers: X86560, AY694184, AY237422, AY237423, AY646853, AY646854, AJ249939, AY237420, AY237419, AY646855, AY603425, AJ786349, DQ058829, DQ125333, AY646856, DQ125334, DQ366344, DQ366345, DQ366346). The amino acid diversity within the SaV protease population was calculated using Shannon's formula (Shannon, 1948):

$$H(i) = -\sum_{x_i} p(x_i) \log_2 p(x_i) \quad (x_i = G, A, I, V, \ldots\ldots),$$

where $H(i)$, $p(x_i)$, and i indicate the amino acid entropy (H) score of a given position, the probability of occurrence of a given amino acid at the position, and the number of the position, respectively. An H score of zero indicates absolute conservation, whereas 4.4 bits indicates complete randomness. The H scores were displayed on the 3-D structure of the SaV protease model constructed above.

We also calculated the Shannon entropy by considering the physicochemical properties of amino acid residues, i.e., the chemical properties and size of side chains as described previously (Oka et al., 2009). For analysis of the diversity in the chemical properties, the amino acid residues were classified into seven groups: acidic (D,E), basic (R,K,H), neutral hydrophilic (S, T, N, Q), aliphatic (G, A, V, I, L, M), aromatic (F, Y, W), thio-containing (C), and imine (P). For analysis of the diversity in the size of side chains, the amino acid residues were classified into four groups: small (G, A, C, S), medium-small (T, V, N, D, I, L, P, M), medium-large (Q, E, R, K), and large (H, F, Y, W). The H scores were plotted on the 3-D structure of the SaV protease model.

SITE-DIRECTED MUTAGENESIS OF THE SaV PROTEASE DOMAIN

The detailed strategy of the mutagenesis for the SaV protease domain has been described previously (Oka et al., 2005b, 2006). Briefly, we used the full-length cDNA clone of the genome of the SaV strain Mc10 (pUC19/SaV Mc10 full-length; GenBank accession number: AY237420) as a starting material for the mutagenesis. We constructed nine SaV Mc10 full-length mutant cDNA clones. Site-directed mutagenesis was performed using a GeneTailor Site-Directed Mutagenesis System (Invitrogen). The oligonucleotides used for the site-directed mutagenesis were as follows (the codons corresponding to changed amino acid(s) are indicated in lowercase): for T1085A, 5′-GTGGTTGTCACAGTTgcaCACGTGGCCTCTGCG-3′; for Y1156A, 5′-ATCACGGTCCAGGGGgctCACCTGCGCATC

ATA-3′; for K1167A, 5′-ATGGATACCCAACAgcgCGTGGGGACT GTGGCAC-3′; for R1168A, 5′-GATACCCAACAAAGgctGGGGAC TGTGGCACAC-3′; for K1167E, 5′-ATGGATACCCAACAgagCGT GGGGACTGTGGCAC-3′; for R1168E, 5′-ATGGATACCCAACAA AGgagGGGGACTGTGGCACAC-3′; for K1167AR1168A, 5′-ATGGATACCCAACAgcggcgGGGGACTGTGGCAC-3′; and for K1167ER1168E, 5′-ATGGATACCCAACAgaggagGGGGACTGTG GCAC-3′. The T1085AY1156A mutant was generated with the above Y1156A primer using methylated DNA of the T1085A as the template. All the mutant clones constructed were subjected to the sequencing of the entire genomic cDNA region to verify the absence of unnecessary mutations leading to amino acid changes.

IN VITRO TRANSCRIPTION-TRANSLATION ASSAY

In vitro transcription-translation with a rabbit reticulocyte system was performed using the TNT T7 Quick for PCR DNA kit (Promega, Madison, WI, USA) as described previously (Oka et al., 2005b). Briefly, a template for the *in vitro* transcription-translation, containing the entire ORF1, was prepared by PCR amplification using the full-length cDNA clone. The primers used for the amplification were as follows. The forward primer containing a T7 promoter sequence (underlined) and a translation initiation codon (bold) was 5′-GGATCCTAA TACGACTCACTATAGGGAACAGCCACC**ATG** gcttccaagccattcta cccaatagag-3′; and the antisense primer containing a stop codon (bold) was 5′-T_{30}**TTA**-ttctaagaacctaacggcccgg. The PCR product (3 μl) was mixed with 20 μl of TNT T7 PCR Quick Master Mix (Promega) and 2 μl of Redivue Pro-mix L- [^{35}S] *in vitro* cell-labeling mix (GE Healthcare Biosciences, Piscataway, NJ, USA). The mixture was incubated at 30°C for 3 or 16 h and subjected to SDS-polyacrylamide gel electrophoresis (SDS-PAGE). The translation products separated by electrophoresis were blotted onto a PVDF membrane (Immobilon-P; Millipore, Bedford, MA, USA) using a semi-dry electroblotting apparatus (ATTO; Tokyo). The radiolabeled proteins were detected by a BAS 2500 Bioimage Analyzer (Fuji Film, Tokyo).

IMMUNOPRECIPITATION

For the detection of NS1 (p11) and NS5 (VPg), which were undetectable with the above assay system, we performed immunoprecipitation before the SDS-PAGE as described previously (Oka et al., 2005b, 2006, 2009). Briefly, 10 μl of the *in vitro* transcription-translation reaction mixture was diluted with 80 μl of RIPA lysis buffer containing 50 mM Tris, pH 7.4, 150 mM NaCl, 0.25% deoxycholic acid, 1% NP40, and 1 mM EDTA (Upstate, Lake Placid, NY, USA) and incubated with 5 μg of anti-A (anti-NS1) or ant-D (anti-NS5) antibodies raised against *E. coli*-expressed recombinant proteins (aa 1–231 for NS1 and aa 941–1055 for NS5; Oka et al., 2005b). After incubation for 1 h on ice, 25 μl of a suspension of Protein A Magnetic Beads (New England Biolabs) and 900 μl of RIPA buffer were added. The mixture was gently rotated at 4°C for 1 h and then washed three times with 1 ml of RIPA lysis buffer. The immunoprecipitated proteins were resuspended in 20 μl of SDS-PAGE sample buffer and heated at 95°C for 5 min prior to analysis with 5 to 20% Tris-Gly polyacrylamide gel. The proteins were blotted onto an Immobilon-P polyvinylidene difluoride membrane (Millipore). Immunoprecipitated radioactive proteins were detected with a Bioimage Analyzer BAS 2500 (Fuji Film).

THE CHEMICAL COMPOUND LIBRARY

Chemical compounds (139,369 compounds, molecular weights 42–2986) were obtained from the Open Innovation Center for Drug Discovery (The University of Tokyo, Tokyo, Japan). The compound library database of this center provides information on the molecular formula, molecular weight, hydrogen-bond donor-acceptor numbers, topological polar surface area (TPSA), and other physicochemical parameters of the compounds for pharmacophore-based *in silico* drug screening.

PHARMACOPHORE-BASED *IN SILICO* SCREENING

Pharmacophore-based *in silico* screening was done using tools available in the MOE. We created a pharmacophore query with a substrate feature using the Pharmacophore Query Editor tool in MOE. Pharmacophore-based *in silico* screening was done by the Pharmacophore Search module in the MOE using the created query.

DRUG SUSCEPTIBILITY ASSAY

The susceptibility of SaV protease to the synthetic small chemical compound was determined by means of an *in vitro* trans cleavage assay as follows. A radiolabeled full-length Mc10 ORF1 polyprotein containing a defective protease (Pro^{mut}; Oka et al., 2005b) or a non-radiolabeled partial Mc10 ORF1 polyprotein (NS6-7-VP1) containing a functional protease (Pro^{wt}; Oka et al., 2006) was separately expressed using the *in vitro* transcription/translation system (Oka et al., 2011). The PCR primer pairs used for the preparation of DNA template for the expression of the NS6-7-VP1 were as follows. The forward primer was 5′-GGATCCTAATACGACTCACTATAGGGAACAGCCACC**ATG**gctc ccacaccaattgttac-3′, including the T7 promoter sequence (underlined) and a start codon (bold); and the antisense primer was 5′-T_{30}**TTA**-ttctaagaacctaacggcccgg, including a stop codon (bold). Twenty microliter of the non-radiolabeled products containing Pro^{wt} was mixed with 1 μl of 2 mM inhibitor candidate in DMSO and incubated for 10 min at room temperature. Then 10 μl of the radiolabeled full-length ORF1 polyprotein (Pro^{mut}) was added to the Pro^{wt}-inhibitor mixture and incubated at 30°C for 20 h, and subjected to the SDS-PAGE analysis as described above. To quantitate the proteolytic activity of the SaV protease, we measured the intensity of the band corresponding to the NS4-NS5 intermediate processing product with Typhoon 7500 (GE Healthcare), due to the lack of overlapping non-specific products of the *in vitro* translation around the NS4-NS5. The chemical compound concentrations resulting in a 50% reduction of the NS4-NS5 intermediate protein production of the drug-free control were determined on the basis of the dose-response curve and defined as the IC_{50} values of the SaV proteolysis activity.

STRUCTURAL MODELING OF SaV PROTEASE DOCKED TO CHEMICAL COMPOUNDS

Structural models of the chemical compounds were constructed using the Molecular Builder tool in MOE. Individual compounds were docked with the SaV protease domain model using the automated ligand docking program ASEDock2005 (Goto et al., 2008) operated in MOE as described above.

RESULTS

STRUCTURAL MODELING OF SaV PROTEASES DOCKED TO THE SUBSTRATE OCTAPEPTIDES

To obtain structural insights into the protease-substrate interactions at the atomic level, we constructed a 3-D model of the intact protease domain of the SaV Mc10 strain, which were docked to octapeptides corresponding to the six authentic cleavage sites (P4–P4′ sites) of the ORF1 polyprotein of the Mc10 strain (see Materials and Methods for details; **Figure 1**). The amino acid sequences of the six octapeptides are very different from each other (**Figure 1A**). Despite the variation, the octapeptides bound to the protease with the same orientation in the clefts of the protease (**Figure 1B**); the P1–P4 amino acid residues bound to the cleft between the N- and C-terminal domains, whereas the P1′–P4′ amino acid residues bound to the cleft on the C-terminal domain. The docking positions were functionally reasonable, because they allowed the cleavage sites of the octapeptides to be placed near the amino acids essential for the catalytic activity of the SaV protease, i.e., H^{31}, E^{52}, C^{116}, and H^{131} (Oka et al., 2005b). Other docking positions caused docking results with very poor docking scores and did not fulfill the functional requirement for the catalytic reaction.

The protease-peptide complex models disclosed two interaction sites that were common to the six bound peptides. First, the substrate amino acid residues at the P4 position were exclusively placed in a thin cleft, termed cleft 1, that was formed by threonine (T), glutamic acid (E), and tyrosine (Y) at positions 30, 52, and 101 of the protease domain (T^{30}, E^{52}, and Y^{101}, respectively; **Figures 2A–C**). An aromatic ring of the phenylalanine (F) or Y at P4 of the octapeptides of the NS1/NS2, NS2/NS3, NS3/NS4, NS5/NS6-7, and NS6-7/VP1 cleavage sites (**Figure 1A**) was positioned such that an aromatic stacking could be generated with the Y^{101} in the protease cleft 1 (**Figures 2A–C**). The steric configuration of the aromatic rings of the P4 amino acid residues in the bound state was very similar except for the Y of the NS5/NS6-7 cleavage site (**Figure 2C**). In the case of the NS4/NS5 peptide, the P4 amino acid is the arginine (R; **Figure 1A**) and was arranged near the side chain of the E^{52} (**Figure 2C**).

Second, the substrate amino acid residues at the P1 site were exclusively placed in a small positively charged cleft, termed cleft 2, that was formed by the histidine (H), H, lysine (K), and R at positions 14, 31, 112, and 113 of the protease domain (H^{14}, H^{31}, K^{112}, and R^{113}, respectively; **Figures 2D–F**). In four out of the six cleavage sequences the P1 amino acid is negatively charged (E; **Figure 1A**) that could interact electrostatically with the side chains of the positively charged cleft 2 of the protease (**Figure 2D**). In the case of the NS2/NS3 and NS3/NS4, the P1 amino acid was glutamine (Q; **Figure 1A**) which is hydrophilic and thus could cause electrostatic interactions via a polarized charge. The steric configuration of the side chains of the P1 amino acid residues at the bound state was very similar (**Figure 2F**). The simulated docking between the protease and the substrate having alanine substitutions at P1 and P4 positions resulted in a docking position similar to that for the wild-type substrate, whereas the docking score was reduced to about 1/2. Collectively, these results suggest that the interactions at the P1 and P4 sites of the substrates play a key role in the substrate recognition, as suggested in the previous experiments (Robel et al., 2008; Oka et al., 2009).

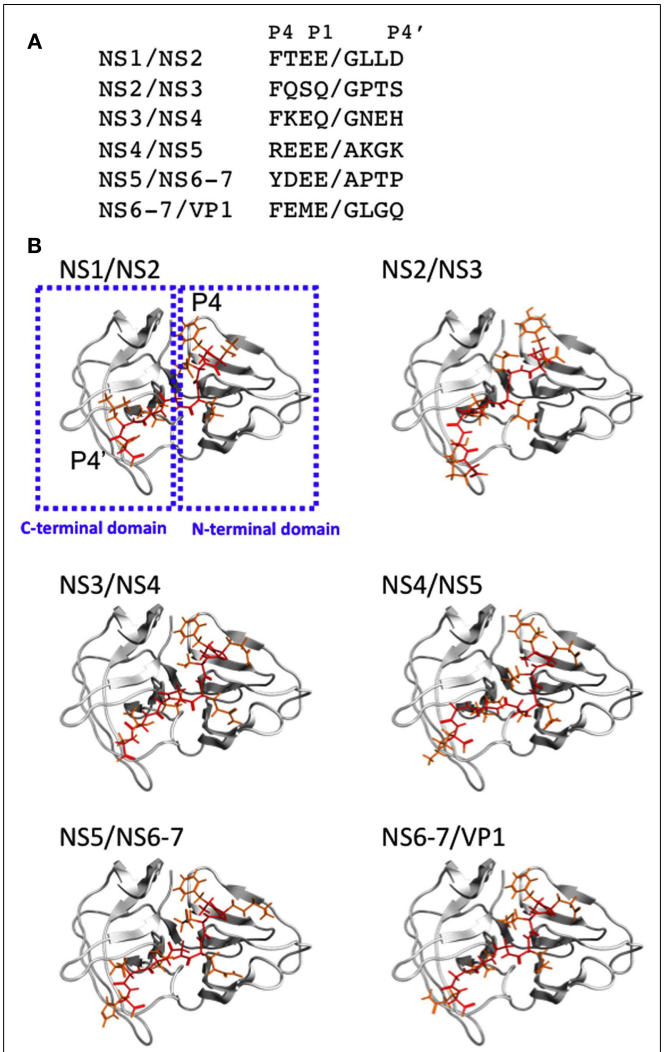


FIGURE 1 | Structural models of SaV protease docked to the substrate octapeptides. (A) Sequences of the six cleavage sites of the SaV ORF1 polyprotein are shown with one-letter amino acid codes. Slashes represent the peptide bonds cleaved by the protease. **(B)** Structural models of the SaV protease-substrate complex. The 3-D structural model of the protease domain was constructed by homology modeling and thermodynamically and physically refined as described previously (Oka et al., 2007). The 3-D structural models of the octapeptides corresponding to the six authentic cleavage sites of the SaV Mc10 ORF1 were constructed by using the Molecular Builder tool in MOE. The optimized protease model was docked to individual octapeptides using the automated ligand docking program ASEDock2005 (Goto et al., 2008) operated in MOE as described previously (Yokoyama et al., 2010). Red and orange sticks indicate main and side chains of the octapeptides, respectively.

AMINO ACID DIVERSITY OF HUMAN SaV PROTEASE

To obtain evolutionary insights into the protease-substrate interactions, we analyzed the amino acid diversity of the protease domain among various human SaV strains in the public database. Full-length human SaV protease domain sequences were collected from GenBank ($N = 19$) and used to calculate the Shannon entropy scores, H (Shannon, 1948), in order to analyze the diversity of individual amino acid residues in the SaV population

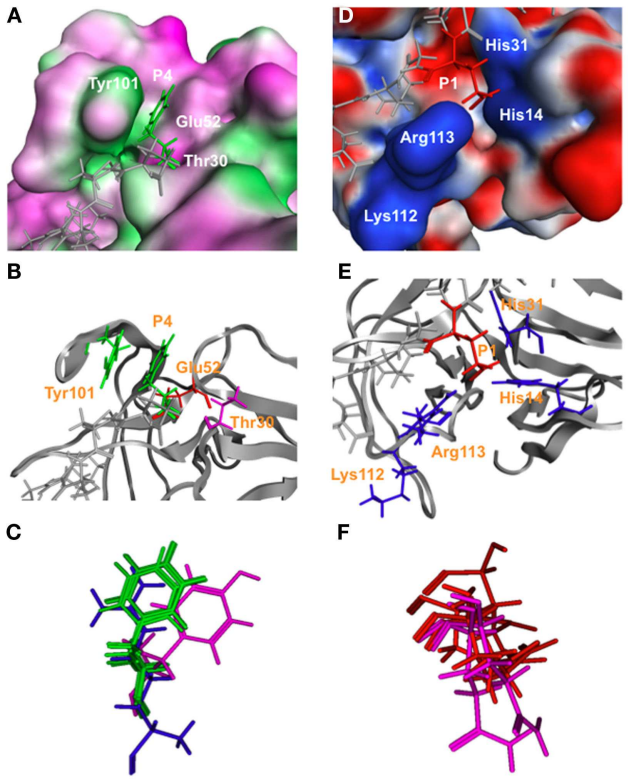


FIGURE 2 | Close-up views of the interaction sites in the protease-substrate complex models. The sites of interactions between the SaV Mc10 protease and the side chains of octapeptides at the P4 **(A–C)** and P1 **(D–F)** sites are highlighted. Upper panels show the cleft formed by T^{30}, E^{52}, and Y^{101}, and the positively charged cleft formed by the H^{14}, H^{31}, K^{112}, and R^{113} of the proteases that are bound to a side chain of the NS6-7/VP1 octapeptide at the P4 **(A)** and P1 **(D)** sites. Middle panels show the relative configurations of side chains of the protease bound to a side chain of the NS6-7/VP1 octapeptide at the P4 **(B)** and P1 **(E)** sites. Bottom panels show the superposed structures of the side chains of the P4 **(C)** and P1 **(F)** amino acid residues of the 6 cleavage sites in the SaV protease-substrate complex models. **(C)** Green, magenta, and blue sticks represent the side chains of phenylalanine, tyrosine, and arginine, respectively, at the P4 site in the substrate-protease complexes. **(F)** Red and magenta sticks represent the side chains of glutamic acid and glutamine, respectively, at the P1 site in the substrate-protease complexes.

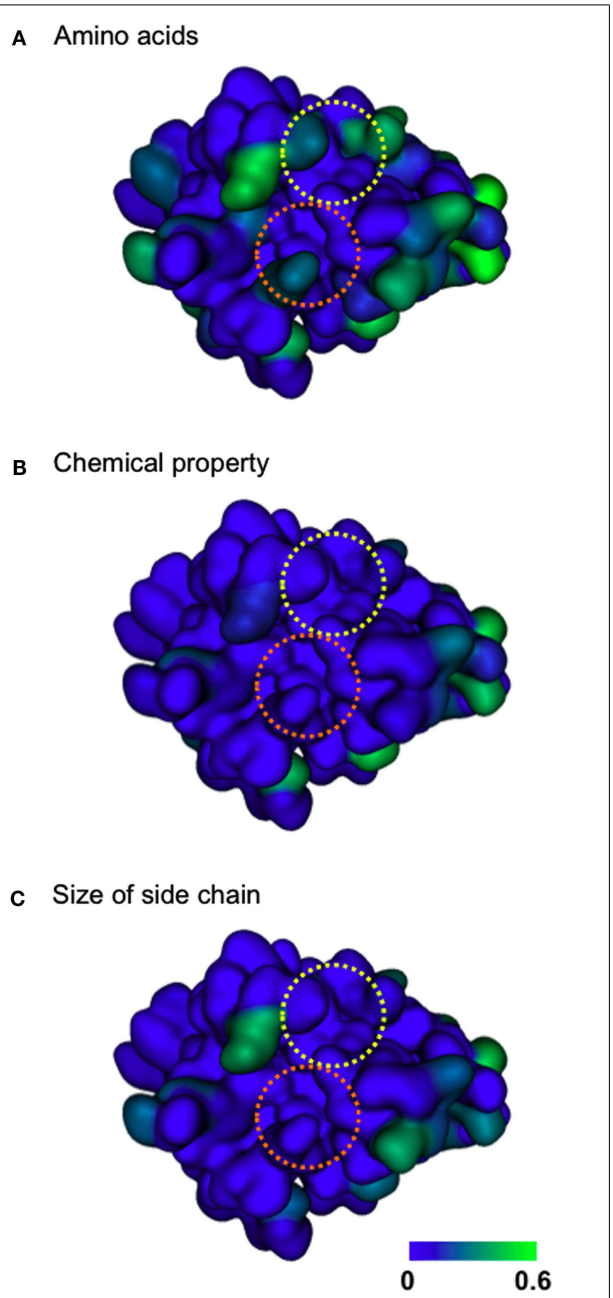


FIGURE 3 | Diversity of SaV protease amino acid residues. The amino acid diversity at individual sites of the SaV protease domain was analyzed with information entropy as described previously (Oka et al., 2009). The Shannon entropy *H* was calculated with Shannon's formula (Shannon, 1948) based on amino acid residues **(A)**, chemical properties **(B)**, and the size of the side chain **(C)** using amino acid sequences of the SaV full-length protease domain from GenBank ($N = 19$). For analysis of the diversity in the chemical properties, the amino acid residues were classified into seven groups: acidic (D,E), basic (R,K,H), neutral hydrophilic (S, T, N, Q), aliphatic (G, A, V, I, L, M), aromatic (F, Y, W), thio-containing (C), and imine (P). For analysis of the diversity in the size of the side chain, the amino acid residues were classified into 4 groups: small (G, A, C, S), medium-small (T, V, N, D, I, L, P, M), medium-large (Q, E, R, K), and large (H, F, Y, W). The *H* scores are plotted on the 3-D structure of the SaV protease model, where an *H* score of zero indicates absolute conservation. Yellow and orange dotted circles indicate clefts 1 and 2, respectively.

as described previously (Oka et al., 2009). The *H* scores generally ranged from 0.0 to 0.6 bits (**Figure 3A**), indicating that the diversity of the SaV protease is relatively small, as seen in many viral enzymes. The variable sites were essentially located on the surface region, indicating that some exposed regions of the SaV protease allow amino acid changes (**Figure 3A**, greenish sites). Although less extensive, some variation was detected at Y^{101} and R^{113} in the clefts 1 and 2, respectively (**Figure 3A**, two dotted circles). However, when the *H* scores were calculated on the basis of chemical properties or the size of the amino acid residues, they were nearly zero throughout the substrate-binding cleft (**Figures 3B,C**). Similarly, the protease amino acid residues, which constitute a large cavity for the binding of entire octapeptides, were sometime variable but highly conserved in the context of the chemical properties

and sizes of the side chains (**Figures 1B** and **3**). Thus the SaV protease appears to restrict extensive changes in the shape and chemical properties of the substrate-binding surface for its survival.

SITE-DIRECTED MUTAGENESIS OF SaV PROTEASE

Consistent with the above structural and diversity data, we previously reported that the E^{52} in cleft 1, as well as H^{14} and H^{31} in cleft 2, are essential to maintain proper processing by SaV protease (Oka et al., 2007). To obtain further insights into the biological roles of clefts 1 and 2 in the proteolysis of the SaV precursor polyprotein, we performed additional site-directed mutagenesis using a full-length clone of SaV Mc10 strain (Oka et al., 2005b). The Mc10 ORF1 encodes a polypeptide of 2278 amino acid residues, where the six cleavage sites have been experimentally determined (Oka et al., 2006; **Figure 4A**). A total of nine mutants of the SaV protease domain were constructed using the Mc10 ORF1. Full-length ORF1 precursor proteins having a single or double mutations in the protease domain were expressed using the *in vitro* transcription-translation system, and the processing products were analyzed by gel electrophoresis as described previously (Oka et al., 2005b, 2006, 2007, 2009). The Mc10 functional protease (Pro^{wt}) and a defective mutant completely lacking the proteolysis activity (Pro^{mut}; Oka et al., 2005b) were used as positive and negative controls of the proteolysis, respectively.

When the ORF1 containing the Pro^{wt} was expressed, nine products corresponded in size to the mature proteins NS1, NS2, NS3, NS4, NS5, and VP1, and relatively stable intermediate proteins,

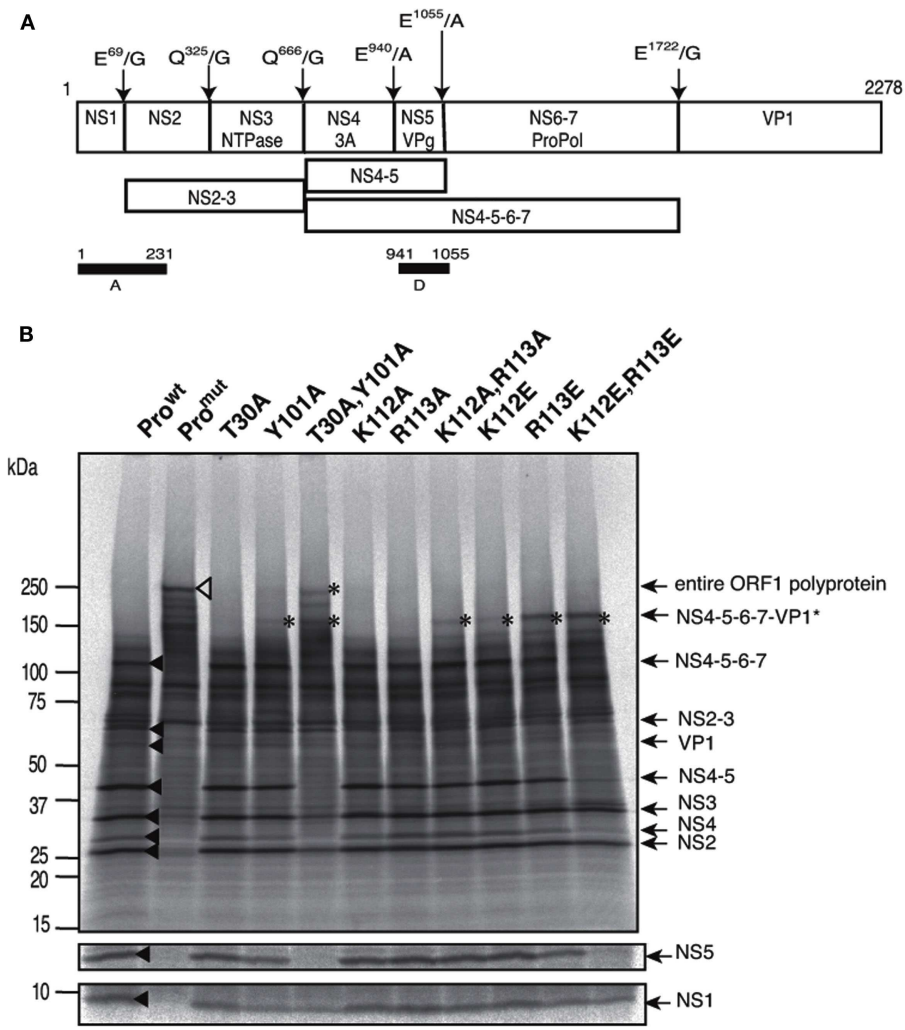


FIGURE 4 | Site-directed mutagenesis of the substrate interaction sites of SaV Mc10 protease. (A) Proteolytic cleavage map of the SaV Mc10 ORF1 polyprotein and the processing intermediates (Oka et al., 2006). Black bars indicate the protein segments, A and D, used to raise polyclonal antibodies for detection of the NS1 and NS5 proteins, respectively. **(B)** SDS-PAGE of ^{35}S-labeled *in vitro* translation products of SaV Mc10 ORF1 containing various protease mutants. NS1 and NS5 were detected by immunoprecipitation using anti-A or anti-D polyclonal antibodies as described previously (Oka et al., 2005b, 2006, 2009). Mc10 ORF1 containing functional protease (Pro^{wt}) and a defective mutant lacking in the proteolysis activity (Pro^{mut}) were included as described previously (Oka et al., 2005b). Newly appearing products when compared to Pro^{wt} are indicated by asterisks. Size markers are shown on the left. Mc10 ORF1-specific proteins (Oka et al., 2005b, 2006) are shown on the right.

such as NS2-3, NS4-5, and NS4-5-6-7 were detected (**Figure 4B**, lane Pro^{wt}, black arrowheads; Oka et al., 2005b, 2006, 2009). These products were undetectable in the Pro^{mut} ORF1 sample, and instead a product corresponding to the ORF1 polyprotein was detected (**Figure 4B**, lane Pro^{mut}, open triangle; Oka et al., 2005b, 2006, 2009). A single alanine substitution at T^{30} in the cleft 1, K^{112} in the cleft 2, or R^{113} in the cleft 2 of viral protease resulted in a processing pattern similar to that of Pro^{wt} (**Figure 4B**, lanes T30A, K112A, and R113A). On the other hand, a single alanine substitution at Y^{101} in the cleft 1 (Y101A), a single acidic substitution at K^{112} or R^{113} in the cleft 2 (K112E and R113E), and double mutations in each cleft (T30AY101A and K112ER113E) resulted in abnormality of the precursor processing, i.e., an increase in accumulation of the full-length ORF1 polyprotein and/or the NS4-5-6-7-VP1 intermediate protein (**Figure 4B**, asterisks). In the samples expressing ORF1 with T30A/Y101A or K112E/R113E double mutations, processing products corresponding to the NS5 and NS4-5 disappeared almost completely (**Figure 4B**, lanes 5 and 11, respectively).

PHARMACOPHORE-BASED *IN SILICO* SCREENING FOR THE LEAD COMPOUNDS OF SaV PROTEASE INHIBITORS

To further assess the role of the clefts 1 and 2 in the ligand binding, we performed a pharmacophore-based *in silico* screening of protease inhibitors. A total of 139,369 compounds (molecular weights 42–2986) were screened for the lead molecules that contain an aromatic-ring-like portion resembling the P4 amino acid, a negatively charged portion resembling the P1 amino acid, and a hydrophobic portion resembling the P1′ amino acid, being arranged at similar 3-D positions with the authentic substrates (**Figure 5**). The hydrophobic portion resembling the P1′ amino acid was included to better mimic the authentic substrate structures. A total of 151 lead compounds matched to the category were then subjected to the *in vitro* trans cleavage assay of the SaV Mc10 ORF1 polyprotein. With this screening, we could obtain three compounds that inhibited processing of the SaV ORF1 at IC_{50} values of 18.4–26.5 μM (**Figure 6**).

We then analyzed how the lead compounds bound to the SaV Mc10 protease by docking simulation (**Figure 7**). As expected, these compounds were predicted to bind to the protease at the same interaction sites by which the authentic substrates bound to the protease. The aromatic-ring-like portion resembling the P4 amino acid bound to the thin cleft formed by T^{30}, E^{52}, and Y^{101} for the binding of the side chain of the P4 amino acid. The negatively charged portion resembling the P1 amino acid bound to the small positively charged pocket formed by the H^{14}, H^{31}, K^{112}, and R^{113} for the binding of the side chain of the P1 amino acid.

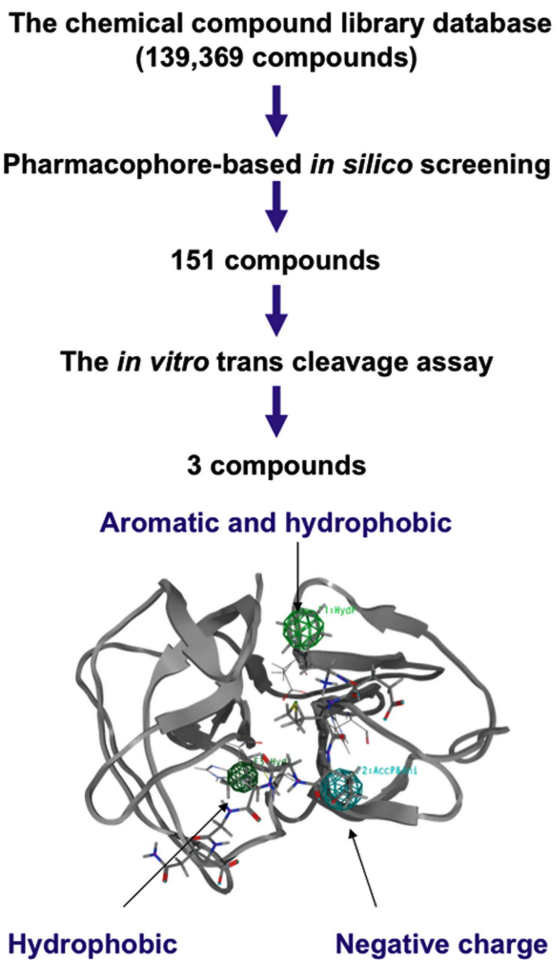


FIGURE 5 | Flow chart for screening chemical compounds against SaV protease. Pharmacophore-based *in silico* screening (Schuster et al., 2006a,b; Kirchmair et al., 2007) was applied to extract the lead compounds having structural features that resembled those of the substrates of SaV protease. One hundred and fifty-one compounds exhibiting some similarities to the authentic substrates were further assessed with respect to their inhibitory activity against SaV protease. With this strategy, we could obtain three compounds that inhibited the processing of the SaV ORF1 polyprotein.

DISCUSSION

The viral proteins that support viral replication and make up the viral particle are often translated as part of polyprotein precursors. Viral protease catalyzes cleavage of the precursor protein and thus plays an essential role in the viral life cycle. In this study, by combining computational and experimental approaches, we studied the structural basis for the substrate recognition by SaV protease. The results obtained in this study were consistent with each other and disclosed novel structural base points of the protease for the attractive interactions with specific structures of ligands.

Using a homology modeling and a docking tool, we first examined the physical interactions of SaV protease and octapeptides corresponding to the six authentic cleavage sites of the SaV ORF1 polyprotein. Despite the marked sequence variation of the octapeptides, they were bound to the protease in the same orientation in the structural models (**Figure 1**). The results suggested that there might be common interaction sites that served as fulcrums to direct the orientation of the octapeptides. Consistently, the models disclosed two interaction sites that were shared with the six peptides and support the stable and functional binding of substrates to the catalytic cavity; the variable side chains at the P4 and P1 sites of the peptides were consistently bound to the two small clefts, termed clefts 1 and 2, respectively (**Figure 2**). The former

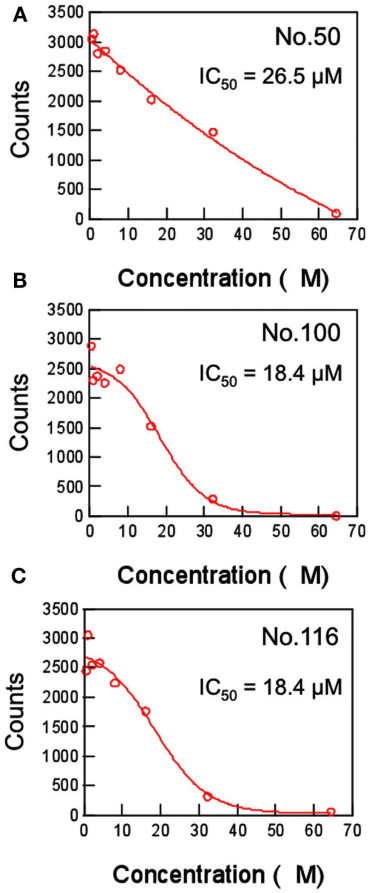


FIGURE 6 | Dose-response curves of the inhibitors against SaV protease. The inhibitory effects of the three chemical compounds that were screened for their structural similarity to the authentic substrates of SaV protease were determined with an *in vitro* trans cleavage assay. A radiolabeled full-length Mc10 ORF1 polyprotein containing a defective protease (Pro^{mut}; Oka et al., 2005b) or a non-radiolabeled partial Mc10 ORF1 polyprotein (NS6-7-VP1) containing a functional protease (Pro^{wt}; Oka et al., 2006) was separately expressed using the *in vitro* transcription/translation system. The translation products were mixed and incubated in the presence of increasing concentrations of the indicated compounds at 30˚C for 20 h. The intensity of the radioactive band corresponding to the NS4-NS5 product was measured with Typhoon 7500 and plotted in relation to the compound concentrations. **(A)** Compound No.50. **(B)** Compound No.100. **(C)** Compound No.116.

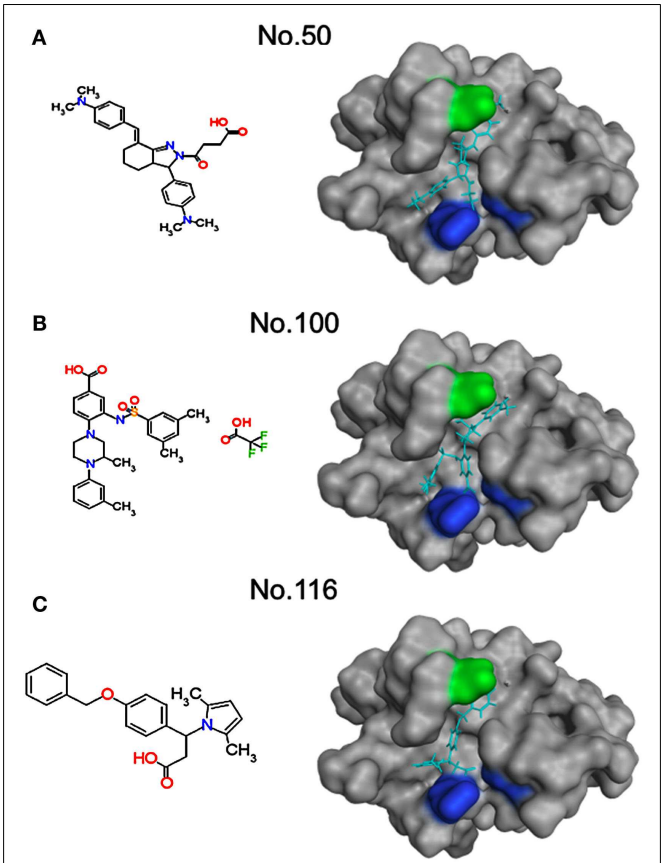


FIGURE 7 | Structural models of SaV protease docked to the inhibitors. Molecular formulas of the inhibitors (left) and structural models of the inhibitor-protease complexes (right) are shown. Molecular models of the three chemical compounds having anti-SaV-protease activity were constructed using the Molecular Builder tool in MOE. Individual compounds were docked to the SaV protease domain model using the automated ligand docking program ASEDock2005 (Goto et al., 2008). Light blue sticks in the protease indicate inhibitors. Greenish and bluish portions of the protease indicate an aromatic and hydrophobic site and positively charged site, respectively. **(A)** Compound No.50. **(B)** Compound No.100. **(C)** Compound No.116.

participated in aromatic stacking interactions, whereas the latter participated in electrostatic interactions. These results are consistent with the previous findings that the P4 and P1 amino acid residues of the substrates play key roles in efficient proteolysis by SaV protease (Robel et al., 2008; Oka et al., 2009) and predicted that these two clefts could play a key role in substrate recognition via interactions with the P4 and P1 amino acid residues of substrates.

This prediction was assessed by several analyses. If the clefts played essential roles in recognition of substrates, spontaneous mutations that alter profoundly the physicochemical properties of the clefts should be suppressed for viral survival. Consistently our Shannon entropy study using protease sequences of various SaV strains from the world shows that the amino acid residues forming the clefts 1 and 2 are variable but highly conserved in terms of the chemical properties or the sizes of side chains (**Figure 3**). The results indicate that these clefts tolerate mutations in nature but resist a range of mutations that markedly alter the chemical properties or the shapes of the cleft surface. The findings are consistent with the above structure-based prediction on the function of the clefts 1 and 2. These clefts are located on the surface of the large cavity of the protease. Therefore, the restrictions in the variation in two clefts are likely to be caused by functional constraints for the essential interactions.

Moreover, we examined whether a range of mutations that markedly alter the physicochemical properties of the clefts indeed could result in aberrant processing of the SaV precursor polyprotein. Our site-directed mutagenesis study showed that a single mutation in cleft 1 (T30A) or in cleft 2 (K112A or R113A) caused little detectable damage in the processing of the viral precursor

polyprotein, showing a tolerance to mutations as indicated by our information entropy study. Notably, however, (i) a single mutation that causes a loss of aromatic stacking interaction (Y101A) in the cleft 1, (ii) a single mutation that causes a loss of the electrostatic interaction in the cleft 2 (K112E or R113E), and (iii) double mutations within the clefts unexceptionally resulted in incomplete processing (**Figure 4**). The results indicate that the abnormal processing was caused only by single mutations that could extensively alter the chemical properties of the clefts. The data agree with the entropy data and again suggest the acceptability of variation in the two clefts under functional constraints.

Finally, we performed *in silico* screening of SaV protease inhibitors on the basis of the above structural and biological information. The screening of the 139,369 compounds *in silico* led to the identification of the 151 compounds that resembled the structural and spatial features of the P4 and P1 amino acid residues of authentic substrates (**Figure 5**). From them, we could experimentally identify the three compounds that inhibited proteolysis of the SaV precursor polyprotein *in vitro* (**Figure 6**). As expected, these compounds were predicted to bind to the SaV protease at the two clefts via similar attractive interactions with the authentic ligands (**Figure 7**). These results provide additional evidence that two clefts on the SaV protease cavity play a key role in the ligand recognition by providing the structural base points for the specific attractive interactions.

Notably, six cleavage sites of SaV precursor polyprotein also differ with respect to their susceptibility to the SaV protease, with the NS2/NS3, NS4/NS5, and NS5/NS6-7 sites being consistently more resistant to the cleavage than the NS1/NS2, NS3/NS4, and NS6-7/VP1 sites (Oka et al., 2005b, 2006, 2009). In this regard, it is of note that the P4 position of the NS4/NS5 site of human SaV is exclusively arginine instead of an aromatic amino acid (**Figure 1A**) and that this arginine is conserved in all human SaV strains reported thus far (Oka et al., 2005b, 2006, 2009). This substitution at P4 position will abolish the aromatic stacking interaction in the cleft 1 and thus will attenuate attractive interactions between protease and the NS4/NS5 cleavage site. This possibility is well consistent with the experimental findings; the cleavage of the NS4/NS5 site is less efficient than that of the other sites (Oka et al., 2005b, 2006, 2009) and is more sensitive to the cleft 1 mutations than the other cleavage sites are (**Figure 4**, lane 5, NS5). Moreover, the attenuation of cleavage of the NS4/NS5 site was reversed simply by replacing the arginine with phenylalanine at the P4 site (Oka et al., 2009). These findings strongly suggest that the well-preserved arginine at the P4 position of the SaV NS4/NS5 cleavage site plays a key role in maintaining the distinct cleavability of precursor polyprotein by SaV protease.

In this study, we disclosed a novel 3-D pharmacophore containing two clefts on the cavity of the SaV protease, which can be used to identify the lead compounds of SaV protease inhibitors. SaV is one of the commonly detected pathogens in the acute gastroenteritis of both children and adults (Johansson et al., 2005; Harada et al., 2009; Iturriza-Gomara et al., 2009; Pang et al., 2009). Diarrhea is one of the greatest causes of mortality in children under age 5 in many countries (Boschi-Pinto et al., 2008), and the outbreaks of the acute gastroenteritis often seriously affects the clinical, economic, and social activities. Therefore, anti-viral compounds against SaV may be beneficial to some at-risk populations or communities. Thus far no anti-SaV inhibitors for the clinical use have been developed. Our findings will provide important clues to the unique specificity of the SaV protease, the regulation of SaV maturation, and the rationale design of anti-SaV inhibitors.

ACKNOWLEDGMENTS

We thank Mami Yamamoto and Kana Miyashita for their technical assistance with the mutagenesis. This work was supported by a grant from the Japan Health Science Foundation for Research on Health Sciences Focusing on Drug Innovation, and grants for Research on Emerging and Re-emerging Infectious Diseases and Food Safety from the Ministry of Health, Labour and Welfare of Japan.

REFERENCES

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. *Science* 294, 93–96.

Belliot, G., Sosnovtsev, S. V., Mitra, T., Hammer, C., Garfield, M., and Green, K. Y. (2003). In vitro proteolytic processing of the MD145 norovirus ORF1 nonstructural polyprotein yields stable precursors and products similar to those detected in calicivirus-infected cells. *J. Virol.* 77, 10957–10974.

Bergmann, E. M., Cherney, M. M., Mckendrick, J., Frormann, S., Luo, C., Malcolm, B. A., Vederas, J. C., and James, M. N. (1999). Crystal structure of an inhibitor complex of the 3C proteinase from hepatitis A virus (HAV) and implications for the polyprotein processing in HAV. *Virology* 265, 153–163.

Boschi-Pinto, C., Velebit, L., and Shibuya, K. (2008). Estimating child mortality due to diarrhoea in developing countries. *Bull. World Health Organ.* 86, 710–717.

Bull, R. A., Hyde, J., Mackenzie, J. M., Hansman, G. S., Oka, T., Takeda, N., and White, P. A. (2011). Comparison of the replication properties of murine and human calicivirus RNA-dependent RNA polymerases. *Virus Genes* 42, 16–27.

Chiba, S., Nakata, S., Numata-Kinoshita, K., and Honma, S. (2000). Sapporo virus: history and recent findings. *J. Infect. Dis.* 181(Suppl. 2), S303–S308.

Chiba, S., Sakuma, Y., Kogasaka, R., Akihara, M., Horino, K., Nakao, T., and Fukui, S. (1979). An outbreak of gastroenteritis associated with calicivirus in an infant home. *J. Med. Virol.* 4, 249–254.

Fullerton, S. W., Blaschke, M., Coutard, B., Gebhardt, J., Gorbalenya, A., Canard, B., Tucker, P. A., and Rohayem, J. (2007). Structural and functional characterization of sapovirus RNA-dependent RNA polymerase. *J. Virol.* 81, 1858–1871.

Goto, J., Kataoka, R., Muta, H., and Hirayama, N. (2008). ASEDock-docking based on alpha spheres and excluded volumes. *J. Chem. Inf. Model.* 48, 583–590.

Guo, M., Chang, K. O., Hardy, M. E., Zhang, Q., Parwani, A. V., and Saif, L. J. (1999). Molecular characterization of a porcine enteric calicivirus genetically related to Sapporo-like human caliciviruses. *J. Virol.* 73, 9625–9631.

Hansman, G. S., Oka, T., Katayama, K., and Takeda, N. (2007). Human sapoviruses: genetic diversity, recombination, and classification. *Rev. Med. Virol.* 17, 133–141.

Harada, S., Okada, M., Yahiro, S., Nishimura, K., Matsuo, S., Miyasaka, J., Nakashima, R., Shimada, Y., Ueno, T., Ikezawa, S., Shinozaki, K., Katayama, K., Wakita, T., Takeda, N., and Oka, T. (2009). Surveillance of pathogens in outpatients with gastroenteritis and characterization of sapovirus strains between 2002 and 2007 in Kumamoto Prefecture, Japan. *J. Med. Virol.* 81, 1117–1127.

Hardy, M. E., Crone, T. J., Brower, J. E., and Ettayebi, K. (2002). Substrate specificity of the Norwalk virus 3C-like proteinase. *Virus Res.* 89, 29–39.

Iturriza-Gomara, M., Elliot, A. J., Dockery, C., Fleming, D. M., and Gray, J. J. (2009). Structured surveillance of infectious intestinal disease in pre-school children in the community: "The Nappy Study." *Epidemiol. Infect.* 137, 922–931.

Johansson, P. J., Bergentoft, K., Larsson, P. A., Magnusson, G., Widell, A., Thorhagen, M., and Hedlund, K. O. (2005). A nosocomial sapovirus-associated outbreak of gastroenteritis in adults. *Scand. J. Infect. Dis.* 37, 200–204.

Kataoka, R., and Goto, J. (2008). ASE-Dock – docking based on the shape of binding site. *Mol. Sci.* 2, NP008.

Kirchmair, J., Ristic, S., Eder, K., Markt, P., Wolber, G., Laggner, C., and Langer, T. (2007). Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* 47, 2182–2196.

Liu, B. L., Clarke, I. N., Caul, E. O., and Lambden, P. R. (1995). Human enteric caliciviruses have a unique genome structure and are distinct from the Norwalk-like viruses. *Arch. Virol.* 140, 1345–1356.

Matthews, D. A., Dragovich, P. S., Webber, S. E., Fuhrman, S. A., Patick, A. K., Zalman, L. S., Hendrickson, T. F., Love, R. A., Prins, T. J., Marakovits, J. T., Zhou, R., Tikhe, J., Ford, C. E., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., Brothers, M. A., Delisle, D. M., and Worland, S. T. (1999). Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc. Natl. Acad. Sci. U.S.A.* 96, 11000–11007.

Mirny, L. A., and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* 291, 177–196.

Mosimann, S. C., Cherney, M. M., Sia, S., Plotch, S., and James, M. N. (1997). Refined X-ray crystallographic structure of the poliovirus 3C gene product. *J. Mol. Biol.* 273, 1032–1047.

Motomura, K., Oka, T., Yokoyama, M., Nakamura, H., Mori, H., Ode, H., Hansman, G. S., Katayama, K., Kanda, T., Tanaka, T., Takeda, N., and Sato, H. (2008). Identification of monomorphic and divergent haplotypes in the 2006–2007 norovirus GII/4 epidemic population by genomewide tracing of evolutionary history. *J. Virol.* 82, 11247–11262.

Motomura, K., Yokoyama, M., Ode, H., Nakamura, H., Mori, H., Kanda, T., Oka, T., Katayama, K., Noda, M., Tanaka, T., Takeda, N., and Sato, H. (2010). Divergent evolution of norovirus GII/4 by genome recombination from May 2006 to February 2009 in Japan. *J. Virol.* 84, 8085–8097.

Naganawa, S., Yokoyama, M., Shiino, T., Suzuki, T., Ishigatsubo, Y., Ueda, A., Shirai, A., Takeno, M., Hayakawa, S., Sato, S., Tochikubo, O., Kiyoura, S., Sawada, K., Ikegami, T., Kanda, T., Kitamura, K., and Sato, H. (2008). Net positive charge of HIV-1 CRF01_AE V3 sequence regulates viral sensitivity to humoral immunity. *PLoS ONE* 3, e3206. doi:10.1371/journal.pone.0003206

Nakamura, K., Someya, Y., Kumasaka, T., Ueno, G., Yamamoto, M., Sato, T., Takeda, N., Miyamura, T., and Tanaka, N. (2005). A norovirus protease structure provides insights into active and substrate binding site integrity. *J. Virol.* 79, 13685–13693.

Noel, J. S., Liu, B. L., Humphrey, C. D., Rodriguez, E. M., Lambden, P. R., Clarke, I. N., Dwyer, D. M., Ando, T., Glass, R. I., and Monroe, S. S. (1997). Parkville virus: a novel genetic variant of human calicivirus in the Sapporo virus clade, associated with an outbreak of gastroenteritis in adults. *J. Med. Virol.* 52, 173–178.

Numata, K., Hardy, M. E., Nakata, S., Chiba, S., and Estes, M. K. (1997). Molecular characterization of morphologically typical human calicivirus Sapporo. *Arch. Virol.* 142, 1537–1552.

Ode, H., Yokoyama, M., Kanda, T., and Sato, H. (2011). Identification of folding preferences of cleavage junctions of HIV-1 precursor proteins for regulation of cleavability. *J. Mol. Model.* 17, 391–399.

Oka, T., Katayama, K., Ogawa, S., Hansman, G. S., Kageyama, T., Miyamura, T., and Takeda, N. (2005a). Cleavage activity of the sapovirus 3C-like protease in *Escherichia coli*. *Arch. Virol.* 150, 2539–2548.

Oka, T., Katayama, K., Ogawa, S., Hansman, G. S., Kageyama, T., Ushijima, H., Miyamura, T., and Takeda, N. (2005b). Proteolytic processing of sapovirus ORF1 polyprotein. *J. Virol.* 79, 7283–7290.

Oka, T., Murakami, K., Wakita, T., and Katayama, K. (2011). Comparative site-directed mutagenesis in the catalytic amino acid triad in calicivirus proteases. *Microbiol. Immunol.* 55, 108–114.

Oka, T., Yamamoto, M., Katayama, K., Hansman, G. S., Ogawa, S., Miyamura, T., and Takeda, N. (2006). Identification of the cleavage sites of sapovirus open reading frame 1 polyprotein. *J. Gen. Virol.* 87, 3329–3338.

Oka, T., Yamamoto, M., Yokoyama, M., Ogawa, S., Hansman, G. S., Katayama, K., Miyashita, K., Takagi, H., Tohya, Y., Sato, H., and Takeda, N. (2007). Highly conserved configuration of catalytic amino acid residues among calicivirus-encoded proteases. *J. Virol.* 81, 6798–6806.

Oka, T., Yokoyama, M., Katayama, K., Tsunemitsu, H., Yamamoto, M., Miyashita, K., Ogawa, S., Motomura, K., Mori, H., Nakamura, H., Wakita, T., Takeda, N., and Sato, H. (2009). Structural and biological constraints on diversity of regions immediately upstream of cleavage sites in calicivirus precursor proteins. *Virology* 394, 119–129.

Pang, X. L., Lee, B. E., Tyrrell, G. J., and Preiksaitis, J. K. (2009). Epidemiology and genotype analysis of sapovirus associated with gastroenteritis outbreaks in Alberta, Canada: 2004–2007. *J. Infect. Dis.* 199, 547–551.

Ponder, J. W., and Case, D. A. (2003). Force fields for protein simulations. *Adv. Protein Chem.* 66, 27–85.

Robel, I., Gebhardt, J., Mesters, J. R., Gorbalenya, A., Coutard, B., Canard, B., Hilgenfeld, R., and Rohayem, J. (2008). Functional characterization of the cleavage specificity of the sapovirus chymotrypsin-like protease. *J. Virol.* 82, 8085–8093.

Robinson, S., Clarke, I. N., Vipond, I. B., Caul, E. O., and Lambden, P. R. (2002). Epidemiology of human Sapporo-like caliciviruses in the South West of England: molecular characterisation of a genetically distinct isolate. *J. Med. Virol.* 67, 282–288.

Sakuragi, J. I., Ode, H., Sakuragi, S., Shioda, T., and Sato, H. (2012). A proposal for a new HIV-1 DLS structural model. *Nucleic Acids Res.* 40, 5012–5022.

Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000). Protein structure modeling for structural genomics. *Nat. Struct. Biol.* 7(Suppl.), 986–990.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68.

Scheffler, U., Rudolph, W., Gebhardt, J., and Rohayem, J. (2007). Differential cleavage of the norovirus polyprotein precursor by two active forms of the viral protease. *J. Gen. Virol.* 88, 2013–2018.

Schuster, D., Laggner, C., Steindl, T. M., Palusczak, A., Hartmann, R. W., and Langer, T. (2006a). Pharmacophore modeling and in silico screening for new P450 19 (aromatase) inhibitors. *J. Chem. Inf. Model.* 46, 1301–1311.

Schuster, D., Maurer, E. M., Laggner, C., Nashev, L. G., Wilckens, T., Langer, T., and Odermatt, A. (2006b). The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J. Med. Chem.* 49, 3454–3466.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423, 623–656.

Shirakawa, K., Takaori-Kondo, A., Yokoyama, M., Izumi, T., Matsui, M., Io, K., Sato, T., Sato, H., and Uchiyama, T. (2008). Phosphorylation of APOBEC3G by protein kinase A regulates its interaction with HIV-1 Vif. *Nat. Struct. Mol. Biol.* 15, 1184–1191.

Sosnovtsev, S. V., Sosnovtseva, S. A., and Green, K. Y. (1998). Cleavage of the feline calicivirus capsid precursor is mediated by a virus-encoded proteinase. *J. Virol.* 72, 3051–3059.

Wirblich, C., Sibilia, M., Boniotti, M. B., Rossi, C., Thiel, H. J., and Meyers, G. (1995). 3C-like protease of rabbit hemorrhagic disease virus: identification of cleavage sites in the ORF1 polyprotein and analysis of cleavage specificity. *J. Virol.* 69, 7159–7168.

Yokoyama, M., Mori, H., and Sato, H. (2010). Allosteric regulation of HIV-1 reverse transcriptase by ATP for nucleotide selection. *PLoS ONE* 5, e8867. doi:10.1371/journal.pone.0008867

Yokoyama, M., Naganawa, S., Yoshimura, K., Matsushita, S., and Sato, H. (2012). Structural dynamics of HIV-1 envelope Gp120 outer domain with V3 loop. *PLoS ONE* 7, e37530. doi:10.1371/journal.pone.0037530

Zeitler, C. E., Estes, M. K., and Venkataram Prasad, B. V. (2006). X-ray crystallographic structure of the Norwalk virus protease at 1.5-A resolution. *J. Virol.* 80, 5050–5058.

Conflict of Interest Statement: 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.

Received: 11 June 2012; paper pending published: 10 July 2012; accepted: 08 August 2012; published online: 05 September 2012.

Citation: Yokoyama M, Oka T, Kojima H, Nagano T, Okabe T, Katayama K, Wakita T, Kanda T and Sato H (2012) Structural basis for specific recognition of substrates by sapovirus protease. Front. Microbio. ***3****:312. doi: 10.3389/fmicb.2012.00312*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 04 September 2012
doi: 10.3389/fmicb.2012.00319

Identifying viral parameters from *in vitro* cell cultures

Shingo Iwami[1,2]*, Kei Sato[3], Rob J. De Boer[4], Kazuyuki Aihara[5,6], Tomoyuki Miura[3] and Yoshio Koyanagi[3]

[1] *Faculty of Sciences, Department of Biology, Kyushu University, Higashi-ku, Fukuoka, Japan*
[2] *Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan*
[3] *Institute for Virus Research, Kyoto University, Kyoto, Japan*
[4] *Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands*
[5] *Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan*
[6] *Graduate School of Information Science and Technology, The University of Tokyo, Meguro-ku, Tokyo, Japan*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Takao Masuda, Tokyo Medical and Dental University, Japan
Yoshinao Kubo, Nagasaki University, Japan

****Correspondence:***
Shingo Iwami, Faculty of Sciences, Department of Biology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka, Fukuoka 812-8581, Japan.
e-mail: siwami@kyushu-u.org

Current *in vitro* cell culture studies of viral replication deliver detailed time courses of several virological variables, like the amount of virions and the number of target cells, measured over several days of the experiment. Each of these time points solely provides a snap-shot of the virus infection kinetics and is brought about by the complex interplay of target cell infection, and viral production and cell death. It remains a challenge to interpret these data quantitatively and to reveal the kinetics of these underlying processes to understand how the viral infection depends on these kinetic properties. In order to decompose the kinetics of virus infection, we introduce a method to "quantitatively" describe the virus infection in *in vitro* cell cultures, and discuss the potential of the mathematical based analyses for experimental virology.

Keywords: virus infection, mathematical modeling, *in vitro* experiment, quantification

INTRODUCTION

The recent rapid development of experimental techniques in molecular biology and cell biology has revealed many new insights into the complexed interactions between viruses and their target cells. Most of these studies are of a qualitative nature and describe the cellular and molecular details of the interactions. To learn more about the quantitative features of virus replication, we can now generate time courses tracking the dynamics of viruses and target cells in experiments. Each of the time points during a series of experiment provides a snap-shot of the number of target cells, the number of infected cells, and the amount of virions in the culture. The whole time course reflects the results of a complex process consisting of consecutive interactions between viruses, their target cells, infected cells, and viral production. It is difficult to translate these data quantitatively into the parameters identifying the multi-composed kinetics of viral infection. To decompose and quantify the kinetics of virus infection, it will be an extremely useful to rely on mathematical modeling, mathematical analysis, and numerical simulation of the experimental data (Ho et al., 1995; Perelson et al., 1997). Modeling the whole time courses mathematically, we can estimate several parameters underlying the kinetics of virus infection (e.g., the burst size and the basic reproductive number). These parameters cannot be obtained directly by experiments only. Comparing the parameter values between viruses allows one to identify the major functional differences between viruses, and to understand why one is more virulent than the other, and why their time courses are so different. This approach is particularly useful for analyzing data from *in vitro* experiments using cell cultures, because we can nowadays obtain frequent samples of several kinetic variables in a relatively simple environment (as compared to an *in vivo* infection). Indeed, it is now possible to fully parameterize our mathematical models on such *in vitro* data, and to realize quite robust quantification of the virus infection kinetics (Mohler et al., 2005; Beauchemin et al., 2008; Iwami et al., 2012).

The importance and significance of modeling work is slowly becoming recognized in the community of experimental virologists. Starting with the landmark papers revealing the turnover of HIV-1 infected cells *in vivo* from the decline in the viral load in patients following initiation of antiretroviral therapy (Ho et al., 1995; Wei et al., 1995), mathematical modeling has evolved into an important tool in modern biology (Perelson, 2001). Here we introduce our recently developed approach to "quantitatively" describe the kinetics of virus infection in cell cultures employing the full time-course of the data. And we will discuss the potential of such approaches combining experimental and mathematical analyses to address unsolved question in virology by identifying viral parameters.

MATERIALS AND METHODS

In vitro cell culture experimental data on the infection of HSC-F cells with SHIV-KS661 were collected over time courses of 10 consecutive days. Each day most of virus (85.4%), and a small percentage of the cells (5.5%) was removed from the culture supernatant for measurement, and fresh medium was added. The measurement consisted of the concentrations of HSC-F cells positive or negative for a viral antigen, Nef, [cells/ml], and the SHIV-KS661viral load [RNA copies/ml] (**Table 1**). The experiment was repeated for two different values of the initial

Table 1 | Experimental data for the *in vitro* experiment.

MOI	Measurement day									
	0	1	2	3	4	5	6	7	8	9
CONCENTRATION OF Nef-NEGATIVE HSC-F CELLS (cells/ml)										
2×10^{-4}	6400000	6570000	6240000	4795608	4826259	1234110	463638	156560	40843	16200
2×10^{-5}	6400000	7300000	7690000	5790000	5233650	6005620	2404116	575240	231420	123641
CONCENTRATION OF Nef-POSITIVE HSC-F CELLS (cells/ml)										
2×10^{-4}	d.l.	d.l.	d.l.	15392	483741	1865890	866362	223440	69157	13800
2×10^{-5}	d.l.	d.l.	d.l.	d.l.	36350	424380	3315884	1394760	468580	46359
TOTAL VIRAL LOAD OF SHIV-KS661 (RNA copies/ml)										
2×10^{-4}	150096	2110000	12000000	322000000	7090000000	26000000000	23400000000	8430000000	1560000000	511000000
2×10^{-5}	16439	160814	621353	17700000	362000000	2180000000	21600000000	21300000000	9000000000	2360000000

d.l., designates samples in which the concentration was below the detection limit.

viral inoculum (MOI: multiplicity of infection). The time courses were analyzed with the model described below.

VIRUS INFECTION

The virus solution of SHIV-KS661 (Shinohara et al., 1999) was prepared in a $CD4^+$ human T lymphoid cell line, M8166 (a subclone of C8166) (Clapham et al., 1987), and was stored in liquid nitrogen until use. The HSC-F cell line (Akari et al., 1996) was cultured in a culture medium (RPMI-1640 supplemented with 10% fetal calf serum) at 37°C and 5% CO_2 in humidified conditions. Each experiment was performed using 2 wells of a 24-well plate with a total suspension volume of 2 ml (1 ml per well) and an initial cell concentration of $T_0 = 6.46 \times 10^6$ cells/ml in each well. Because the initial cell concentration is close to the carrying capacity of 24-well plates, and HSC-F cells replicate slowly, in the absence of SHIV-KS661 infection, the population of target cells, changes very little on the timescale of our experiment (data not shown). We therefore neglected the effects of potential regeneration of HSC-F cells when constructing the mathematical model. For virus infection, cultures of HSC-F cells were inoculated with two different MOIs [MOI 2.0×10^{-4} and MOI 2.0×10^{-5}, where a MOI of 1 means one 50% tissue culture infectious dose ($TCID_{50}$) per cell] of SHIV-KS661, and were incubated at 37°C. Four hours after inoculation, the cells were washed to remove the remaining viruses and were replaced into a fresh culture medium. The culture supernatant was harvested daily for 10 days, and was replaced with fresh medium. On a daily basis 5.5% of the cells in the culture were harvested to measure the number of target cells and infected cells. Cells were counted by staining them with an anti-SIV Nef monoclonal antibody (04–001, Santa Cruz Biotechnology, Santa Cruz, CA) labeled by Zenon Alexa Fluor 488 (Invitrogen, Carlsbad, CA), as previously described (Iwami et al., 2012). Each harvested supernatant, including 85.4% of the culture virus was stored at −80°C, and the amount of viral RNA was quantified by RT-PCR, as previously described (Iwami et al., 2012).

MATHEMATICAL MODELING

To describe the *in vitro* kinetics of virus infection, we used a classical mathematical model that is used widely for analyzing viral kinetics (Perelson and Nelson, 1999; Nowak and May, 2000; Perelson, 2001):

$$\frac{dT}{dt} = -\beta TV, \quad \frac{dI}{dt} = \beta TV - \delta I, \quad \frac{dV}{dt} = pI - cV, \qquad (1)$$

where T and I are the numbers of target (susceptible) cells, and infected (virus-producing) cells per ml of medium, respectively, and V is the number of RNA copies of virus per ml of medium. The parameters δ, c, β and p represent the death rate of infected cells, the degradation rate of viral RNA, the rate constant for infection of target cells by virus, and the viral production rate of an infected cell, respectively. The basic model (1) is a simplified version of the previously developed mathematical model in (Iwami et al., 2012), because we here use the time-course data including only viral RNA (but not both viral RNA and infectivity). A schematic of the basic model is shown in **Figure 1**.

DATA FITTING

Due to the daily harvesting of cells and virus, in our model the concentrations of target and infected cells must be reduced by 5.5% per day, and the viral loads (RNA copies) have to be reduced by 85.4% per day. We approximate these losses by adding

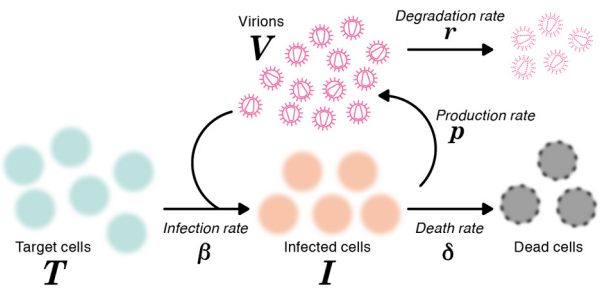


FIGURE 1 | A schematic representation of the mathematical model. The variables T and I are the number of target and infected cells, respectively, and V is the number of RNA copies of virus. The parameters δ, c, β and p represent the death rate of infected cells, the degradation rate of viral RNA, the rate constant for infection of target cells by virus, and the viral production rate of an infected cell, respectively.

continuous exponential decay terms, $-dT$, $-dI$, and $-rV$, to the equations, respectively, where $d = 0.057$ per day to account for the harvesting of cells, and $r = 1.93$ per day for the collection of virus. The degradation rate of virus was estimated to be $c = 0.039$ per day in separate experiments (data not shown). The remaining three parameters (δ, β, p), along with the 6 initial ($t = 0$) values for the variables (three for each of the two MOI values), were determined by fitting Equation (1) to the data. We simultaneously fit Equation (1) to the concentrations of Nef-negative and Nef-positive HSC-F cells and the viral loads for both MOIs, using nonlinear least-squares regression [using the FindMinimum package of *Mathematica8.0* that minimizes the sum of squared residuals (SSR)]. Experimental measurements below the detection limit were excluded when computing the SSR.

RESULTS AND DISCUSSION

In total we obtained 53 data points for quantifying the kinetics of SHIV-KS661 *in vitro* cell cultures. Using a previously established estimate for the degradation rates of RNA (c) *in vitro* culture, we estimated the values of the three remaining unknown parameters (δ, β, p) and the six initial values. The parameter estimates obtained by fitting Equation (1) to the full *in vitro* dataset simultaneously as described in "Materials and Methods" are given in **Tables 2** and **3**. These estimates are similar to our previous parameter estimates in (Iwami et al., 2012). The behavior of the model using these best-fit parameter estimates is shown together with the data in **Figure 2**, which reveals that the relatively simple model of Equation (1) describes these *in vitro* data very well. This suggests that the parameters that were estimated are representative for the various processes underlying the viral kinetics. Let us discuss what we can learn from these data.

HALF-LIFE OF INFECTED CELLS (log2/δ)

The death rate of infected cells was estimated to be $\delta = 1.75$ per day. Since in differential equations the time to death is exponentially distributed (Holder and Beauchemin, 2011), this death rate corresponds to a half-life of $\log 2/\delta = 0.40d$, and an average life-span of $1/\delta = 0.57\,d$, of productively infected HSC-F cells. Because the Nef protein is primarily produced at a late phase of the viral replication in a cell, and since we do not distinguish between an early eclipse phase and a late phase of virus production in our model, the "infected cells" that our model describes should largely correspond to cells at a relatively late stage of infection (Iwami et al., 2012). The half-life that we estimate should therefore apply primarily for infected cells at a late stage of infection, and need not apply for cells in the early eclipse phase.

BURST SIZE (p/δ)

The viral production rate of an infected cell was estimated to be $p = 3.26 \times 10^4$ RNA copies per day. Because an infected cell in the model produces virus over an average of $1/\delta$ days, the total viral burst size can be estimated as $p/\delta = 1.87 \times 10^4$ RNA copies per cell. This *in vitro* estimate is in reasonable agreement with recent *in vivo* estimates obtained using single-cycle SIV (Chen et al., 2007). The total burst size is defined as the total number of virions produced by any one infected cell during its life-time (Nowak and May, 2000; Beauchemin et al., 2008; Iwami et al., 2012) (see **Figure 3**), and is often considered as a normalized viral replication property reflecting the trade-off between viral production (p) and its cytopathic effects (δ).

BASIC REPRODUCTIVE NUMBER (R_0)

The average number of newly infected cells produced from any one infected cell, under conditions where the most of the target cells are uninfected, is known as the basic reproductive number R_0, and is an important parameter predicting the course of infection (Nowak et al., 1997; Nowak and May, 2000; Ribeiro et al., 2010). Any one infected cell produces a progeny of $p/(\delta + d)$ viruses before the cell dies, or is removed from the culture, and each produced virus will infect target cells at a constant rate β, until the virus is cleared or harvested (i.e., over $1/(r + c)$ days on average). At the beginning of the experiment there are T_0 target cells. Thus, using our parameter estimates, the reproductive number is calculated as $R_0 = \beta p T_0/[(\delta + d)(r + c)] = 5.10$ in *in vitro* culture experiments (Nowak and May, 2000; Beauchemin et al., 2008; Iwami et al., 2012) (see **Figure 3**). The basic reproductive number characterizes the course of the infection in cell

Table 3 | Fitted initial values for the *in vitro* experiment.

Variable	Unit	Fitted initial value at MOI of	
		2×10^{-4}	2×10^{-5}
T(0)	cells/ml	8.36×10^6	8.18×10^6
I(0)	cells/ml	1.13	3.45×10^{-4}
V(0)	RNA copies/ml	1.50×10^5	1.41×10^4

Table 2 | Parameters values and derived quantities.

Parameter Name	Symbol	Unit	Value
PARAMETERS OBTAINED FROM SIMULTANEOUS FIT TO FULL *in vitro* DATASET			
Rate constant for infection	β	$(\text{RNA/ml} \cdot \text{day})^{-1}$	8.61×10^{-11}
Death rate of infected cells	δ	day^{-1}	1.75
Production rate of total virus	p	RNA copies $\cdot$ day^{-1}	3.26×10^4
QUANTITIES DERIVED FROM FITTED VALUES			
Half-life of infected cells	$\log 2/\delta$	days	0.40
Viral burst size	p/δ	RNA copies	1.87×10^4
Basic reproductive number of virus	R_0	–	5.10

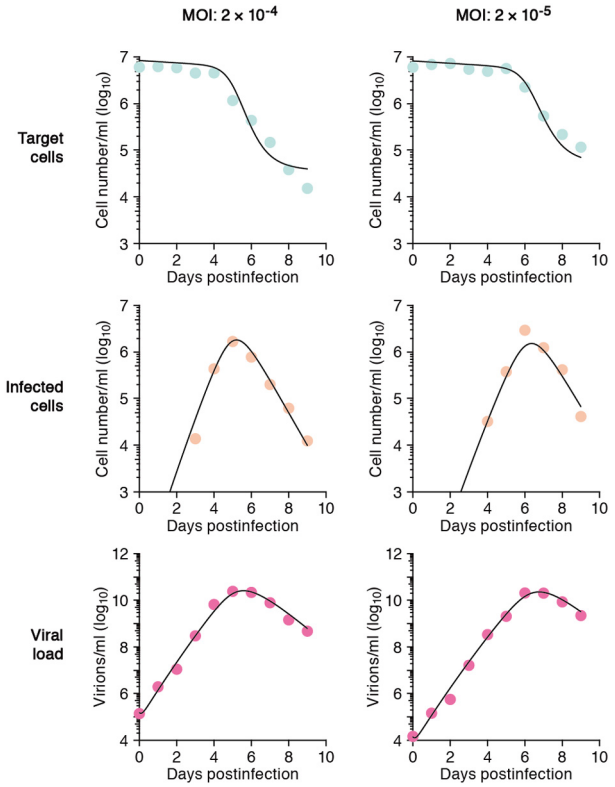


FIGURE 2 | The mathematical model describes the data well. HSC-F cells were inoculated with SHIV-KS661 24 h before $t = 0$ and each *in vitro* experimental quantity was measured daily from $t = 0$ d to 9 *d*. The curves depict the best fit of the model (Equation 1), to the experimental data of SHIV-KS661 infection *in vitro* (symbols) for the target cells, infected cells, and the viral load for the two different experiments conducted at different MOIs. All data were fitted simultaneously as described in "Materials and Methods."

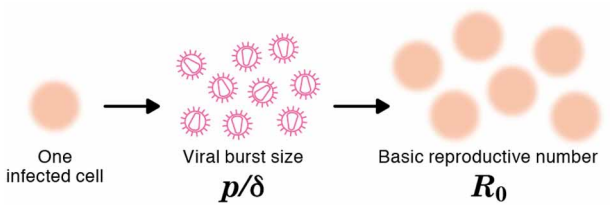


FIGURE 3 | The burst size and the basic reproductive number of a virus. The burst size is defined as the expected number of virions produced by one infected cell over its life-time (e.g., $p/\delta = 9$ in the figure). The basic reproductive number, R_0, is defined as the expected number of newly infected cells resulting from one infected cell during its life-time (e.g., $R_0 = 6$ in the figure).

culture. For example, one can predict the fraction of target cells that will be removed by the infection through the recursive relation $1 - f_I = e^{-R_0 f_I}$, which is called the "final size equation" (Anderson, 1991; Iwami et al., 2012). Here the parameter f_I corresponds to the fraction of target cells that are eventually removed by the infection (i.e., $f_I = 1 - T(\infty)/T_0$). Using the $R_0 = 5.1$ we find that $f_I = 0.9937$, and that the fraction of surviving target cells at the end of the infection should approach $1\text{-}f_I = 0.0063$. In our experiments this implies a final target cell population of approximately $T(\infty) = 4.03 \times 10^4$ cells/ml. This value agrees well with the final size of Nef-negative HSC-F cells in the MOI 2.0×10^{-4} experiment, where $T(9) = 1.62 \times 10^4$ cells/ml. Thus, the basic reproductive number provides valuable information about the expected course of infection. Note that at the MOI of 2.0×10^{-5} the infection is so slow that the final target cell value has not yet been approached at day 9 (see **Figure 2**).

CONCLUSION

Combining mathematical modeling with experimental data, we have been able to estimate several parameters defining the kinetics of SHIV-KS661 infecting HSC-F cells, from just two time courses of an *in vitro* infection. For this it was essential that we had the full time-course of the infection available for fitting the model. The data before the peak of virus infection, i.e., the up-slope of the number of viral RNA copies in the culture supernatant, reflects virus production, while the data after the peak, i.e., the down-slopes of the viral load and the infected cells, reflect the death of infected cells and viral clearance. Thus to reliably estimate the kinetic parameters, one needs to collect time-course data throughout the infection.

PERSPECTIVE

Our results of an SHIV infection in an *in vitro* cell culture are a simple example of a quantitative analysis of virus infection dynamics employing on mathematical and computational methods. Our approach can be applied regardless of viral family and genus. To further explore how our approach of modeling time-course data can be used in future work we will discuss a number of hypothetical examples, emphasizing how quantitative estimates can be used to address unsolved question in virology.

IDENTIFYING THE MAJOR DIFFERENCES AMONG SEVERAL VIRAL STRAINS

After fitting time-course data from different virus strains, one can compare the estimated parameters of each viral strain, such as its half-life of infected cells, burst size and basic reproductive number, to reveal the quantitatively largest differences between the strains (Mitchell et al., 2011). For example, let us denote SHIV-KS661 as "virus-A," which brings about a half-life of 0.4 days in infected cells, and consider a less cytopathic variant "virus-B" extending the half-life 1.5-fold to 0.6 days. The expected time-courses of these two variants are depicted in **Figure 4**, and reveal a major difference in the target cell dynamics and minor differences in the number of infected cells and the viral load (compare the solid line with the dashed lines, depicting virus A and B, respectively). If one were to fit the *in silico* data from virus A and B in **Figure 4** with our mathematical model, one would correctly conclude that the half-life of infected cells of virus-B is 1.5 times longer than that of cells infected with virus-A, and therefore that virus B is less cytopathic than virus A. It is difficult to arrive at that result by just visual inspection of the data, however. The effect of cytopathicity on the time courses of target cells, infected cells, and virus load are difficult to predict intuitively. Additionally, there is no experimental technique available

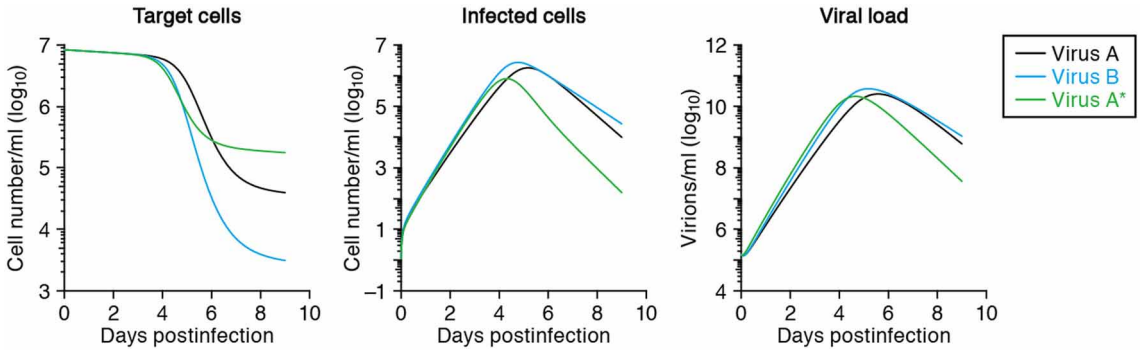


FIGURE 4 | Predicted virus infection kinetics with different parameters. The solid curves show the predicted kinetics of target cells (left), infected cells (middle) and virus load (right) following infection with SHIV-KS661 at an MOI 2×10^{-4} using the parameters in **Tables 2** and **3**. The time $t = 0$ corresponds to 24 h after inoculation of SHIV-KS661 to HSC-F. In the text we call this virus-A. The dashed and dotted curves show those of the variants virus-B and virus-A*, respectively. Virus-B is less cytopathic and has a 1.5-fold decreased death rate of infected cells ($\delta = 1.16$). Cells infected with the more virulent "mutant" virus-A* die 3-fold faster than those infected with the wild-type A (i.e., $\delta = 5.25$ vs. $\delta = 1.75$, respectively), but produce 2-fold more virus ($p = 6.72 \times 10^4$) per day than those infected with "wild-type" virus-A.

to measure quantities like the cytopathicity directly as an absolute value. For the production rate, the burst size and the basic reproductive number of the virus, similar arguments apply, and one has to rely on modeling to identify quantitative differences between viral strains.

IDENTIFYING THE FUNCTION OF VIRAL PROTEINS OR AMINO ACIDS IN INFECTION

Using molecular cell biology techniques, we are currently able to investigate the function of individual viral proteins in several aspects of viral replication. However, it remains difficult to interconnect those particular results and to integrate the roles of different molecules in terms of the overall parameters defining a virus infection, like a replication rate or a burst size. By modeling *in vitro* time courses, and comparing the estimated parameters between a wild-type virus and several particular mutants, one can quantify the role of every amino acid mutation on the several parameters defining a virus infection. For example, if one were to take SHIV-KS661 as a "wild-type virus-A," with an estimated half-life of infected cells of 0.40 days and a viral production rate of 3.26×10^4 RNA copies per day (see **Table 2**), and find by fitting that a more virulent mutant "virus-A*" has a 3-fold shorter half-life of its infected cells, but a 2.0-fold increased production rate of 6.52×10^4 RNA copies per day, one would be able to conclude that this particular mutation decreases the total viral production per generation to 2/3 of that of the wild-type. Thus, the more virulent virus is less fit, i.e., has a lower R_0, because the total burst sizes of virus-A and virus-A* are 1.87×10^4 and 1.25×10^4 RNA copies per generation, respectively. For the function of the mutated protein one would be able to conclude that it plays a role in the production of novel viral particles, and that increased production apparently brings about a shorter expected life-span of infected cells. The solid and dotted curves in **Figure 4** show the virus kinetics predicted by Equation (1) for virus-A (solid line) and virus-A* (dotted line), respectively. Similar approaches allow us to also investigate the functions of multiple mutations in possibly several proteins quantitatively.

FINDING THE TARGET OF NOVEL ANTIVIRAL COMPOUNDS

Calculating and comparing parameter estimates in the absence and presence of an antiviral compound, allows one to investigate the function of the compound in a very similar manner (Baccam et al., 2006; Beauchemin et al., 2008). For instance, if the daily viral production rate is reduced to half but the half-life of infected cells has remained similar, one concludes that the compound inhibits viral production without affecting cytopathicity. In addition, if a dose-dependent basic reproductive number, R_o, were obtained, one would estimate how effectively the compound inhibits total viral replication. From the value of $1 - 1/R_0$ (Anderson, 1991), one can calculate the critical compound concentration at which the infection should die out. Note that this value is not the same as the conventional IC_{50}, the half maximal (50%) inhibitory concentration. Identifying the precise mode of action of novel compounds may help the development of novel antiviral drugs.

FUTURE DIRECTION

As discussed above, an approach of combining experiments with mathematical modeling has broad applications in virology. One possible extension of our model is to also consider the "eclipse" phase of the infection of a cell to allow for a period in which no virus is produced and the cell may have a different death rate (Baccam et al., 2006; Beauchemin et al., 2008; Iwami et al., 2012). Another extension is to divide the viral population into infectious and non-infectious virus, because the virus that is produced by most of the cells is non-infectious (Schulze-Horsel et al., 2009; Iwami et al., 2012). In Equation (1), it is assumed all virus is infectious, and the non-infectious fraction is in fact incorporated by a lower infection rate β. If one were to have data on the amounts of infectious and non-infectious virions, and/or on the fraction of infected cells in the eclipse phase, one can extend the mathematical model accordingly and obtain even more detailed quantification of the characteristics of any virus studied in particular culture conditions. Furthermore, it is challenging but very interesting to distinguish cell-free and cell-to-cell infection,

which are two different mode of viral infection, and to quantify the efficacy of each mode. Sourisseau et al. reported that in a continuously shaken culture in a HIV replication assay cell-to-cell infection is blocked (Sourisseau et al., 2007). Combining a novel mathematical model including both a cell-free and a cell-to-cell infection mode, and fitting that to shaken and non-shaken HIV replication assays, we might be able to quantitatively identify the two infection modes. Summarizing, our method of modeling time courses of viral infection is effectively capable of describing the data, and therefore provides a new approach of characterizing and comparing viruses in a quantitative manner to better understand their infection kinetics under *in vivo* circumstances.

ACKNOWLEDGMENTS

This work was supported by JST PRESTO program (Shingo Iwami) and by the Aihara Innovative Mathematical Modeling Project, the Japan Society for the Promotion of Science (JSPS) through the "Funding Program for World-Leading Innovative R & D on Science and Technology (FIRST Program)," initiated by Council for Science and Technology Policy (CSTP) (Shingo Iwami, Kei Sato, and Kazuyuki Aihara).

REFERENCES

Akari, H., Mori, K., Terao, K., Otani, I., Fukasawa, M., Mukai, R., and Yoshikawa, Y. (1996). *In vitro* immortalization of Old World monkey T lymphocytes with Herpesvirus saimiri: its susceptibility to infection with simian immunodeficiency viruses. *Virology* 218, 382–388.

Anderson, R. M. (1991). The Kermack-McKendrick epidemic threshold theorem. *Bull. Math. Biol.* 53, 3–32.

Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G., and Perelson, A. S. (2006). Kinetics of influenza A virus infection in humans. *J. Virol.* 80, 7590–7599.

Beauchemin, C. A., McSharry, J. J., Drusano, G. L., Nguyen, J. T., Went, G. T., Ribeiro, R. M., and Perelson, A. S. (2008). Modeling amantadine treatment of influenza A virus *in vitro*. *J. Theor. Biol.* 254, 439–451.

Chen, H. Y., Di Mascio, M., Perelson, A. S., Ho, D. D., and Zhang, L. (2007). Determination of virus burst size *in vivo* using a single-cycle SIV in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19079–19084.

Clapham, P. R., Weiss, R. A., Dalgleish, A. G., Exley, M., Whitby, D., and Hogg, N. (1987). Human immunodeficiency virus infection of monocytic and T-lymphocytic cells receptor modulation and differentiation induced by phorbol ester. *Virology* 158, 44–51.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* 373, 123–126.

Holder, B. P., and Beauchemin, C. A. A. (2011). Exploring the effect of biological delays in kinetic models of influenza within a host or cell culture. *BMC Public Health* 11, S10.

Iwami, S., Holder, B. P., Beauchemin, C. A., Morita, S., Tada, T., Sato, K., Igarashi, T., and Miura, T. (2012). Quantification system for the viral dynamics of a highly pathogenic simian/human immunodeficiency virus based on an *in vitro* experiment and a mathematical model. *Retrovirology* 9, 18.

Mitchell, H., Levin, D., Forrest, S., Beauchemin, C. A., Tipper, J., Knight, J., Donart, N., Layton, R. C., Pyles, J., Gao, P., Harrod, K. S., Perelson, A. S., and Koster, F. (2011). Higher level of replication efficiency of 2009 (H1N1) pandemic influenza virus than those of seasonal and avian strains: kinetics from epithelial cell culture and computational modeling. *J. Virol.* 85, 1125–1135.

Mohler, L., Flockerzi, D., Sann, H., and Reichl, U. (2005). Mathematical model of influenza A virus production in large-scale microcarrier culture. *Biotechnol. Bioeng.* 90, 46–58.

Nowak, M. A., Lloyd, A. L., Vasquez, G. M., Wiltrout, T. A., Wahl, L. M., Bischofberger, N., Williams, J., Kinter, A., Fauci, A. S., Hirsch, V. M., and Lifson, J. D. (1997). Viral dynamics of primary viremia and antiretroviral therapy in simian immunodeficiency virus infection, *J. Virol.* 71, 7518–7525.

Nowak, M. A., and May, R. M. (2000). *Virus Dynamics*. New York, NY: Oxford University Press.

Perelson, A. S. (2001). Modelling viral and immune system dynamics. *Nat. Rev. Immunol.* 2, 28–36.

Perelson, A. S., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M., and Ho, D. D. (1997). Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* 387, 188–191.

Perelson, A. S., and Nelson, P. W. (1999). Mathematical analysis of HIV-1 dynamics *in vivo*. *SIAM Rev.* 41, 3–44.

Ribeiro, R. M., Qin, L., Chavez, L. L., Li, D., Self, S. G., and Perelson, A. S. (2010). Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. *J. Virol.* 84, 6096–6102.

Schulze-Horsel, J., Schulze, M., Agalaridis, G., Genzel, Y., and Reichl, U. (2009). Infection dynamics and virus-induced apoptosis in cell culture-based influenza vaccine production—Flow cytometry and mathematical modeling. *Vaccine* 27, 2712–2722.

Shinohara, K., Sakai, K., Ando, S., Ami, Y., Yoshino, N., Takahashi, E., Someya, K., Suzaki, Y., Nakasone, T., Sasaki, Y., Kaizu, M., Lu, Y., and Honda, M. (1999). A highly pathogenic simian/human immunodeficiency virus with genetic changes in cynomolgus monkey. *J. Gen. Virol.* 80, 1231–1240.

Sourisseau, M., Sol-Foulon, N., Porrot, F., Blanchet, F., and Schwartz, O. (2007). Inefficient human immunodeficiency virus replication in mobile lymphocytes. *J. Virol.* 81, 1000–1012.

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. S., and Shaw, G. M. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 373, 117–122.

Conflict of Interest Statement: 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.

Received: 14 July 2012; accepted: 16 August 2012; published online: 04 September 2012.

*Citation: Iwami S, Sato K, De Boer RJ, Aihara K, Miura T and Koyanagi Y (2012) Identifying viral parameters from in vitro cell cultures. Front. Microbio. **3**:319. doi: 10.3389/fmicb.2012.00319*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

MINI REVIEW ARTICLE
published: 17 October 2012
doi: 10.3389/fmicb.2012.00369

Functional constraints on HIV-1 capsid: their impacts on the viral immune escape potency

Taichiro Takemura* and Tsutomu Murakami

AIDS Research Center, National Institute of Infectious Diseases, Tokyo, Japan

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Masako Nomaguchi, The University of Tokushima Graduate School, Japan
Anjali Joshi, Texas Tech University Helath Sciences Center, USA

****Correspondence:***
Taichiro Takemura, AIDS Research Center, National Institute of Infectious Diseases, 1-23-1 Toyama Shinjuku-ku, Tokyo 162-8640, Japan.
e-mail: taichiro@nih.go.jp

In mature HIV-1 particles, viral capsid (CA) proteins form the conical core structure that encapsidates two copies of the viral RNA genome. After fusion of the viral envelope and cellular membranes, the CA core enters into the cytoplasm of the target cells. CA proteins then interact with a variety of viral other protein as well as host factors, which may either support or inhibit replication of the virus. Recent studies have revealed that CA proteins are important not only for the uncoating step but also for the later nuclear import step. Identification of proteins that interact with CA to fulfill these functions is, therefore, important for understanding the unknown HIV-1 replication machinery. CA proteins can also be targets of the host immune response. Notably, some HLA-restricted cytotoxic T-lymphocyte (CTL) responses that recognize CA functional regions can greatly contribute to delay in AIDS progression. The multi-functionality of the CA protein may limit the flexible virus evolution and reduce the possibility of an escape mutant arising. The presence of many functional regions in CA protein may make it a potential target for effective therapies.

Keywords: HIV-1, capsid, host factor, immune response, functional constraints

INTRODUCTION

The HIV-1 *gag* gene encodes the Gag protein, major structural component of virus particles (Vogt, 1997; Scarlata and Carter, 2003; Engelman and Cherepanov, 2012). The Gag protein consists of six functionally different proteins. Capsid (CA) is the largest component of Gag protein, and forms core structure of the mature HIV-1 particle. Recent studies have revealed that the CA protein has multiple roles in the virus-host interaction at the cellular or individual levels. In this mini-review, we are summarizing the interaction of the CA and host cellular proteins such as cyclophilin A (CypA), Nuclear pore proteins (Nups), TRIM5alpha, and/or host immune response.

HIV-1 CAPSID PROTEINS CONSTITUTE THE VIRAL CORE STRUCTURE

The HIV-1 Gag proteins are synthesized as Pr55Gag polyprotein in cytoplasm of the virus-producing cell, and are then translocated to the plasma membrane. Subsequently, they are co-assembled into virus particles, which bud and are then released from the plasma membrane. Right after the budding and release step from the virus producing cells, virus particles undergo a process of maturation (Morikawa, 2003). The viral protease (PR) cleaves the Gag polyprotein into six proteins: matrix (MA), CA, nucleocapsid (NC), p6, p2, and p1. Virus morphology dramatically changes as a result of the maturation process. In the mature virus particles, CA proteins form a conical core structure encapsidating two copies of the viral RNA genome associated by NC proteins. The CA protein has N-terminal and C-terminal domains (NTD and CTD, respectively), and the short flexible linker connects between the two regions (Gitti et al., 1996; Momany et al., 1996; Gamble et al., 1997). The CA proteins assemble into the small units containing five or six monomers (**Figure 1A**) (Li et al., 2000; Pornillos et al., 2009, 2011; Yeager, 2011). The core structure is composed of approximately 250 units of hexamers, and 12 units of pentamers at the both conical ends.

The mature virus particles can then infect to the new target cells. The binding of the Env and receptors/co-receptors leads to fusion of the viral envelope and cellular membranes, and subsequently, the CA core enters the cytoplasm of the target cell. The CA core interacts with a variety of cellular proteins at this step (Mascarenhas and Musier-Forsyth, 2009). Before penetration into new target cells, the viral core should be stable to protect the viral RNA genome from the outer environment (Koh et al., 2000). However, after penetration, the core must be destabilized to uncoat and release the viral genome at the correct time for replication. Although spatial and temporal regulation of the uncoating process is not yet well understood, this process presumably depends on the interaction of CA with several host factors. The identification of such host proteins is, therefore, likely to be essential for understanding the HIV-1 replication cycle.

INTERACTION OF CA AND HOST FACTORS

CYCLOPHILIN A

Cyclophilin A (CypA) is a host cellular protein that carries peptidyl-prolyl cis-trans isomerase (PPIase) activity and is abundantly expressed in various types of cell, including T-lymphocytes. CypA incorporates into HIV-1 particles via interaction with pr55Gag protein (Luban et al., 1993; Franke et al., 1994; Luban, 1996), binding at a site in the CA protein NTD, termed the CypA-binding loop (**Figures 1B** and **2**). The CypA-binding loop is a proline-rich loop located between helices 4 and 5 in the CA protein, with the proline residue at position 90 considered to be the most important amino acid for CypA-binding (Grättinger et al., 1999). Interference in CA-CypA-binding reduces HIV-1

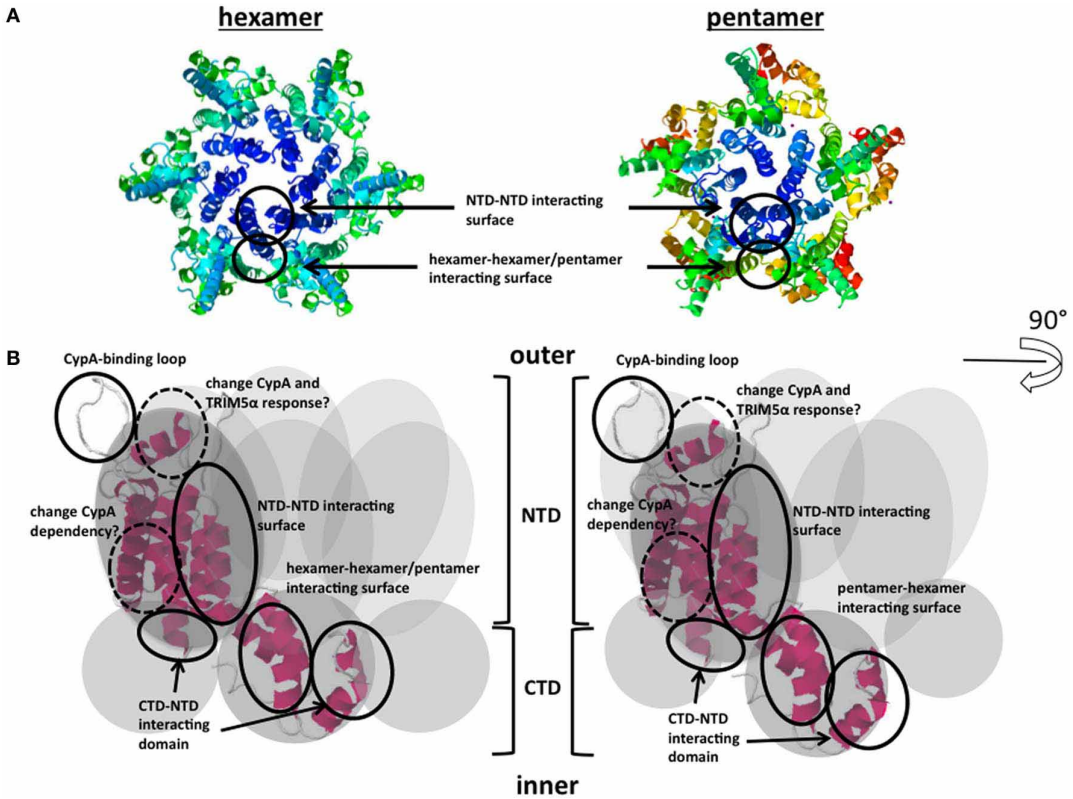


FIGURE 1 | CA proteins form hexamers and pentamers and schematic model of CA interacting domains. (A) Hexameric/pentameric models proposed from structural analysis of the HIV-1 core. Each core comprises approximately 250 hexamers and exactly 12 pentamers. The accession numbers of protein Databases are 3H47, 3P05. **(B)** Functional surfaces in the CA protein. The gray circles indicate monomers of CA. NTD-NTD or CTD-NTD interacting surfaces, CypA-binding loops, and hexamer-hexamer/pentamer interacting surfaces are shown. CA; viral capsid; NTD; N-terminal domain, CTD; C-terminal domain. The accession numbers of protein Database are 3DIK. The detailed map including each component is shown in **Figure 2**.

infection efficiency (Franke et al., 1994; Luban, 1996). Initial studies to determine the role of CypA in HIV-1 infection focused on CypA incorporated into virions, but later studies showed that the target cellular CypA is more important for HIV-1 infection (Sokolskaja et al., 2004; Hatziioannou et al., 2005). The exact role of CypA in the HIV-1 replication is still unclear and remains a subject of debate (Luban, 2007; Mascarenhas and Musier-Forsyth, 2009). One hypothesis is that the proline isomerase activity induces a conformational change in the core structure, which results in its destabilization and efficient viral uncoating in the target cell cytoplasm. Interestingly, CypA is required for the replication HIV-1 group M virus, but not always necessary for the HIV-1 group O or other simian immunodefficiency virus (SIV) replication (Braaten et al., 1996). Besides, Takeuchi et al. (2012) reported the opposing response to the CypA in HIV or SIV infection in different host cell species. Identification of the molecular basis of CypA-dependent HIV-1 infection may also contribute to understand the evolution of the HIV-1.

Nup358 AND OTHER FACTORS SUPPORTING NUCLEAR IMPORT

The lentiviruses, including HIV-1, are able to infect non-dividing cells such as macrophages. The viral genome, consisting the pre-integration complex (PIC), must move into the nucleus of the target cell even in the non-dividing state, but molecular mechanism of the nuclear import step has yet to be determined. Numerous studies have examined the requirements for specific cellular proteins during HIV-1 nuclear import, including nuclear transport proteins such as some importin family proteins and Nups. It has been proposed that the PIC goes through the nuclear pore on the nuclear membrane (De Iaco and Luban, 2011). One of the Nups, Nup358 (also known as RanBP2 or RAN binding protein 2), is a cellular co-factor for the HIV-1 infection at the PIC nuclear import step (Hutten et al., 2009; Zhang et al., 2010; Ocwieja et al., 2011; Schaller et al., 2011). Nup358 is a relatively large protein (358 kDa) located on the cytoplasmic surface of the nuclear pore. Nup358 controls the cell cycle, nuclear export, and transportin-dependent nuclear import. Interestingly, Nup358 has a cyclophilin-like domain on its C-terminal end. Schaller et al. (2011) revealed that Nup358 directly binds to HIV-1 CA protein, with the viral acceptor being the CypA-binding loop described above. The knockdown of Nup358 in target cells impairs HIV-1 infection at the nuclear import step, and the migration of the PIC through the nuclear pore depends on interaction of Nup358 and CA (Schaller et al., 2011). Although the direct interaction of CA and Nup358 has now been identified, it is still unclear how Nup358 supports HIV-1 nuclear

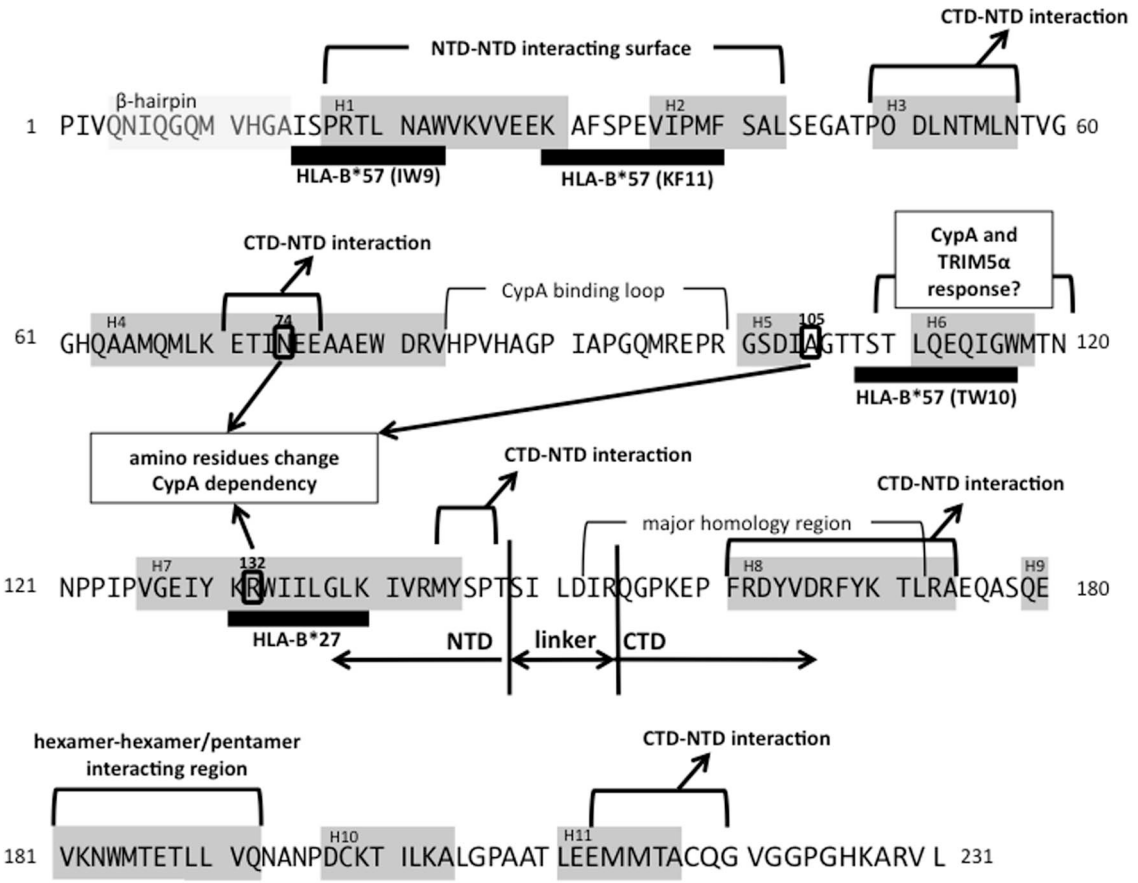


FIGURE 2 | Mapping of CA functional domains and two major protective CTL alleles. Gray boxes show β-hairpin (amino acids 4–14) or helix structures (H1–H11). Thick black bars indicate the CTL epitopes restricted by each allele (Llano et al., 2009). The NTD-NTD, CTD-NTD, or hexamer-hexamer/pentamer interacting regions are indicated. Three amino residues (N74, A105, and R132), which are reported to change CypA dependency, are shown in circles (Schneidewind et al., 2007; Yang and Aiken, 2007; Ambrose et al., 2012). The reference amino acids sequences are from HXB2 (accession number; AB50258.1).

import. Furthermore, transportin-SR2 (TNPO3) was identified as a HIV-1 co-factor that supports the nuclear import step by a series of genome-wide siRNA screens, and TNPO3 interacts with CA (Brass et al., 2008; Krishnan et al., 2010; Lee et al., 2010; De Iaco and Luban, 2011). These observations suggest that the functional analysis of CA should be expanded to include nuclear import steps. Also, a recent study proposed a model in which reverse transcription and uncoating processes are regulated by each other or occurring at almost same time, at the perinuclear location in the target cells (Hulme et al., 2011). Further analysis into the role of CA in these steps, and their interactions with other host proteins including Nups should help to clarify the molecular mechanism of HIV-1 replication.

TRIM5alpha

HIV-1 infection is strikingly restricted in the non-human primate cells. It had been predicted that the CA targeting dominant-acting inhibitory factor(s) expressed in these cells (Kootstra et al., 2003; Goff, 2004; Towers, 2007; Luban, 2012). Stremlau et al. (2004) identified TRIM5alpha as a species-specific HIV-1 restriction factor from the cDNA library of rhesus monkey cells. The other ortholog of TRIM5 gene (known as TRIM-Cyp), which restricts HIV-1 infection, was also identified from owl monkey cells (Nisole et al., 2004; Sayah et al., 2004). TRIM5alpha is a member of the tripartite motif-containing protein family, and the N-teminal domain has E3 ubiquitin ligase activity (Diaz-Griffero et al., 2006; Stremlau et al., 2006; Luban, 2012). Since the TRIM-Cyp carries a CypA-like domain in the C-terminal, it can targets the HIV-1 CA via the CypA-binding loop (Sayah et al., 2004). Although, the C-terminal domain of TRIM5alpha also has a function to recognize CA, its mechanism had not been clarified despite of the numerous studies. Ganser-Pornillos et al. (2011) proposed the TRIM5alpha lattice model, in which a spontaneously formed cellular TRIM5alpha hexameric lattice recognizes a surface of the incoming CA core. The detailed mechanism of TRIM5alpha- or TRIM-Cyp-mediated HIV-1 restriction after recognition of the core structure is still unclear. Two distinct restriction stages have been observed in both TRIM5alpha- or TRIM-Cyp-mediated HIV-1 restriction (Anderson et al., 2006; Yap et al., 2006). The unknown host factor may contribute to their restriction mechanism.

CA PROTEIN AND THE HOST IMMUNE RESPONSE

CAN THE CA ACTIVATE HOST INNATE IMMUNE SENSOR?

Manel et al. (2010) reported that the interaction of CypA and newly synthesized Gag proteins induces the type I interferon response to activate T-cells. They showed that the CypA-Gag interaction in dendritic cells activates the IRF3 pathway. This was the first report of an interaction between HIV-1 Gag (or CA) and CypA playing a role in inducing the host innate immune response. More recently, Pertel et al. (2011) showed that the interaction of host cellular TRIM5 and CA stimulates AP-1 and NF-κB signaling via the TAK-1 (also known as MAP3K7) pathway. The proposed mechanism for this immune sensor is rather complicated. In the presence of the heterodimeric E2 ubiquitin-conjugating enzyme, UBC13/UEV1A, TRIM5alpha catalyzes the synthesis of unattached K63-linked ubiquitin chains. The free K63-linked ubiquitin chain activates the TAK1 kinase complex, and the TAK-1-mediated signal then activates the inflammatory cytokine transcription via the NF-κB and AP-1 pathway. These two findings suggest that the Gag protein or core structure may be recognized by the TRIM5alpha or CypA that is acting as a pathogen recognition receptor for the host innate immune response. Further work in this area may contribute to the development of new therapeutic strategies utilizing the host immune response.

MAJOR PROTECTIVE CYTOTOXIC T-LYMPHOCYTE (CTL) ALLELES TARGETING CA PROTEIN

The host cytotoxic T-lymphocyte (CTL) response is a major effector to control HIV-1 replication *in vivo* (Borrow et al., 1994; Goulder and Watkins, 2008). CD8+ CTLs recognize the antigenic peptides in the context of the class I major histocompatibility complex (MHC). The CTL escape mutations occurring HIV-1 infection has been well documented (Leslie et al., 2004; Gao et al., 2005). In many cases, the replication fitness of the escape mutants is lower than that of the parental viruses (Leslie et al., 2004). Since the CTLs recognize the target viral peptides presented by the MHC on the surface of the virus-infected cells, the efficiency of the CTL response closely depends on the host MHC (human leukocyte antigen, HLA, in human) alleles. Numerous studies show that some HLA alleles, such as HLA-B*27 and HLA-B*57, have stronger protective effects than that of other HLA alleles (Carrington and O'Brien, 2003; Gao et al., 2005; Goulder and Watkins, 2008). HIV-1-specific CD8+ T-cell responses restricted by these alleles provide a probable mechanism for the protection of HIV-1 infected carriers from disease progression. Notably, HLA-B*27 and HLA-B*57 targeting epitopes are located in the viral CA protein (mapped in **Figure 2**). The major escape variant from HLA-B*27, R132K, has been altered CypA dependency for viral infection (Schneidewind et al., 2007). The three major highly conserved epitopes of HLA-B*57, IW9, KF11, and TW10 are located distant regions in CA. In these epitopes, IW9 and KF11 locate the region that has been identified as the CA NTD-NTD interacting surface. Escape mutations in those regions may confer a structural disadvantage on the virus, one which impairs infectivity (Llano et al., 2009; Pornillos et al., 2009; Brennan et al., 2012). TW10 locates the outer surface of the CA core structure, and Brockman et al. (2007) showed that escape mutants in this region also alter the CypA dependency for its replication. In addition, Battivelli et al. (2011) reported that the mutations in TW10 increase the sensitivity to the potent host cellular restriction factor TRIM5alpha. Thus, escape mutations in CA often induce impairment of viral replication as a result of a failure of interactions with viral or host factors. More generally, it could be said that the interaction between CA and viral or host proteins limits the evolutionary flexibility of the virus. The CTL efficiency is not solely depending on the choice of the epitopes. Even though various factors impact on effective CTL responses such as presentation efficiency of the epitopes, the low evolutional flexibility of the target epitope regions could be important factor for selecting effective CTL responses.

PERSPECTIVE

For optimization of anti-HIV-1 therapies, much remains to be studied about the function of each HIV-1 encoded protein and its interacting partner. It is widely known that escape mutants arising under selective pressure from the host immune response or anti-viral therapies often lose the fitness for replication. Recent advances in structural and genomic analyses expand our understanding of viral and host protein interactions. When the host immune response targets such a multifunctional protein, like CA, the possibilities for generating a successful escape mutation may be limited. In fact, two major protective CTL alleles target a region that is functionally important for HIV-1 replication. It can be considered that the multifunctionality of the protein limits the "robustness" of the HIV-1 as a living organism. From this point of view, further functional analysis of the CA protein, as well as CA-host protein interactions, may contribute to establish better therapies for HIV-1.

REFERENCES

Ambrose, Z., Lee, K., Ndjomou, J., Xu, H., Oztop, I., Matous, J., et al. (2012). Human immunodeficiency virus type 1 capsid mutation N74D alters cyclophilin A dependence and impairs macrophage infection. *J. Virol.* 86, 4708–4714.

Anderson, J. L., Campbell, E. M., Wu, X., Vandegraaff, N., Engelman, A., and Hope, T. J. (2006). Proteasome inhibition reveals that a functional preintegration complex intermediate can be generated during restriction by diverse TRIM5 proteins. *J. Virol.* 80, 9754–9760.

Battivelli, E., Migraine, J., Lecossier, D., Yeni, P., Clavel, F., and Hance, A. J. (2011). Gag cytotoxic T lymphocyte escape mutations can increase sensitivity of HIV-1 to human TRIM5alpha, linking intrinsic and acquired immunity. *J. Virol.* 85, 11846–11854.

Braaten, D., Franke, E. K., and Luban, J. (1996). 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.

Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., et al. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921–926.

Brennan, C. A., Ibarrondo, F. J., Sugar, C. A., Hausner, M. A., Shih, R., Ng, H. L., et al. (2012). Early HLA-B*57-restricted CD8+ T lymphocyte responses predict HIV-1 disease progression. *J. Virol.* 86, 10505–10516.

Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M., and Oldstone, M.

B. (1994). Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* 68, 6103–6110.

Brockman, M. A., Schneidewind, A., Lahaie, M., Schmidt, A., Miura, T., Desouza, I., et al. (2007). Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A. *J. Virol.* 81, 12608–12618.

Carrington, M., and O'Brien, S. J. (2003). The influence of HLA genotype on AIDS. *Annu. Rev. Med.* 54, 535–551.

De Iaco, A., and Luban, J. (2011). Inhibition of HIV-1 infection by TNPO3 depletion is determined by capsid and detectable after viral cDNA enters the nucleus. *Retrovirology* 6, 8–98.

Diaz-Griffero, F., Li, X., Javanbakht, H., Song, B., Welikala, S., Stremlau, M., et al. (2006). Rapid turnover and polyubiquitylation of the retroviral restriction factor TRIM5. *Virology* 349, 300–315.

Engelman, A., and Cherepanov, P. (2012). The structural biology of HIV-1, mechanistic and therapeutic insights. *Nat. Rev. Microbiol.* 10, 279–290.

Franke, E. K., Yuan, H. E., and Luban, J. (1994). Specific incorporation of cyclophilin A into HIV-1 virions. *Nature* 372, 359–362.

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.

Ganser-Pornillos, B. K., Chandrasekaran, V., Pornillos, O., Sodroski, J., Sundquist, W. I., and Yeager, M. (2011). Hexagonal assembly of a restricting TRIM5alpha protein. *Proc. Natl. Acad. Sci. U.S.A.* 108, 534–539.

Gao, X., Bashirova, A., Iversen, A. K., Phair, J., Goedert, J. J., Buchbinder, S., et al. (2005). AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis. *Nat. Med.* 11, 1290–1292.

Gitti, R. K., Lee, B. M., Walker, J., Summers, M. F., Yoo, S., and Sundquist, W. I. (1996). Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science* 273, 231–235.

Goff, S. P. (2004). Retrovirus restriction factors. *Mol. Cell* 16, 849–859.

Goulder, P. J., and Watkins, D. I. (2008). Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nat. Rev. Immunol.* 8, 619–630.

Grättinger, M., Hohenberg, H., Thomas, D., Wilk, T., Müller, B., and Kräusslich, H. G. (1999). *In vitro* assembly properties of wild-type and cyclophilin-binding defective human immunodeficiency virus capsid proteins in the presence and absence of cyclophilin A. *Virology* 257, 247–260.

Hatziioannou, T., Perez-Caballero, D., Cowan, S., and Bieniasz, P. D. (2005). Cyclophilin interactions with incoming human immunodeficiency virus type 1 capsids with opposing effects on infectivity in human cells. *J. Virol.* 79, 176–183.

Hulme, A. E., Perez, O., and Hope, T. J. (2011). Complementary assays reveal a relationship between HIV-1 uncoating and reverse transcription. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9975–9980.

Hutten, S., Wälde, S., Spillner, C., Hauber, J., and Kehlenbach, R. H. (2009). The nuclear pore component Nup358 promotes transportin-dependent nuclear import. *J. Cell Sci.* 122, 1100–1110.

Koh, K., Miyaura, M., Yoshida, A., Sakurai, A., Fujita, M., and Adachi, A. (2000). Cell-dependent gag mutants of HIV-1 are crucially defective at the stage of uncoating/reverse transcription in non-permissive cells. *Microbes Infect.* 2, 1419–1423.

Kootstra, N. A., Munk, C., Tonnu, N., Landau, N. R., and Verma, I. M. (2003). Abrogation of postentry restriction of HIV-1-based lentiviral vector transduction in simian cells. *Proc. Natl. Acad. Sci. U.S.A.* 100, 1298–1303.

Krishnan, L., Matreyek, K. A., Oztop, I., Lee, K., Tipper, C. H., Li, X., et al. (2010). The requirement for cellular transportin 3 (TNPO3 or TRN-SR2) during infection maps to human immunodeficiency virus type 1 capsid and not integrase. *J. Virol.* 84, 397–406.

Lee, K., Ambrose, Z., Martin, T. D., Oztop, I., Mulky, A., Julias, J. G., et al. (2010). Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* 7, 221–233.

Leslie, A. J., Pfafferott, K. J., Chetty, P., Draenert, R., Addo, M. M., Feeney, M., et al. (2004). HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.* 10, 282–289.

Li, S., Hill, C. P., Sundquist, W. I., and Finch, J. T. (2000). Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* 407, 409–413.

Llano, A., Frahm, N., and Brander, C. (2009). "How to optimally define optimal Cytotoxic T lymphocyte epitopes in HIV infection?" in *HIV Molecular Immunology*, eds C. Kuiken, T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, S. Wolinsky, and B. Korber (Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory), I-3–5.

Luban, J. (1996). Absconding with the chaperone: essential cyclophilin-Gag interaction in HIV-1 virions. *Cell* 87, 1157–1159.

Luban, J. (2007). Cyclophilin A, TRIM5, and resistance to human immunodeficiency virus type 1 infection. *J. Virol.* 81, 1054–1061.

Luban, J. (2012). TRIM5 and the regulation of HIV-1 infectivity. *Mol. Biol. Int.* 2012, 426840.

Luban, J., Bossolt, K. L., Franke, E. K., Kalpana, G. V., and Goff, S. P. (1993). Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. *Cell* 73, 1067–1078.

Manel, N., Hogstad, B., Wang, Y., Levy, D. E., Unutmaz, D., and Littman, D. R. (2010). A cryptic sensor for HIV-1 activates antiviral innate immunity in dendritic cells. *Nature* 467, 214–217.

Mascarenhas, A. P., and Musier-Forsyth, K. (2009). The capsid protein of human immunodeficiency virus: interactions of HIV-1 capsid with host protein factors. *FEBS J.* 276, 6118–6127.

Momany, C., Kovari, L. C., Prongay, A. J., Keller, W., Gitti, R. K., Lee, B. M., et al. (1996). Crystal structure of dimeric HIV-1 capsid protein. *Nat. Struct. Biol.* 3, 763–770.

Morikawa, Y. (2003). HIV capsid assembly. *Curr. HIV Res.* 1, 1–14.

Nisole, S., Lynch, C., Stoye, J. P., and Yap, M. W. (2004). A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. *Proc. Natl. Acad. Sci. U.S.A.* 101, 13324–13328.

Ocwieja, K. E., Brady, T. L., Ronen, K., Huegel, A., Roth, S. L., Schaller, T., et al. (2011). HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog.* 7:e1001313. doi: 10.1371/journal.ppat.1001313

Pertel, T., Hausmann, S., Morger, D., Züger, S., Guerra, J., Lascano, J., et al. (2011). TRIM5 is an innate immune sensor for the retrovirus capsid lattice. *Nature* 472, 361–365.

Pornillos, O., Ganser-Pornillos, B. K., Kelly, B. N., Hua, Y., Whitby, F. G., Stout, C. D., et al. (2009). X-ray structures of the hexameric building block of the HIV capsid. *Cell* 137, 1282–1292.

Pornillos, O., Ganser-Pornillos, B. K., and Yeager, M. (2011). Atomic-level modelling of the HIV capsid. *Nature* 469, 424–427.

Sayah, D. M., Sokolskaja, E., Berthoux, L., and Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* 430, 569–573.

Scarlata, S., and Carter, C. (2003). Role of HIV-1 Gag domains in viral assembly. *Biochim. Biophys. Acta* 1614, 62–72.

Schaller, T., Ocwieja, K. E., Rasaiyaah, J., Price, A. J., Brady, T. L., Roth, S. L., et al. (2011). HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog.* 7:e1002439. doi: 10.1371/journal.ppat. 1002439

Schneidewind, A., Brockman, M. A., Yang, R., Adam, R. I., Li, B., Le Gall, S., et al. (2007). Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. *J. Virol.* 81, 12382–12393.

Sokolskaja, E., Sayah, D. M., and Luban, J. (2004).Target cell cyclophilin A modulates human immunodeficiency virus type 1 infectivity. *J. Virol.* 78, 12800–12808.

Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P., and Sodroski, J. (2004). The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old world monkey. *Nature* 427, 848–853.

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.

Takeuchi, H., Ishii, H., Kuwano, T., Inagaki, N., Akari, H., and Matano, T. (2012). Host cell species-specific effect of cyclosporine A on simian immunodeficiency virus replication. *Retrovirology* 9, 3.

Towers, G. J. (2007). The control of viral infection by tripartite motif proteins and cyclophilin A. *Retrovirology* 4, 40.

Vogt, V. M. (1997). "Retroviral virions and genomes," in *Retroviruses*, eds J. M. Coffin, S. H. Hughes, and H. E. Varmus (Woodbury, NY: Cold

Spring Harbor Laboratory Press), 27–70.

Yang, R., and Aiken, C. (2007). A mutation in alpha helix 3 of CA renders human immunodeficiency virus type 1 cyclosporin A resistant and dependent: rescue by a second-site substitution in a distal region of CA. *J. Virol.* 81, 3749–3756.

Yap, M. W., Dodding, M. P., and Stoye, J. P. (2006). Trim-cyclophilin A fusion proteins can restrict human immunodeficiency virus type 1 infection at two distinct phases in the viral life cycle. *J. Virol.* 80, 4061–4067.

Yeager, M. (2011). Design of *in vitro* symmetric complexes and analysis by hybrid methods reveal mechanisms of HIV capsid assembly. *J. Mol. Biol.* 410, 534–552.

Zhang, R., Mehla, R., and Chauhan, A. (2010). Perturbation of host nuclear membrane component RanBP2 impairs the nuclear import of human immunodeficiency virus-1 preintegration complex (DNA). *PLoS ONE* 5:e15620. doi: 10.1371/journal.pone.0015620

Conflict of Interest Statement: 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.

Received: 30 July 2012; paper pending published: 07 August 2012; accepted: 28 September 2012; published online: 17 October 2012.

*Citation: Takemura T and Murakami T (2012) Functional constraints on HIV-1 capsid: their impacts on the viral immune escape potency. Front. Microbio. **3**:369. doi: 10.3389/fmicb.2012.00369*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

MINI REVIEW ARTICLE
published: 29 June 2012
doi: 10.3389/fmicb.2012.00234

Association of MHC-I genotypes with disease progression in HIV/SIV infections

Takushi Nomura*[1,2] *and Tetsuro Matano*[1,2] *

[1] *AIDS Research Center, National Institute of Infectious Diseases, Tokyo, Japan*
[2] *The Institute of Medical Science, The University of Tokyo, Tokyo, Japan*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Takamasa Ueno, Kumamoto University, Japan
Hirofumi Akari, Kyoto University, Japan

***Correspondence:**
Tetsuro Matano, AIDS Research Center, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan.
e-mail: tmatano@nih.go.jp

Virus-specific cytotoxic T lymphocytes (CTLs) are major effectors in acquired immune responses against viral infection. Virus-specific CTLs recognize specific viral peptides presented by major histocompatibility complex class-I (MHC-I) on the surface of virus-infected target cells via their T cell receptor (TCR) and eliminate target cells by both direct and indirect mechanisms. In human immunodeficiency virus (HIV) and simian immunodeficiency virus (SIV) infections, host immune responses fail to contain the virus and allow persistent viral replication, leading to AIDS progression. CTL responses exert strong suppressive pressure on HIV/SIV replication and cumulative studies have indicated association of HLA/MHC-I genotypes with rapid or slow AIDS progression.

Keywords: CTL, HIV, HLA, Mamu, MHC-I, MHC-I haplotype, SIV

INTRODUCTION

Innate and acquired immune responses play an important role in the control of infectious pathogens. Pathogenic microbes are able to escape from the host innate immune responses and replicate in the hosts. After the acute growth phase, pathogen-specific neutralizing antibody and cytotoxic T lymphocyte (CTL) responses are induced and prevent the onset of pathogenic manifestations in most of acute infectious diseases. In HIV and simian immunodeficiency virus (SIV) infections, these acquired immune responses are induced but fail to contain the virus and allow persistent viral replication, leading to AIDS progression, while persistent SIVsm infection of natural hosts, sooty mangabeys, does not result in disease onset (Silvestri et al., 2003). Effective neutralizing antibody responses are not efficiently induced in the acute phase (Burton et al., 2004). In contrast, virus-specific CTL responses play a main role in the reduction of viral loads from the peak to the set-point levels (Borrow et al., 1994; Koup et al., 1994; Matano et al., 1998; Jin et al., 1999; Schmitz et al., 1999). Previous studies suggest that, among various viral antigen-specific CTL responses, those directed against the viral structural protein Gag contribute to the control of viral replication (Edwards et al., 2002; Zuniga et al., 2006; Borghans et al., 2007; Kiepiela et al., 2007).

In virus-infected cells, antigenic peptides that are processed from viral proteins via the proteasome pathway and bound to MHC-I (HLA class I) molecules are presented on the cell surface. CTLs recognize antigenic peptide (epitope)-MHC-I complexes on the cell surface by their TCRs and eliminate the virus-infected cells by inducing apoptosis or lysis. Because presentation of antigenic peptides is restricted by MHC-I molecules, CTL efficacy is affected by MHC-I (HLA class I) genotypes.

ASSOCIATION OF HLA ALLELES WITH HIV PROGRESSION

HIV-infected individuals without anti-retroviral therapy (ART) mostly develop AIDS in 5–10 years after HIV exposure (Lui et al., 1988; Farewell et al., 1992). Humans have a single polymorphic HLA-A, HLA-B, and HLA-C locus per chromosome. A number of studies on HIV-infected individuals reported the association of HLA genotypes with disease progression (Tang et al., 2002; Kiepiela et al., 2004; Wang et al., 2009; Leslie et al., 2010). Indeed, association of *HLA-B*57* (Migueles et al., 2000; Altfeld et al., 2003; Miura et al., 2009) and *HLA-B*27* (Goulder et al., 1997; Feeney et al., 2004; Altfeld et al., 2006; Schneidewind et al., 2007) with lower viral loads in the chronic phase and slow disease progression has been indicated. HLA-B*57-restricted $Gag_{240-249}$ TW10 (TSTLQEQIGW) and HLA-B*27-restricted $Gag_{263-272}$ KK10 (KRWIILGLNK) epitope-specific CTL responses exert strong suppressive pressure on HIV replication and often select for viral genome mutations resulting in viral escape from these CTL recognition with viral fitness costs (Goulder et al., 1997; Feeney et al., 2004). Some HIV-infected individuals possessing those HLA alleles associating with slower disease progression control viral replication for long periods, while the frequency of such elite controllers is under 1% (Lambotte et al., 2005; Grabar et al., 2009). In contrast, HLA genotypes such as *HLA-B*35* associating with rapid disease progression have also been reported (Carrington et al., 1999; Gao et al., 2001). *HLA-B*35* subtypes are divided into *HLA-B*35-Px* and *HLA-B*35-Py* based on the specificity of binding ability to epitope peptides in the P9 pocket. The former group, *HLA-B*35-Px* alleles including *HLA-B*3502*, *B*3503*, and *B*3504* associate with rapid disease progression, whereas the latter *HLA-B*35-Py* alleles including *HLA-B*3501* and *HLA-B*3508* associate with relatively slower progression (Gao et al., 2001). Such differences in disease progression among *HLA-B* subtypes are also known in *HLA-B*58* (Leslie et al., 2010).

ANIMAL AIDS MODELS

Robust non-human primate AIDS models showing high pathogenic homology to human HIV infections are essential for

AIDS research. While it is difficult to analyze the early phase in human HIV infection, animal models have considerable advantages in immunological analysis in the acute phase. Furthermore, comparisons among the hosts infected with the same virus strain are possible in animal AIDS models, although highly diversified HIVs are prevalent in humans. An important characteristic of HIV infection is selective loss of memory $CCR5^+$ $CD4^+$ T lymphocytes in the acute phase leading to persistent virus replication (Connor et al., 1997; Zhang et al., 1999; Brenchley et al., 2004). HIV tropism for $CCR5^+$ $CD4^+$ memory cells is considered as one central mechanism for persistent infection. R5-tropic SIVmac251/SIVmac239 or SIVsmE660/SIVsmE543-3 infection of rhesus macaques inducing the acute, selective loss of memory $CD4^+$ T lymphocytes is currently considered the best AIDS model for analysis of AIDS pathogenesis and evaluation of vaccine efficacy (Veazey et al., 1998; Nishimura et al., 2004; Bontrop and Watkins, 2005; Mattapallil et al., 2005; Morgan et al., 2008). Recent studies indicated an association of restriction factor TRIM5α genotypes with disease progression in macaques infected with pathogenic SIVs such as SIVsmE660/SIVsmE543-3 but not in SIVmac239 infection (Kirmaier et al., 2010; Lim et al., 2010; de Groot et al., 2011; Fenizia et al., 2011; Letvin et al., 2011; Reynolds et al., 2011; Yeh et al., 2011). Macaque AIDS models of chimeric simian-human immunodeficiency virus (SHIV) infection are also known. Infection with X4-tropic SHIVs such as SHIV89.6P results in acute $CD4^+$ T cell depletion, while R5-tropic SHIVs such as SHIV162P3 induce persistent infection leading to chronic disease progression (Tsai et al., 2007; Nishimura et al., 2010; Zhuang et al., 2011). These SHIVs are useful especially for the analysis of Env-specific antibody responses (Ng et al., 2010; Watkins et al., 2011).

GENETIC FEATURES OF MHC-I IN MACAQUES

Human classical MHC-I alleles are composed of a single polymorphic HLA-A, HLA-B, and HLA-C locus per chromosome. MHC-I haplotypes in rhesus macaques, however, have variable numbers of Mamu-A and Mamu-B loci (Boyson et al., 1996; Adams and Parham, 2001; Daza-Vamenta et al., 2004; Kulski et al., 2004; Otting et al., 2005; **Figure 1**). A number of studies described SIV infections in macaques sharing one or two MHC-I alleles, while few studies have examined SIV infection in macaques sharing an MHC-I haplotype.

PROTECTIVE MHC-I ALLELES IN INDIAN RHESUS MACAQUES AGAINST SIV INFECTION

Simian immunodeficiency virus infections of Indian rhesus macaques are widely used as an AIDS model. *Mamu-A*01*, *Mamu-B*08*, and *Mamu-B*17* are known as protective alleles and macaques possessing these alleles tend to show slow disease progression after SIVmac251/SIVmac239 challenge (Muhl et al., 2002; Mothe et al., 2003; Yant et al., 2006; Loffredo et al., 2007b). Fourteen Mamu-A*01-restricted SIVmac239 CTL epitopes have been reported (Allen et al., 2001; Mothe et al., 2002b). Mamu-A*01-restricted Tat_{28-35} SL8 (STPESANL)-specific and $Gag_{181-189}$ CM9 (CTPYDINQM)-specific CTL responses are induced dominantly in SIVmac239 infection. Both epitope-specific CTLs show strong suppressive capacity against SIVmac239 replication *in vitro* (Loffredo et al., 2005), while the latter but not the former play a major role in suppression of viral replication *in vivo* (O'Connor et al., 2002; Loffredo et al., 2007c). In SHIV89.6P infection, Mamu-A*01-positive macaques elicit CM9-specific CTL responses and show slower disease progression than Mamu-A*01-negative animals (Zhang et al., 2002). Eight Mamu-B*08-restricted SIVmac239 CTL epitopes have been reported; previous studies indicated that $Vif_{123-131}$ RL9 (RRAIRGEQL), $Vif_{172-179}$ RL8 (RRDNRRGL), and $Nef_{137-146}$ RL10 (RRHRILDIYL) epitope-specific CTL responses contribute to viral control (Loffredo et al., 2007a; Loffredo et al., 2008; Valentine et al., 2009; Mudd et al., 2012). SIVmac239 Vif_{66-73} HW8 (HLEVQGYW), $Nef_{165-173}$ IW9 (IRYPKTFGW), and $Nef_{195-203}$ MW9 (MHPAQTSQW) have been reported as Mamu-B*17-restricted CTL epitopes (Mothe et al., 2002a). In addition, cRW9 (RHLAFKCLW) in an alternate reading frame is known as a cryptic epitope (Maness et al., 2007). The cRW9-coding region [nucleotides 6889–6915 in SIVmac239 (accession number M33262)] is located in the same open reading frame that encodes exon 1 of the Rev protein but is downstream of the splice donor site. So, it is not predicted to be translated under normal biological circumstances. However, SIVmac239-infected Mamu-B*17-positive macaques efficiently induce cRW9-specific CTL responses.

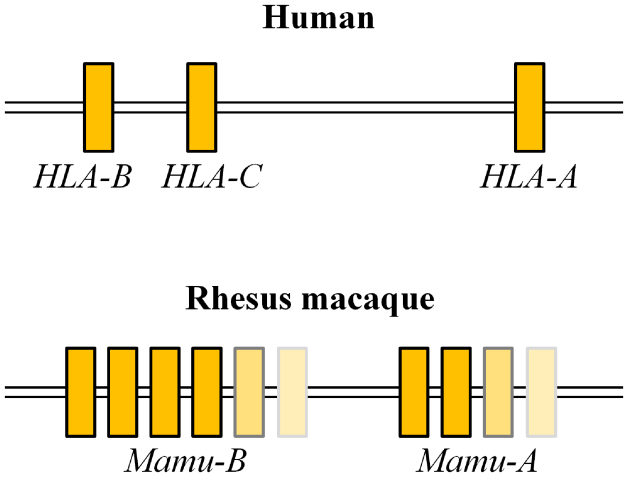


FIGURE 1 | Comparison of genome structures of MHC-I sub-regions in humans and rhesus macaques. Humans have a single polymorphic HLA-A, HLA-B, and HLA-C locus per chromosome, while rhesus macaques have variable numbers of Mamu-A and Mamu-B loci per chromosome.

ASSOCIATION OF MHC-I HAPLOTYPES WITH DISEASE PROGRESSION AFTER SIVmac239 CHALLENGE IN BURMESE RHESUS MACAQUES

We accumulated groups of Burmese rhesus macaques sharing individual MHC-I haplotypes (Tanaka-Takahashi et al., 2007; Naruse et al., 2010). SIVmac239 challenge of Burmese rhesus macaques mostly results in persistent viremia (geometric means of setpoint plasma viral loads: about 10^5 copies/ml) leading to AIDS (mean survival periods: about 2 years; Nomura et al., 2012). Further analysis revealed the association of MHC-I haplotypes with disease progression after SIVmac239 challenge.

Table 1 | Association of MHC-I haplotypes with disease progression in SIV infection (Nomura et al., 2012).

MHC-I haplotypes	Mean survival periods	Geometric means of setpoint plasma viral loads (copies/ml)	Peripheral $CD4^+$ T cell decline	Predominant CTL responses
90-120-Ia	>40 months	10^4	Slow	Gag/Nef
90-010-Ie	23 months	10^5	Intermediate	Nef
90-120-Ib	24 months	10^5	Intermediate	Nef
90-088-Ij	15 months	10^6	Rapid	-

In our study (Nomura et al., 2012), the group of Burmese rhesus macaques possessing MHC-I haplotype *90-010-Ie* (dominant MHC-I alleles: *A1*066:01* and *B*005:02*) exhibited a typical pattern of disease progression after SIVmac239 challenge (**Table 1**). These animals showed predominant Nef-specific CTL responses, approximately 10^5 copies/ml of setpoint plasma viral loads (geometric means), and 2 years of mean survival periods. Another group of macaques possessing *90-120-Ib* (dominant MHC-I alleles: *A1*018:08* and *B*036:03*) showed similar setpoint viral loads and survival periods. However, the group of Burmese rhesus macaques possessing MHC-I haplotype *90-088-Ij* (dominant MHC-I alleles: *A1*008:01* and *B*007:02*) showed higher setpoint plasma viral loads (geometric means: about 10^6 copies/ml) and shorter survival periods (means: about 15 months; **Table 1**). These animals mostly showed poor CTL responses.

In contrast, the group of Burmese rhesus macaques possessing MHC-I haplotype *90-120-Ia* (dominant MHC-I alleles: *A1*043:01* and *B*061:03*), referred to as A^+ animals, showed lower setpoint plasma viral loads (geometric means: about 10^4 copies/ml) and slower disease progression (means of survival periods: more than 40 months; **Table 1**). These animals predominantly elicited Gag-specific and Nef-specific CTL responses after SIVmac239 challenge. Mamu-A1*043:01-restricted $Gag_{206-216}$ (IINEEAADWDL) and Mamu-A1*065:01-restricted $Gag_{241-249}$ (SSVDEQIQW) were determined as dominant CTL epitopes. SIVmac239-infected A^+ animals selected viral escape mutations from these epitope-specific CTL responses with viral fitness costs in the chronic phase (Kobayashi et al., 2005; Kawada et al., 2006). These mutations are GagL216S, a mutation leading to a leucine (L)-to-serine (S) substitution at the 216th amino acid in SIVmac239 Gag, and GagD244E, aspartic acid (D)-to-glutamic acid (E) at the 244th, or GagI247L, isoleucine [I]-to-L at the 247th. A^+ animals immunized with a prophylactic prime-boost vaccine consisting of a DNA prime followed by a boost with a recombinant Sendai virus vector expressing SIVmac239 Gag controlled an SIVmac239 challenge (Matano et al., 2004). However, vaccinated A^+ animals failed to control a challenge with a mutant SIVmac239 carrying GagL216S and GagD244E, indicating that $Gag_{206-216}$-specific and $Gag_{241-249}$-specific CTL responses are responsible for the control of the wild-type SIVmac239 replication (Kawada et al., 2006, 2008). Interestingly, the Mamu-A1*065:01-restricted SIVmac239 $Gag_{241-249}$ epitope is located in a region corresponding to the HLA-B*57-restricted HIV $Gag_{240-249}$ epitope TW10 and TW10-specific CTL responses have also been indicated to exert strong suppressive pressure on HIV replication. An SIVmac239 $Gag_{241-249}$-specific CTL escape mutation, GagD244E, results in loss of viral fitness similarly with an HIV TW10-specific CTL escape mutation. Both of the Mamu-A1*065:01-restricted SIVmac239 $Gag_{241-249}$ epitope and the HLA-B*57-restricted HIV TW10 epitope are considered to have the same anchor residues, S at position 2 and tryptophan (W) at the carboxyl terminus. Additionally, anchor residues of CTL epitopes presented by Mamu-B*17/Mamu-B*08 were indicated to be similar to those restricted by HLA-B*57/HLA-B*27 (Loffredo et al., 2009; Wu et al., 2011).

CONCLUDING REMARKS

Human HLA genotypes largely affect disease progression in HIV infection, reflecting that CTL responses play a central role in suppression of HIV replication. Animal AIDS models are required for understanding of the interaction between highly diversified viruses and the hosts with polymorphic MHC-I genotypes. SIV infection of Indian rhesus macaques are widely used as an AIDS model, and association of certain MHC-I alleles with slower disease progression has been indicated. We have recently reported SIV infection of Burmese rhesus macaques as a robust AIDS model and indicated association of MHC-I haplotypes with disease progression. Accumulation of those macaque groups sharing MHC-I haplotypes could lead to constitution of a more sophisticated AIDS model facilitating analysis of virus-host immune interaction.

REFERENCES

Adams, E. J., and Parham, P. (2001). Species-specific evolution of MHC class I genes in the higher primates. *Immunol. Rev.* 183, 41–64.

Allen, T. M., Mothe, B. R., Sidney, J., Jing, P., Dzuris, J. L., Liebl, M. E., Vogel, T. U., O'Connor, D. H., Wang, X., Wussow, M. C., Thomson, J. A., Altman, J. D., Watkins, D. I., and Sette, A. (2001). CD8(+) lymphocytes from simian immunodeficiency virus-infected rhesus macaques recognize 14 different epitopes bound by the major histocompatibility complex class I molecule mamu-A*01: implications for vaccine design and testing. *J. Virol.* 75, 738–749.

Altfeld, M., Addo, M. M., Rosenberg, E. S., Hecht, F. M., Lee, P. K., Vogel, M., Yu, X. G., Draenert, R., Johnston, M. N., Strick, D., Allen, T. M., Feeney, M. E., Kahn, J. O., Sekaly, R. P., Levy, J. A., Rockstroh, J. K., Goulder, P. J., and Walker, B. D. (2003). Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection. *AIDS* 17, 2581–2591.

Altfeld, M., Kalife, E. T., Qi, Y., Streeck, H., Lichterfeld, M., Johnston, M. N., Burgett, N., Swartz, M. E., Yang, A., Alter, G., Yu, X. G., Meier, A., Rockstroh, J. K., Allen, T. M., Jessen, H., Rosenberg, E. S., Carrington, M., and Walker, B. D. (2006). HLA alleles associated with delayed progression to AIDS contribute strongly to the initial CD8(+) T cell response against HIV-1. *PLoS Med.* 3, e403. doi: 10.1371/journal.pmed.0030403

Bontrop, R. E., and Watkins, D. I. (2005). MHC polymorphism: AIDS susceptibility in non-human primates. *Trends Immunol.* 26, 227–233.

Borghans, J. A., Molgaard, A., de Boer, R. J., and Kesmir, C. (2007). HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24. *PLoS ONE* 2, e920. doi: 10.1371/journal.pone.0000920s

Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M., and Oldstone, M. B. (1994). Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* 68, 6103–6110.

Boyson, J. E., Shufflebotham, C., Cadavid, L. F., Urvater, J. A., Knapp, L. A., Hughes, A. L., and Watkins, D. I. (1996). The MHC class I genes of the rhesus monkey. Different evolutionary histories of MHC class I and II genes in primates. *J. Immunol.* 156, 4656–4665.

Brenchley, J. M., Schacker, T. W., Ruff, L. E., Price, D. A., Taylor, J. H., Beilman, G. J., Nguyen, P. L., Khoruts, A., Larson, M., Haase, A. T., and Douek, D. C. (2004). CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J. Exp. Med.* 200, 749–759.

Burton, D. R., Desrosiers, R. C., Doms, R. W., Koff, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. (2004). HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.* 5, 233–236.

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., and O'Brien, S. J. (1999). HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283, 1748–1752.

Connor, R. I., Sheridan, K. E., Ceradini, D., Choe, S., and Landau, N. R. (1997). Change in coreceptor use correlates with disease progression in HIV-1-infected individuals. *J. Exp. Med.* 185, 621–628.

Daza-Vamenta, R., Glusman, G., Rowen, L., Guthrie, B., and Geraghty, D. E. (2004). Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res.* 14, 1501–1515.

de Groot, N. G., Heijmans, C. M., Koopman, G., Verschoor, E. J., Bogers, W. M., and Bontrop, R. E. (2011). TRIM5 allelic polymorphism in macaque species/populations of different geographic origins: its impact on SIV vaccine studies. *Tissue Antigens* 78, 256–262.

Edwards, B. H., Bansal, A., Sabbaj, S., Bakari, J., Mulligan, M. J., and Goepfert, P. A. (2002). Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. *J. Virol.* 76, 2298–2305.

Farewell, V. T., Coates, R. A., Fanning, M. M., MacFadden, D. K., Read, S. E., Shepherd, F. A., and Struthers, C. A. (1992). The probability of progression to AIDS in a cohort of male sexual contacts of men with HIV disease. *Int. J. Epidemiol.* 21, 131–135.

Feeney, M. E., Tang, Y., Roosevelt, K. A., Leslie, A. J., McIntosh, K., Karthas, N., Walker, B. D., and Goulder, P. J. (2004). Immune escape precedes breakthrough human immunodeficiency virus type 1 viremia and broadening of the cytotoxic T-lymphocyte response in an HLA-B27-positive long-term-nonprogressing child. *J. Virol.* 78, 8927–8930.

Fenizia, C., Keele, B. F., Nichols, D., Cornara, S., Binello, N., Vaccari, M., Pegu, P., Robert-Guroff, M., Ma, Z. M., Miller, C. J., Venzon, D., Hirsch, V., and Franchini, G. (2011). TRIM5alpha does not affect simian immunodeficiency virus SIV(mac251) replication in vaccinated or unvaccinated Indian rhesus macaques following intrarectal challenge exposure. *J. Virol.* 85, 12399–12409.

Gao, X., Nelson, G. W., Karacki, P., Martin, M. P., Phair, J., Kaslow, R., Goedert, J. J., Buchbinder, S., Hoots, K., Vlahov, D., O'Brien, S. J., and Carrington, M. (2001). Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N. Engl. J. Med.* 344, 1668–1675.

Goulder, P. J., Phillips, R. E., Colbert, R. A., McAdam, S., Ogg, G., Nowak, M. A., Giangrande, P., Luzzi, G., Morgan, B., Edwards, A., McMichael, A. J., and Rowland-Jones, S. (1997). Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat. Med.* 3, 212–217.

Grabar, S., Selinger-Leneman, H., Abgrall, S., Pialoux, G., Weiss, L., and Costagliola, D. (2009). Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. *AIDS* 23, 1163–1169.

Jin, X., Bauer, D. E., Tuttleton, S. E., Lewin, S., Gettie, A., Blanchard, J., Irwin, C. E., Safrit, J. T., Mittler, J., Weinberger, L., Kostrikis, L. G., Zhang, L., Perelson, A. S., and Ho, D. D. (1999). Dramatic rise in plasma viremia after CD8(+) T cell depletion in simian immunodeficiency virus-infected macaques. *J. Exp. Med.* 189, 991–998.

Kawada, M., Igarashi, H., Takeda, A., Tsukamoto, T., Yamamoto, H., Dohki, S., Takiguchi, M., and Matano, T. (2006). Involvement of multiple epitope-specific cytotoxic T-lymphocyte responses in vaccine-based control of simian immunodeficiency virus replication in rhesus macaques. *J. Virol.* 80, 1949–1958.

Kawada, M., Tsukamoto, T., Yamamoto, H., Iwamoto, N., Kurihara, K., Takeda, A., Moriya, C., Takeuchi, H., Akari, H., and Matano, T. (2008). Gag-specific cytotoxic T-lymphocyte-based control of primary simian immunodeficiency virus replication in a vaccine trial. *J. Virol.* 82, 10199–10206.

Kiepiela, P., Leslie, A. J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferott, K. J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M. M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L. D., Szinger, J., Day, C., Klenerman, P., Mullins, J., Korber, B., Coovadia, H. M., Walker, B. D., and Goulder, P. J. (2004). Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 432, 769–775.

Kiepiela, P., Ngumbela, K., Thobakgale, C., Ramduth, D., Honeyborne, I., Moodley, E., Reddy, S., de Pierres, C., Mncube, Z., Mkhwanazi, N., Bishop, K., van der Stok, M., Nair, K., Khan, N., Crawford, H., Payne, R., Leslie, A., Prado, J., Prendergast, A., Frater, J., McCarthy, N., Brander, C., Learn, G. H., Nickle, D., Rousseau, C., Coovadia, H., Mullins, J. I., Heckerman, D., Walker, B. D., and Goulder, P. (2007). CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat. Med.* 13, 46–53.

Kirmaier, A., Wu, F., Newman, R. M., Hall, L. R., Morgan, J. S., O'Connor, S., Marx, P. A., Meythaler, M., Goldstein, S., Buckler-White, A., Kaur, A., Hirsch, V. M., and Johnson, W. E. (2010). TRIM5 suppresses cross-species transmission of a primate immunodeficiency virus and selects for emergence of resistant variants in the new species. *PLoS Biol.* 8. doi: 10.1371/journal.pbio.1000462

Kobayashi, M., Igarashi, H., Takeda, A., Kato, M., and Matano, T. (2005). Reversion *in vivo* after inoculation of a molecular proviral DNA clone of simian immunodeficiency virus with a cytotoxic-T-lymphocyte escape mutation. *J. Virol.* 79, 11529–11532.

Koup, R. A., Safrit, J. T., Cao, Y., Andrews, C. A., McLeod, G., Borkowsky, W., Farthing, C., and Ho, D. D. (1994). Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* 68, 4650–4655.

Kulski, J. K., Anzai, T., Shiina, T., and Inoko, H. (2004). Rhesus macaque class I duplicon structures, organization, and evolution within the alpha block of the major histocompatibility complex. *Mol. Biol. Evol.* 21, 2079–2091.

Lambotte, O., Boufassa, F., Madec, Y., Nguyen, A., Goujard, C., Meyer, L., Rouzioux, C., Venet, A., and Delfraissy, J. F. (2005). HIV controllers: a homogeneous group of HIV-1-infected patients with spontaneous control of viral replication. *Clin. Infect. Dis.* 41, 1053–1056.

Leslie, A., Matthews, P. C., Listgarten, J., Carlson, J. M., Kadie, C., Ndung'u, T., Brander, C., Coovadia, H., Walker, B. D., Heckerman, D., and Goulder, P. J. (2010). Additive contribution of HLA class I alleles in the immune control of HIV-1 infection. *J. Virol.* 84, 9879–9888.

Letvin, N. L., Rao, S. S., Montefiori, D. C., Seaman, M. S., Sun, Y., Lim, S. Y., Yeh, W. W., Asmal, M., Gelman, R. S., Shen, L., Whitney, J. B., Seoighe, C., Lacerda, M., Keating, S., Norris, P. J., Hudgens, M. G., Gilbert, P. B., Buzby, A. P., Mach, L. V., Zhang, J., Balachandran, H., Shaw, G. M., Schmidt, S. D., Todd, J. P., Dodson, A., Mascola, J. R., and Nabel, G. J. (2011). Immune and genetic correlates of vaccine protection against mucosal infection by SIV in monkeys. *Sci. Transl. Med.* 3, 81ra36.

Lim, S. Y., Rogers, T., Chan, T., Whitney, J. B., Kim, J., Sodroski, J., and Letvin, N. L. (2010). TRIM5alpha modulates immunodeficiency virus control in rhesus monkeys. *PLoS Pathog.* 6, e1000738. doi: 10.1371/journal.ppat.1000738

Loffredo, J. T., Rakasz, E. G., Giraldo, J. P., Spencer, S. P., Grafton, K. K., Martin, S. R., Napoe, G., Yant, L. J., Wilson, N. A., and Watkins,

D. I. (2005). Tat(28-35)SL8-specific CD8+ T lymphocytes are more effective than Gag(181–189)CM9-specific CD8+ T lymphocytes at suppressing simian immunodeficiency virus replication in a functional *in vitro* assay. *J. Virol.* 79, 14986–14991.

Loffredo, J. T., Bean, A. T., Beal, D. R., Leon, E. J., May, G. E., Piaskowski, S. M., Furlott, J. R., Reed, J., Musani, S. K., Rakasz, E. G., Friedrich, T. C., Wilson, N. A., Allison, D. B., and Watkins, D. I. (2008). Patterns of CD8+ immunodominance may influence the ability of Mamu-B*08-positive macaques to naturally control simian immunodeficiency virus SIVmac239 replication. *J. Virol.* 82, 1723–1738.

Loffredo, J. T., Friedrich, T. C., Leon, E. J., Stephany, J. J., Rodrigues, D. S., Spencer, S. P., Bean, A. T., Beal, D. R., Burwitz, B. J., Rudersdorf, R. A., Wallace, L. T., Piaskowski, S. M., May, G. E., Sidney, J., Gostick, E., Wilson, N. A., Price, D. A., Kallas, E. G., Piontkivska, H., Hughes, A. L., Sette, A., and Watkins, D. I. (2007a). CD8+ T cells from SIV elite controller macaques recognize Mamu-B*08-bound epitopes and select for widespread viral variation. *PLoS ONE* 2, e1152. doi: 10.1371/journal.pone.0001152

Loffredo, J. T., Maxwell, J., Qi, Y., Glidden, C. E., Borchardt, G. J., Soma, T., Bean, A. T., Beal, D. R., Wilson, N. A., Rehrauer, W. M., Lifson, J. D., Carrington, M., and Watkins, D. I. (2007b). Mamu-B*08-positive macaques control simian immunodeficiency virus replication. *J. Virol.* 81, 8827–8832.

Loffredo, J. T., Burwitz, B. J., Rakasz, E. G., Spencer, S. P., Stephany, J. J., Vela, J. P., Martin, S. R., Reed, J., Piaskowski, S. M., Furlott, J., Weisgrau, K. L., Rodrigues, D. S., Soma, T., Napoe, G., Friedrich, T. C., Wilson, N. A., Kallas, E. G., and Watkins, D. I. (2007c). The antiviral efficacy of simian immunodeficiency virus-specific CD8+ T cells is unrelated to epitope specificity and is abrogated by viral escape. *J. Virol.* 81, 2624–2634.

Loffredo, J. T., Sidney, J., Bean, A. T., Beal, D. R., Bardet, W., Wahl, A., Hawkins, O. E., Piaskowski, S., Wilson, N. A., Hildebrand, W. H., Watkins, D. I., and Sette, A. (2009). Two MHC class I molecules associated with elite control of immunodeficiency virus replication, Mamu-B*08 and HLA-B*2705, bind peptides with sequence similarity. *J. Immunol.* 182, 7763–7775.

Lui, K. J., Darrow, W. W., and Rutherford, G. W., 3rd (1988). A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science* 240, 1333–1335.

Maness, N. J., Valentine, L. E., May, G. E., Reed, J., Piaskowski, S. M., Soma, T., Furlott, J., Rakasz, E. G., Friedrich, T. C., Price, D. A., Gostick, E., Hughes, A. L., Sidney, J., Sette, A., Wilson, N. A., and Watkins, D. I. (2007). AIDS virus specific CD8+ T lymphocytes against an immunodominant cryptic epitope select for viral escape. *J. Exp. Med.* 204, 2505–2512.

Matano, T., Kobayashi, M., Igarashi, H., Takeda, A., Nakamura, H., Kano, M., Sugimoto, C., Mori, K., Iida, A., Hirata, T., Hasegawa, M., Yuasa, T., Miyazawa, M., Takahashi, Y., Yasunami, M., Kimura, A., O'Connor, D. H., Watkins, D. I., and Nagai, Y. (2004). Cytotoxic T lymphocyte-based control of simian immunodeficiency virus replication in a preclinical AIDS vaccine trial. *J. Exp. Med.* 199, 1709–1718.

Matano, T., Shibata, R., Siemon, C., Connors, M., Lane, H. C., and Martin, M. A. (1998). Administration of an anti-CD8 monoclonal antibody interferes with the clearance of chimeric simian/human immunodeficiency virus during primary infections of rhesus macaques. *J. Virol.* 72, 164–169.

Mattapallil, J. J., Douek, D. C., Hill, B., Nishimura, Y., Martin, M., and Roederer, M. (2005). Massive infection and loss of memory CD4+ T cells in multiple tissues during acute SIV infection. *Nature* 434, 1093–1097.

Migueles, S. A., Sabbaghian, M. S., Shupert, W. L., Bettinotti, M. P., Marincola, F. M., Martino, L., Hallahan, C. W., Selig, S. M., Schwartz, D., Sullivan, J., and Connors, M. (2000). HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. U.S.A.* 97, 2709–2714.

Miura, T., Brockman, M. A., Schneidewind, A., Lobritz, M., Pereyra, F., Rathod, A., Block, B. L., Brumme, Z. L., Brumme, C. J., Baker, B., Rothchild, A. C., Li, B., Trocha, A., Cutrell, E., Frahm, N., Brander, C., Toth, I., Arts, E. J., Allen, T. M., and Walker, B. D. (2009). HLA-B57/B*5801 human immunodeficiency virus type 1 elite controllers select for rare gag variants associated with reduced viral replication capacity and strong cytotoxic T-lymphocyte [corrected] recognition. *J. Virol.* 83, 2743–2755.

Morgan, C., Marthas, M., Miller, C., Duerr, A., Cheng-Mayer, C., Desrosiers, R., Flores, J., Haigwood, N., Hu, S. L., Johnson, R. P., Lifson, J., Montefiori, D., Moore, J., Robert-Guroff, M., Robinson, H., Self, S., and Corey, L. (2008). The use of nonhuman primate models in HIV vaccine development. *PLoS Med.* 5, e173. doi: 10.1371/journal.pmed.0050173

Mothe, B. R., Sidney, J., Dzuris, J. L., Liebl, M. E., Fuenger, S., Watkins, D. I., and Sette, A. (2002a). Characterization of the peptide-binding specificity of Mamu-B*17 and identification of Mamu-B*17-restricted epitopes derived from simian immunodeficiency virus proteins. *J. Immunol.* 169, 210–219.

Mothe, B. R., Horton, H., Carter, D. K., Allen, T. M., Liebl, M. E., Skinner, P., Vogel, T. U., Fuenger, S., Vielhuber, K., Rehrauer, W., Wilson, N., Franchini, G., Altman, J. D., Haase, A., Picker, L. J., Allison, D. B., and Watkins, D. I. (2002b). Dominance of CD8 responses specific for epitopes bound by a single major histocompatibility complex class I molecule during the acute phase of viral infection. *J. Virol.* 76, 875–884.

Mothe, B. R., Weinfurter, J., Wang, C., Rehrauer, W., Wilson, N., Allen, T. M., Allison, D. B., and Watkins, D. I. (2003). Expression of the major histocompatibility complex class I molecule Mamu-A*01 is associated with control of simian immunodeficiency virus SIVmac239 replication. *J. Virol.* 77, 2736–2740.

Mudd, P. A., Ericsen, A. J., Burwitz, B. J., Wilson, N. A., O'Connor, D. H., Hughes, A. L., and Watkins, D. I. (2012). Escape from CD8+ T cell responses in Mamu-B*00801+ macaques differentiates progressors from elite controllers. *J. Immunol.* 188, 3364–3370.

Muhl, T., Krawczak, M., Ten Haaft, P., Hunsmann, G., and Sauermann, U. (2002). MHC class I alleles influence set-point viral load and survival time in simian immunodeficiency virus-infected rhesus monkeys. *J. Immunol.* 169, 3438–3446.

Naruse, T. K., Chen, Z., Yanagida, R., Yamashita, T., Saito, Y., Mori, K., Akari, H., Yasutomi, Y., Miyazawa, M., Matano, T., and Kimura, A. (2010). Diversity of MHC class I genes in Burmese-origin rhesus macaques. *Immunogenetics* 62, 601–611.

Ng, C. T., Jaworski, J. P., Jayaraman, P., Sutton, W. F., Delio, P., Kuller, L., Anderson, D., Landucci, G., Richardson, B. A., Burton, D. R., Forthal, D. N., and Haigwood, N. L. (2010). Passive neutralizing antibody controls SHIV viremia and enhances B cell responses in infant macaques. *Nat. Med.* 16, 1117–1119.

Nishimura, Y., Igarashi, T., Donau, O. K., Buckler-White, A., Buckler, C., Lafont, B. A., Goeken, R. M., Goldstein, S., Hirsch, V. M., and Martin, M. A. (2004). Highly pathogenic SHIVs and SIVs target different CD4+ T cell subsets in rhesus monkeys, explaining their divergent clinical courses. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12324–12329.

Nishimura, Y., Shingai, M., Willey, R., Sadjadpour, R., Lee, W. R., Brown, C. R., Brenchley, J. M., Buckler-White, A., Petros, R., Eckhaus, M., Hoffman, V., Igarashi, T., and Martin, M. A. (2010). Generation of the pathogenic R5-tropic simian/human immunodeficiency virus SHIVAD8 by serial passaging in rhesus macaques. *J. Virol.* 84, 4769–4781.

Nomura, T., Yamamoto, H., Shiino, T., Takahashi, N., Nakane, T., Iwamoto, N., Ishii, H., Tsukamoto, T., Kawada, M., Matsuoka, S., Takeda, A., Terahara, K., Tsunetsugu-Yokota, Y., Iwata-Yoshikawa, N., Hasegawa, H., Sata, T., Naruse, T. K., Kimura, A., and Matano, T. (2012). Association of major histocompatibility complex class I haplotypes with disease progression after simian immunodeficiency virus challenge in Burmese rhesus macaques. *J. Virol.* 86, 6481–6490.

O'Connor, D. H., Allen, T. M., Vogel, T. U., Jing, P., DeSouza, I. P., Dodds, E., Dunphy, E. J., Melsaether, C., Mothe, B., Yamamoto, H., Horton, H., Wilson, N., Hughes, A. L., and Watkins, D. I. (2002). Acute phase cytotoxic T lymphocyte escape is a hallmark of simian immunodeficiency virus infection. *Nat. Med.* 8, 493–499.

Otting, N., Heijmans, C. M., Noort, R. C., de Groot, N. G., Doxiadis, G. G., van Rood, J. J., Watkins, D. I., and Bontrop, R. E. (2005). Unparalleled complexity of the MHC class I region in rhesus macaques. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1626–1631.

Reynolds, M. R., Sacha, J. B., Weiler, A. M., Borchardt, G. J., Glidden, C. E., Sheppard, N. C., Norante, F. A., Castrovinci, P. A., Harris, J. J., Robertson, H. T., Friedrich, T. C., McDermott, A. B., Wilson, N. A., Allison, D. B., Koff, W. C., Johnson, W. E., and Watkins, D. I. (2011). The TRIM5{alpha} genotype of rhesus macaques affects acquisition of simian immunodeficiency virus SIVsmE660 infection after repeated limiting-dose intrarectal challenge. *J. Virol.* 85, 9637–9640.

Schmitz, J. E., Kuroda, M. J., Santra, S., Sasseville, V. G., Simon, M. A., Lifton, M. A., Racz, P., Tenner-Racz, K., Dalesandro, M., Scallon, B. J., Ghrayeb, J., Forman, M. A., Montefiori, D. C., Rieber, E. P., Letvin, N. L., and Reimann, K. A. (1999). Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. *Science* 283, 857–860.

Schneidewind, A., Brockman, M. A., Yang, R., Adam, R. I., Li, B., Le Gall, S., Rinaldo, C. R., Craggs, S. L., Allgaier, R. L., Power, K. A., Kuntzen, T., Tung, C. S., LaBute, M. X., Mueller, S. M., Harrer, T., McMichael, A. J., Goulder, P. J., Aiken, C., Brander, C., Kelleher, A. D., and Allen, T. M. (2007). Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. *J. Virol.* 81, 12382–12393.

Silvestri, G., Sodora, D. L., Koup, R. A., Paiardini, M., O'Neil, S. P., McClure, H. M., Staprans, S. I., and Feinberg, M. B. (2003). Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. *Immunity* 18, 441–452.

Tanaka-Takahashi, Y., Yasunami, M., Naruse, T., Hinohara, K., Matano, T., Mori, K., Miyazawa, M., Honda, M., Yasutomi, Y., Nagai, Y., and Kimura, A. (2007). Reference strand-mediated conformation analysis-based typing of multiple alleles in the rhesus macaque MHC class I Mamu-A and Mamu-B loci. *Electrophoresis* 28, 918–924.

Tang, J., Tang, S., Lobashevsky, E., Myracle, A. D., Fideli, U., Aldrovandi, G., Allen, S., Musonda, R., and Kaslow, R. A. (2002). Favorable and unfavorable HLA class I alleles and haplotypes in Zambians predominantly infected with clade C human immunodeficiency virus type 1. *J. Virol.* 76, 8276–8284.

Tsai, L., Trunova, N., Gettie, A., Mohri, H., Bohm, R., Saifuddin, M., and Cheng-Mayer, C. (2007). Efficient repeated low-dose intravaginal infection with X4 and R5 SHIVs in rhesus macaque: implications for HIV-1 transmission in humans. *Virology* 362, 207–216.

Valentine, L. E., Loffredo, J. T., Bean, A. T., Leon, E. J., MacNair, C. E., Beal, D. R., Piaskowski, S. M., Klimentidis, Y. C., Lank, S. M., Wiseman, R. W., Weinfurter, J. T., May, G. E., Rakasz, E. G., Wilson, N. A., Friedrich, T. C., O'Connor, D. H., Allison, D. B., and Watkins, D. I. (2009). Infection with "escaped" virus variants impairs control of simian immunodeficiency virus SIVmac239 replication in Mamu-B*08-positive macaques. *J. Virol.* 83, 11514–11527.

Veazey, R. S., DeMaria, M., Chalifoux, L. V., Shvetz, D. E., Pauley, D. R., Knight, H. L., Rosenzweig, M., Johnson, R. P., Desrosiers, R. C., and Lackner, A. A. (1998). Gastrointestinal tract as a major site of CD4+ T cell depletion and viral replication in SIV infection. *Science* 280, 427–431.

Wang, Y. E., Li, B., Carlson, J. M., Streeck, H., Gladden, A. D., Goodman, R., Schneidewind, A., Power, K. A., Toth, I., Frahm, N., Alter, G., Brander, C., Carrington, M., Walker, B. D., Altfeld, M., Heckerman, D., and Allen, T. M. (2009). Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. *J. Virol.* 83, 1845–1855.

Watkins, J. D., Diaz-Rodriguez, J., Siddappa, N. B., Corti, D., and Ruprecht, R. M. (2011). Efficiency of neutralizing antibodies targeting the CD4-binding site: influence of conformational masking by the V2 loop in R5-tropic clade C simian-human immunodeficiency virus. *J. Virol.* 85, 12811–12814.

Wu, Y., Gao, F., Liu, J., Qi, J., Gostick, E., Price, D. A., and Gao, G. F. (2011). Structural basis of diverse peptide accommodation by the rhesus macaque MHC class I molecule Mamu-B*17: insights into immune protection from simian immunodeficiency virus. *J. Immunol.* 187, 6382–6392.

Yant, L. J., Friedrich, T. C., Johnson, R. C., May, G. E., Maness, N. J., Enz, A. M., Lifson, J. D., O'Connor, D. H., Carrington, M., and Watkins, D. I. (2006). The high-frequency major histocompatibility complex class I allele Mamu-B*17 is associated with control of simian immunodeficiency virus SIVmac239 replication. *J. Virol.* 80, 5074–5077.

Yeh, W. W., Rao, S. S., Lim, S. Y., Zhang, J., Hraber, P. T., Brassard, L. M., Luedemann, C., Todd, J. P., Dodson, A., Shen, L., Buzby, A. P., Whitney, J. B., Korber, B. T., Nabel, G. J., Mascola, J. R., and Letvin, N. L. (2011). The TRIM5 gene modulates penile mucosal acquisition of simian immunodeficiency virus in rhesus monkeys. *J. Virol.* 85, 10389–10398.

Zhang, Z., Schuler, T., Zupancic, M., Wietgrefe, S., Staskus, K. A., Reimann, K. A., Reinhart, T. A., Rogan, M., Cavert, W., Miller, C. J., Veazey, R. S., Notermans, D., Little, S., Danner, S. A., Richman, D. D., Havlir, D., Wong, J., Jordan, H. L., Schacker, T. W., Racz, P., Tenner-Racz, K., Letvin, N. L., Wolinsky, S., and Haase, A. T. (1999). Sexual transmission and propagation of SIV and HIV in resting and activated CD4+ T cells. *Science* 286, 1353–1357.

Zhang, Z. Q., Fu, T. M., Casimiro, D. R., Davies, M. E., Liang, X., Schleif, W. A., Handt, L., Tussey, L., Chen, M., Tang, A., Wilson, K. A., Trigona, W. L., Freed, D. C., Tan, C. Y., Horton, M., Emini, E. A., and Shiver, J. W. (2002). Mamu-A*01 allele-mediated attenuation of disease progression in simian-human immunodeficiency virus infection. *J. Virol.* 76, 12845–12854.

Zhuang, K., Finzi, A., Tasca, S., Shakirzyanova, M., Knight, H., Westmoreland, S., Sodroski, J., and Cheng-Mayer, C. (2011). Adoption of an "open" envelope conformation facilitating CD4 binding and structural remodeling precedes coreceptor switch in R5 SHIV-infected macaques. *PLoS ONE* 6, e21350. doi: 10.1371/journal.pone.0021350

Zuniga, R., Lucchetti, A., Galvan, P., Sanchez, S., Sanchez, C., Hernandez, A., Sanchez, H., Frahm, N., Linde, C. H., Hewitt, H. S., Hildebrand, W., Altfeld, M., Allen, T. M., Walker, B. D., Korber, B. T., Leitner, T., Sanchez, J., and Brander, C. (2006). Relative dominance of Gag p24-specific cytotoxic T lymphocytes is associated with human immunodeficiency virus control. *J. Virol.* 80, 3122–3125.

Conflict of Interest Statement: 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.

Received: 01 May 2012; paper pending published: 16 May 2012; accepted: 11 June 2012; published online: 29 June 2012.

*Citation: Nomura T and Matano T (2012) Association of MHC-I genotypes with disease progression in HIV/SIV infections. Front. Microbio. **3**:234. doi: 10.3389/fmicb.2012.00234*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 31 December 2012
doi: 10.3389/fmicb.2012.00429

Molecular recognition of paired receptors in the immune system

Kimiko Kuroki*[1], *Atsushi Furukawa*[1,2] *and Katsumi Maenaka*[1,2]

[1] *Laboratory of Biomolecular Science, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*
[2] *Core Research for Evolutional Sciences and Technology, Japan Science and Technology Agency, Saitama, Japan*

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Hidekatsu Iha, Oita University, Japan
Takamasa Ueno, Kumamoto University, Japan
Hisashi Arase, Osaka University, Japan

****Correspondence:***
Katsumi Maenaka, Laboratory of Biomolecular Science, Faculty of Pharmaceutical Sciences, Hokkaido University, Kita-12, Nishi-6, Kita-ku, Sapporo 060-0812, Japan.
e-mail: maenaka@pharm.hokudai.ac.jp

Cell surface receptors are responsible for regulating cellular function on the front line, the cell membrane. Interestingly, accumulating evidence clearly reveals that the members of cell surface receptor families have very similar extracellular ligand-binding regions but opposite signaling systems, either inhibitory or stimulatory. These receptors are designated as paired receptors. Paired receptors often recognize not only physiological ligands but also non-self ligands, such as viral and bacterial products, to fight infections. In this review, we introduce several representative examples of paired receptors, focusing on two major structural superfamilies, the immunoglobulin-like and the C-type lectin-like receptors, and explain how these receptors distinguish self and non-self ligands to maintain homeostasis in the immune system. We further discuss the evolutionary aspects of these receptors as well as the potential drug targets for regulating diseases.

Keywords: paired receptor, immunoglobulin-like receptor, c-type lectin-like receptor, infectious diseases, tumorigenesis, ITIM, ITAM, structural biology

PAIRED RECEPTORS

Paired receptors are related membrane proteins that are mainly expressed on immune cells. They share significantly conserved amino acid sequences within the extracellular domains, but have both activating and inhibitory members. The inhibitory receptors possess the immunoreceptor tyrosine-based inhibitory motif (ITIM) within the cytoplasmic region. In contrast, the activating receptors have short cytoplasmic regions and a positively charged residue (Arg or Lys) in the transmembrane domain to associate with an adaptor protein possessing the immunoreceptor tyrosine-based activation motif (ITAM). They are located in small gene clusters on a chromosome, and are usually expressed on overlapping immune cells.

Although most of the inhibitory receptors can bind to the endogenous ligands, it has been somewhat more difficult to identify the ligands of the activating receptors (**Table 1**), and their functions have not been determined. The typical immune cell receptor-ligand interaction is quite weak (K_d in the μM range; **Table 2**), with fast association and dissociation. When the activating and inhibitory receptors specifically recognize the same ligand, the inhibitory receptors usually have higher affinity (**Table 2**). This indicates that the inhibitory receptors function in the maintenance of immunological tolerance by the recognition of self-ligands, such as major histocompatibility complex class I molecules [MHCIs, also referred to as human leukocyte antigen class I molecules (HLAIs)]. Especially, in natural killer (NK) cells, which lack a gene arrangement system to recognize foreign antigens, the inhibitory receptors for MHCIs recognize and eliminate cells that fail to express MHCIs, due to viral infections or tumor formation. This hypothetical mechanism is known as the "missing-self hypothesis" (Ljunggren and Karre, 1990). Based on this hypothesis, Valiante et al. (1997) suggested a model in which each NK cell expresses at least one inhibitory receptor for the classical MHCIs, to avoid killing healthy self-cells. However, a subpopulation of human NK cells that lack inhibitory receptors for self-MHCIs was identified (Cooley et al., 2007). Furthermore, there are populations of "licensed" and "unlicensed" NK cells, which are exposed and not exposed to self ligands for inhibitory receptors, respectively. Unlicensed NK cells are basically hyporesponsive, but seem to have important roles in tumor elimination and viral clearance.

The immune system is considered to be tightly regulated by the balance between the activating and inhibitory signals through these paired receptors, and dysregulation of this balance often causes autoimmunity, allergy, and various infectious diseases.

IMMUNOGLOBULIN-LIKE RECEPTORS

The immunoglobulin (Ig)-like receptors include the killer cell Ig-like receptors (KIRs), leukocyte Ig-like receptors (LILRs), murine paired Ig-like receptors (PIRs), Fc receptor, leukocyte-associated inhibitory receptors (LAIRs), NKp46, and so on. They have several conserved extracellular domains possessing a characteristic Ig-fold, consisting of 70–110 amino acids with a sandwich-like structure formed by two sheets of antiparallel β strands. A large number of Ig-like receptor genes are located within the leukocyte receptor complex (LRC) on human chromosome 19 and on mouse chromosome 7, the syntenic region in the mouse. Here, we discuss the ligands and the molecular recognition of several Ig-like receptors.

KILLER CELL IMMUNOGLOBULIN-LIKE RECEPTORS

The KIRs are expressed on NK cells and some subsets of T cells, and consist of 15 functional inhibitory (KIR2DL and KIR3DL) and

Table 1 | Paired receptors and their ligands.

Receptor	Function	Endogenous ligand	Non-self ligand
KIR2DL1	Inhibitory	HLA-C (group 2)	
KIR2DL2	Inhibitory	HLA-C (group 1, a subset of group 2), some HLA-B	
KIR2DL3	Inhibitory	HLA-C (group 1, a subset of group 2), some HLA-B	
KIR2DL4	Activating?	HLA-G	
KIR2DL5	Inhibitory	?	
KIR3DL1	Inhibitory	HLA-Bw4, a subset of HLA-A	
KIR3DL2	Inhibitory	HLA-A3, A11	CpG ODN
KIR3DL3	Inhibitory	?	
KIR2DS1	Activating	HLA-C (group 2)	
KIR2DS2	Activating	HLA-C (group 1)?	
KIR2DS3	Activating	HLA-C (group 2)?	
KIR2DS4	Activating	A subset of HLA-Cw4, A11	
KIR2DS5	Activating	?	
KIR3DS1	Activating	HLA-Bw4, a subset of HLA-A?	
LILRB1	Inhibitory	HLA-A, B, C, E, F, G	CMV UL18
LILRB2	Inhibitory	HLA-A, B, C, E, F, G, ANGPTLs (Zheng et al., 2012)	
LILRB3-5	Inhibitory	?	
LILRA1	Activating	HLA-B27, HLA-CfHC	
LILRA2, 4–6	Activating	?	
LILRA3	Soluble	HLA-CfHC, HLA-A, HLA-G	
PIR-B	Inhibitory	MHCI (Nakamura et al., 2004)	
		Nogo, MAG, OMgp (Atwal et al., 2008)	
		ANGPTLs (Zheng et al., 2012)	
PIR-A	Activating	MHCI (Nakamura et al., 2004)	
PILRα	Inhibitory	CD99, PANP, NPDC1, COLEC12	HSV-1 gB
PILRβ	Activating	CD99	HSV-1 gB
SIRPα	Inhibitory	CD47, SP-A, SP-D	
SIRPβ	Activating	SP-D	
SIRPγ	No signal	CD47	
DCIR	Inhibitory	?	
DCAR	Activating	?	
NKRP1-A	Activating	LLT1	
NKRP1-D	Inhibitory	Clrb (Iizuka et al., 2003)	
NKRP1-C, F	Activating	Clrg (Iizuka et al., 2003)	
Ly49A, C, I, etc	Inhibitory	H-2	CMV m157
Ly49D, H, etc	Activating		CMV m157, others?
CD94/NKG2A	Inhibitory	HLA-E (Qa-1)	
CD94/NKG2C, E	Activating	HLA-E (Qa-1)	Others?
MAIR-I	Inhibitory	?	
MAIR-II	Activating	?	
CD200R1	Inhibitory	CD200 (Wright et al., 2003)	
CD200R3,4	Activating	?	

ANGPTLs, angiopoietin-like proteins; fHC, free heavy chain; Nogo, neurite outgrowth inhibitor; MAG, myelin-associated glycoprotein; OMgp, oligodendrocyte myelin glycoprotein; Clrb, C-type lectin-related molecule b.

activating (KIR2DS and KIR3DS) receptors. The *KIRs* are divided into two major haplotypes (haplotype A and haplotype B) and are encoded together with the *LILRs* (described later), forming a gene cluster within the LRC (**Figure 1A**). The *KIR* family is highly polymorphic, with not only nucleotide sequence polymorphisms but also the presence/absence of each locus. The KIRs basically recognize the classical MHCIs (HLA-A, -B, or -C) in an allele-specific fashion. The KIRs are classified into two structural groups, KIR2D and KIR3D, which have two and three Ig-like domains (D1–D2, D0–D2, or D0–D1–D2) in the extracellular region, respectively (**Figure 2**).

KIR2DL1 specifically binds to HLA–C group 2 molecules (Asn77 and Lys80), while KIR2DL2/2DL3 bind to HLA–C group 1 molecules (Ser77 and Asn80; Parham, 2005). The ligands of the KIR2DSs reportedly recognize the same MHCI molecules as those bound by their related inhibitory KIRs (Parham, 2005). In contrast

Table 2 | Examples of binding affinities of receptor-ligand interactions.

Receptor	Ligand	K_d (μM)	Reference
LILRB1	HLA-G1	2.0	Shiroishi et al. (2003)
LILRB1	HLA-B35	8.8	Shiroishi et al. (2003)
LILRB1	HLA-Cw4	6.5	Shiroishi et al. (2003)
LILRB1	UL18	0.0021	Chapman et al. (1999)
LILRB2	HLA-A11	45	Shiroishi et al. (2003)
LILRB2	HLA-G1	4.8	Shiroishi et al. (2003)
LILRB2	HLA-B35	26	Shiroishi et al. (2003)
LILRB2	HLA-Cw4	14	Shiroishi et al. (2003)
LILRB2	HLA-Cw7	26	Shiroishi et al. (2003)
KIR2DL1	HLA-Cw4	7.2	Stewart et al. (2005)
KIR2DS1	HLA-Cw4	30	Stewart et al. (2005)
KIR2DL3	HLA-Cw7	7.0	Maenaka et al. (1999)
KIR3DS1	HLA-B27	7.0	Li et al. (2010)
PILRα	CD99	2.2	Tabata et al. (2008)
PILRβ	CD99	85	Tabata et al. (2008)
SIRPα	CD47	∼2.0	Brooke et al. (2004)
SIRPγ	CD47	∼23	Brooke et al. (2004)
NKRP1	LLT1	48	Kamishikiryo et al. (2011)
NKG2A/CD94	HLA-E	0.8–12.4	Kaiser et al. (2008)
NKG2C/CD94	HLA-E	5.2–18.2	Kaiser et al. (2008)

to the T cell receptors (TCRs), which recognize a wide area of the bound peptide and its surrounding area (α1 and α2 helices) of the MHCIs (**Figure 3A**), KIR2Ds bind to the C-terminal site of the bound peptide (Maenaka et al., 1999; Boyington et al., 2000; Fan et al., 2001). This peptide-dependent recognition (**Figure 3B**) is relatively less specific than that of the TCRs.

KIR3DL1 binds to HLA–B with the Bw4 epitope, determined by amino acid positions 77–83 (Parham, 2005). KIR3DL2 recognizes some HLA-A alleles (Parham, 2005). A recent structural study of the KIR3D-MHCI complex revealed that, while the additional N-terminal Ig-like domain (D0) bound to the bottom of the α2 and α3 domains, the C-terminal two Ig-like domains (D1 and D2) exhibited essentially the same binding mode as the KIR2Ds (Vivian et al., 2011). This explains the common peptide-dependent MHCI recognition of the KIR members. On the other hand, the KIRs have both inhibitory and activating members, and basically the activating KIRs exhibit much lower or non-detectable affinity to MHCIs than the inhibitory ones. It is potentially possible that some peptides can bind more strongly to the activating KIRs than the inhibitory ones, even though Stewart et al. (2005) demonstrated that most (or maybe all) peptides did not follow this characterization. Interestingly, recent reports demonstrated that the peptide mutations are likely to play a pivotal role in regulating human immunodeficiency virus (HIV) infection, by mediating KIR recognition (Thananchai et al., 2009; Alter et al., 2011). This illustrates some similarity between the KIR and TCR functions, but in the opposite way (KIR may have a more inhibitory role, but that of TCR is stimulatory).

Unexpectedly, KIR3DL2 was recently found to bind to the microbial CpG oligonucleotide (ODN), and the D0 domain is primarily involved in this recognition. The internalization of the KIR3DL2-ODN complex causes the activation of NK cells through toll-like receptor 9 (TLR9) signals (Sivori et al., 2010). As a novel ligand recognition system, KIR would directly bind to microorganisms and take advantage of non-self ligands in order to regulate the host immune system.

Recently, KIR2DS2 and KIRDS4 were reported to be up-regulated after hematopoietic cell transportation, and their up-regulations were significant in cytomegalovirus viremia (Gallez-Hawkins et al., 2011). This suggested that the expression levels of KIRs are also important for controlling NK cell or T cell function. Furthermore, the KIR expression on cord blood T cells was induced during a human congenital infection with *Trypanosoma cruzi*, possibly by epigenetic mechanisms.

LEUKOCYTE IMMUNOGLOBULIN-LIKE RECEPTORS

The Leukocyte immunoglobulin-like receptor (LILR; LIR, ILT, CD85) family was initially identified as the cellular counter structure to the viral UL18 protein, an MHCI homolog expressed by human cytomegalovirus (hCMV; Cosman et al., 1997). To date, 13 *LILR* family genes, including two pseudogenes (*LILRP1* and *LILRP2*), have been identified. The LILR family members can be divided into three classes: the inhibitory LILRs (LILRB1, -B2, -B3, -B4, -B5) with ITIM-like sequences, the activating LILRs (LILRA1, -A2, -A4, -A5, -A6) with a positively charged Arg residue in the transmembrane domain pairing with the FcRγ chain containing an ITAM, and the soluble LILR (LILRA3) with no transmembrane region. The LILRs have a broad cellular distribution that includes NK, T, and B lymphocytes, as well as myelomonocytic cells such as macrophages, mast cells, and dendritic cells.

The *LILR* family genes are encoded within the LRC on human chromosome 19q13 (**Figure 1A**). In the syntenic region of the mouse, at the proximal end of chromosome 7, the *LILR* gene-orthologous *Pir-a* and *Pir-b* are located. There are two clusters of *LILR* genes (*LILR* centromeric and *LILR* telomeric) that are transcribed in opposite directions (**Figure 1A**). The similarity of their amino acid sequences with that encoded by *KIR* suggests that the *LILRs* and *KIRs* are related by a recent gene duplication event. *LILR* genes have been found in a wide variety of species and are more stable in number, in contrast to the *KIR* genes. However, in addition to the deletion of the *LILRA3* gene which is absent in some individuals, recent studies revealed that *LILRA3* and *LILRA6* show high levels of genetic diversity, with decreased copy numbers in Asians (Hirayasu et al., 2008) and increased variability in Africans (Sudmant et al., 2010).

Although the KIR and LILR protein families are structurally and functionally comparable, there are some distinguishing characteristics. LILRB1 and LILRB2 bind to a variety of MHCIs through two N-terminal extracellular domains (D1 and D2; Borges et al., 1997; Colonna et al., 1997; Cosman et al., 1997). They recognize MHCIs on target cells to mediate inhibitory signals, to prevent the killing of normal cells expressing MHCIs. In other words, abnormal cells expressing few or no MHCIs can activate the cellular function of LILRB1/B2-positive leukocytes, due to the lack of the LILRB1/B2-mediated inhibitory signal. LILRA1 and LILRA3 also bind to some MHCIs (Allen et al., 2001; Ryu et al., 2011). Whereas most KIRs recognize discrete polymorphic epitopes within the α1 and α2 domains of MHCIs (**Figure 3B**),

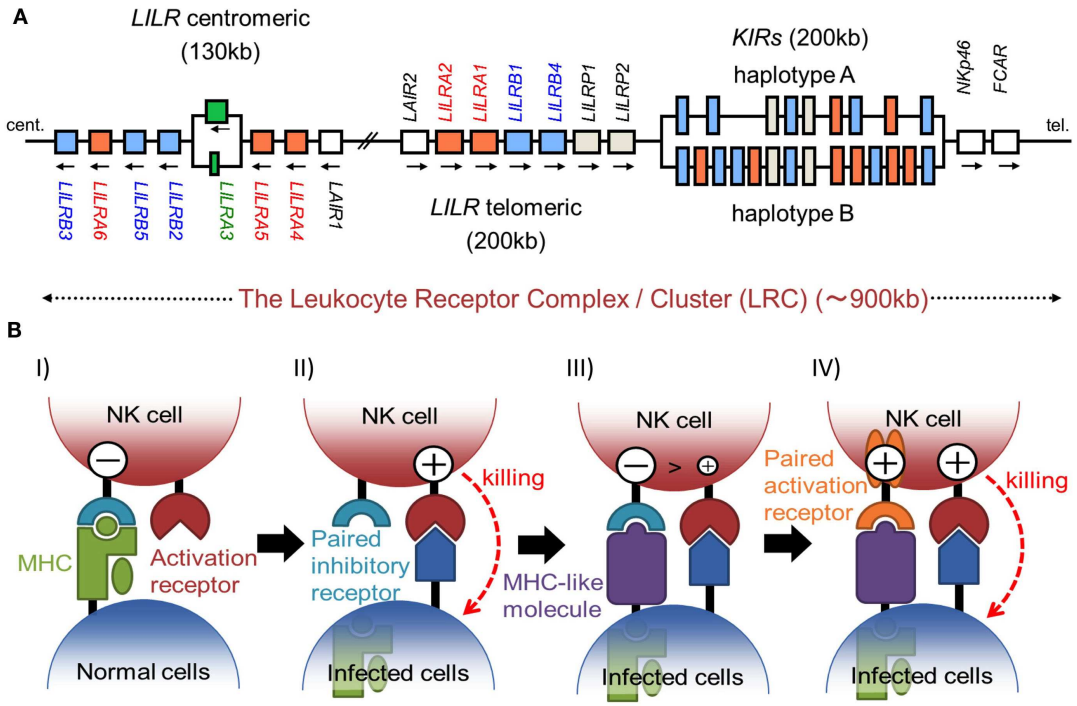


FIGURE 1 | (A) Schematic representation of the LRC on human chromosome 19q13.4. A large number of Ig-like receptor genes, including two clusters of *LILR* loci and a cluster of *KIR* loci, are encoded within the LRC. Arrows indicate the direction of transcription for each gene. These loci have evolved by multiple duplications, and the two *LILR* clusters are likely to have been generated by the inverse duplication of an ancient one. **(B)** The hypothesis of paired receptor family evolution. I) The NK cell possesses at least one inhibitory receptor (cyan), and the inhibitory signals through it protect the normal self-cells from NK cell killing. II) In the infected cells, the low level expression of MHCIs induces the NK killing, in a system called the "missing-self hypothesis." III) In order to escape the host NK cytotoxicity, some viruses acquired the expression of MHCI-like molecules (purple), which bind to the inhibitory receptors. IV) On the other hand, NK cells express activation receptors (orange), which evolved from the related inhibitory receptors to trigger NK cell activation.

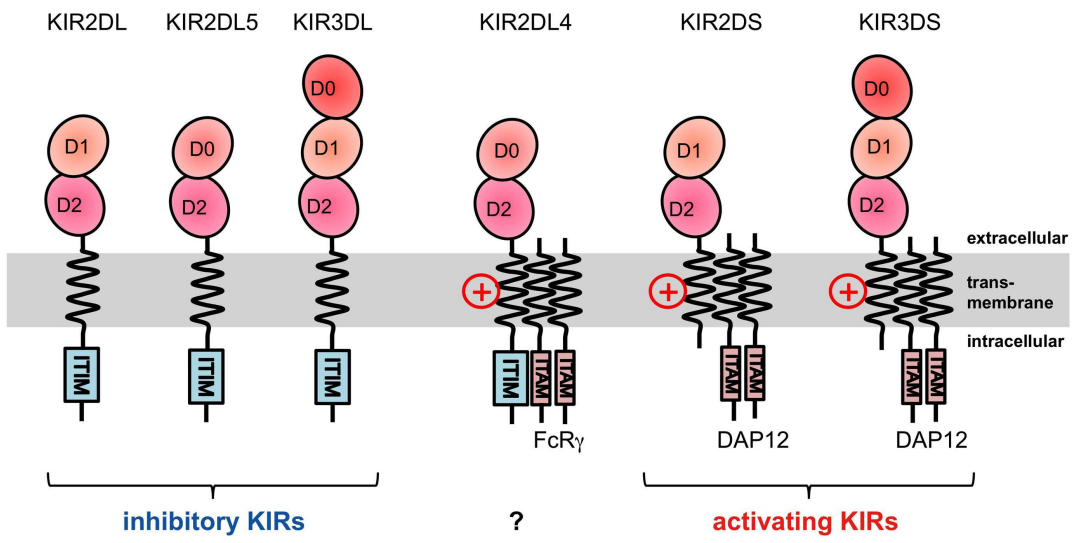


FIGURE 2 | Domain configuration of the KIRs. The extracellular Ig-like domain is classified into three types, D0, D1, and D2, dependent on the sequence homology. KIR2DL4 possesses an ITIM motif, but also associates with the FcRγ chain.

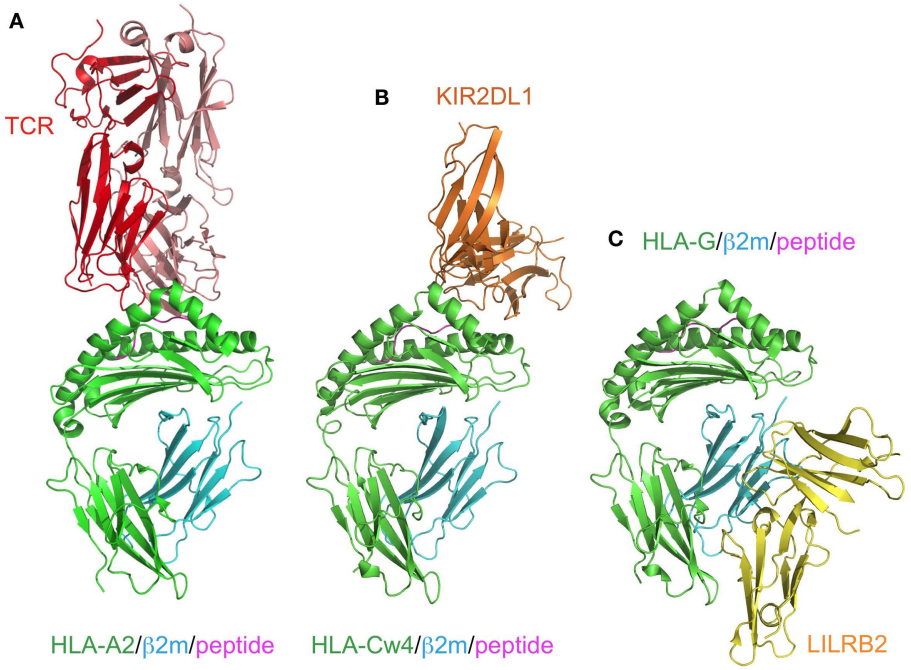


FIGURE 3 | Comparison of the recognition modes of the MHCI/MHCI receptors. The complex structures of HLA-A2 and TCR (red; A, PDB ID:2VLR), HLA-Cw4 and KIR2DL1 (orange; B, PDB ID:1IM9), and HLA-G and LILRB2 (yellow; C, PDB ID: 2DYP). The MHCIs (heavy chain in green, β2m in cyan, peptide in magenta) are recognized in different manners. **(A)** TCR binds to the center of the peptide and the α1–α3 domain of HLA-A2. **(B)** KIR2DL1 binds to both the α1 and α2 helices of HLA and the C-terminal end of the peptide. This binding region contains the 77N/S and 80K/N residues, which determine the ligand specificity. **(C)** LILRB2 binds to the α3 domain and β2m, which are conserved regions among the MHCIs.

a characteristic consistent with their narrow binding specificities, LILRB1/B2 binding is mediated via a site in the conserved α3 and β2m domains of the MHCI molecule (**Figure 3C**; Willcox et al., 2003; Shiroishi et al., 2006). Moreover, LILRB1/B2 effectively compete with CD8 for MHCI binding, and modulate $CD8^+$ T cell activation by blocking CD8 binding as well as by recruiting inhibitory molecules through their ITIMs (Shiroishi et al., 2003). This system was also observed in the binding of the mouse LILRB homolog, PIR-B, to MHCIs (Shiroishi et al., 2003).

As described above, LILRB1 also binds to the UL18 protein with much higher affinity than the MHCIs (Cosman et al., 1997), but LILRB2 and LILRA1 do not (Borges et al., 1997). UL18 is a highly glycosylated protein (**Figure 4A**) sharing 25% sequence identity with MHCIs (Beck and Barrell, 1988). Although the LILRB1/UL18 complex structure revealed that the binding mode was conserved with those of the LILRB1/MHCIs (**Figures 4A,B**), the residues within the α3 domain differed and created a more favorable binding region at the interface with LILRB1 (Yang and Bjorkman, 2008). Moreover, the 13 potential *N*-glycosylation sites effectively covered the interaction sites for the potential UL18 binding partners, including the α1–α2 domain binding to KIR and TCR and the part of α3 domain binding to CD8. On the other hand, only the binding region of LILIRB1 remains exposed (**Figure 4A**). These structural features demonstrated how the viral protein UL18 can effectively compete with the host ligands for LILRB1, to regulate the host's immune response.

PAIRED TYPE 2 IMMUNOGLOBULIN-LIKE RECEPTORS

The Paired type 2 immunoglobulin-like receptors (PILRs) are expressed mainly on immune cells and have one Ig-like domain in the extracellular region, with either an ITIM in the intracellular domain (inhibitory receptor, PILRα) or a positively charged amino acid in the transmembrane region associated with the activating subunit, DNAX activating protein of 12 kDa (DAP12; activating receptor, PILRβ; Fournier et al., 2000). The PILRs recognize sialylated *O*-linked sugar-modified mucin and mucin-like molecules, such as CD99 (Wang et al., 2008), PILR-associating neural protein (PANP; Kogure et al., 2011), and two newly identified ligands, neuronal differentiation and proliferation factor-1 (NPDC1), and collectin-12 (COLEC12; Sun et al., 2012), as physiological ligands. Notably, the PILR-ligand family is further expanding, as the report by Sun et al. (2012) described that the PILRs can also bind to immune cells that do not express any identified PILR-ligands. As mentioned above, PILRα shows higher affinity to these ligands than PILRβ, which is typical for paired receptors in the immune system (**Table 2**). On the other hand, recent reports revealed that PILRα is the receptor for herpes simplex virus 1 (HSV-1), by binding to its glycoprotein B (gB; Satoh et al., 2008). HSV-1 gB can utilize the inhibitory member, PILRα, as an entry receptor, and thus it is probably beneficial for HSV-1 infection. The PILRs recognize physiological and viral proteins in similar but non-identical manners, and notably exhibit both sugar- and peptide-dependent binding modes, which are quite unique for sugar-protein interactions. Even though the

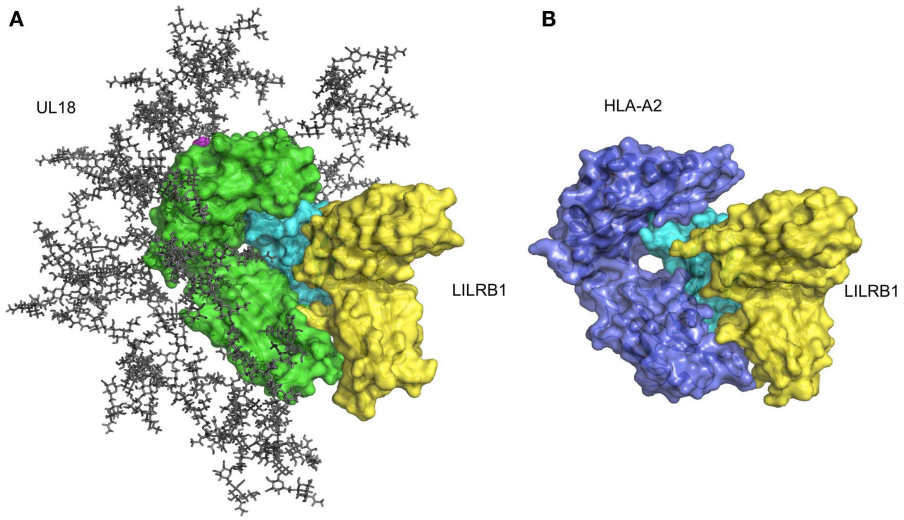


FIGURE 4 | Structure of UL18/LILRB1 and comparison with the HLA-A2/LILRB1 complex. (A) The crystal structure of the UL18/LILRB1 complex (UL18 in green, β2m in cyan, peptide in magenta, LILRB1 in yellow; PDB ID: 3D2U) with complex carbohydrate models attached to the 13 potential *N*-glycosylation sites. The α1–α2 domains recognized by TCR or KIRs are highly glycosylated, and steric hindrance inhibits effective interactions. Meanwhile, the interface with LILRB1 still is exposed. **(B)** The crystal structure of the HLA-A2/LILRB1 complex (HLA-A2 and peptide in purple, β2m in cyan, LILRB1 in yellow; PDB ID: 1P7Q).

structural information about the PILRs is still lacking, mutagenesis studies suggested that the PILR-ligand recognition modes are somewhat similar to those of the sialic acid-binding Ig-like lectins (Siglec) family, which can bind to sialic acid (Tabata et al., 2008; Wang et al., 2008). However, the significant and unusual characteristic, that the PILRs cannot bind to either sugars or peptides only, and require both for binding, is largely unknown and rather unique. Future studies, especially crystallographic and NMR analyses, will clarify the molecular mechanisms of these binding systems.

SIGNAL REGULATORY PROTEINS

The Signal regulatory proteins (SIRP) family has three members, SIRPα, SIRPβ, and SIRPγ. SIRPα (SHPS-1, BIT, CD172A) is broadly expressed on myeloid cells, such as neutrophils, macrophages, and dendritic cells, as well as on neurons (Adams et al., 1998). SIRPα interacts with CD47, which expresses in hemopoietic cells, epithelial cells, and endothelial cells, as well as in brain and mesenchymal cells, resulting in the transmission of the inhibitory signal through the SH2-domain-containing protein tyrosine phosphatases 1 and 2 (SHP1 and 2, respectively), and finally causes reduced phagocytosis activity in macrophages and cytokine production in various cells (Barclay and Brown, 2006). In this sense, SIRPα plays an important role in immune suppression. In contrast, SIRPβ can potentially generate activation signaling by associating with DAP12, but it cannot interact with CD47. In addition, SIRPγ binds to CD47 with 10-fold lower affinity as compared to SIRPα, but lacks a signaling motif (Barclay and Brown, 2006). An X-ray crystallographic analysis indicated that the difference in the binding affinities of the SIRPs with CD47 is due to the subtle differences in the loops, with direct and indirect effects (Hatherley et al., 2008).

The other function of the SIRPs is to bind with Surfactant Protein D (Sp-D). Sp-D is an important component of the pulmonary surfactant involved in host innate immunity, and is capable of binding most Gram-negative bacteria as well as several Gram-positive bacteria, leading to increased opsonization of bacteria. On the other hand, the binding of Sp-D to SIRPα transmits the immune suppression signals, resulting in decreased cytokine production (Gardai et al., 2003). Interestingly, the SIRPα binding to SP-D was competed with lipopolysaccharide (LPS; Fournier et al., 2012). These observations indicated that the anti-inflammation signals through SIRPα are present in the absence of pathogens, and once pathogens are present, Sp-D binds preferably to LPS or other bacterial carbohydrates, and then induces the host innate immunity. Under such conditions, the absence of ligand binding to SIRPα also elicits an increase in inflammation (Gardai et al., 2003). One recent report demonstrated that SIRPβ also binds to Sp-D, but with slightly lower affinity as compared to SIRPα (Fournier et al., 2012). These results suggested that SIRPs exhibit self/non-self discrimination and cooperatively modulate the immuneresponses.

C-TYPE LECTIN-LIKE RECEPTORS

C-type lectin-like receptors (CLRs) are expressed on the cell surface of various immune cells, to regulate the innate immune systems. The term "C-type lectin" means Ca^{2+} dependent carbohydrate-binding lectin. The CLRs contain a conserved motif, either EPN (Glu-Pro-Asn) or QPD (Gln-Pro-Asp). This motif is located in a structurally conserved loop, which is stabilized by a disulfide bond with another conserved loop. The EPN motif confers specificity for mannose-based ligands, whereas the QPD motif is typical of the galactose-specific Carbohydrate Recognition Domain (CRD; Zelensky and Gready, 2005).

The carbonyl side chains of these amino acid residues coordinate Ca^{2+}, form hydrogen bonds with individual monosaccharides, and determine binding specificity. Due to the versatile recognition ability of the CLRs (described in below), they are known as pathogen associated molecular patterns (PAMPs) recognition receptors. The CLRs are primarily involved in detecting pathogens and subsequently triggering signaling pathways to evoke various immune reactions. Meanwhile, a few CLRs are known to act as immuno-repressive receptors. Interestingly, it was demonstrated that Macrophage inducible C-type lectin (Mincle, also called CLEC4E) recognized not only the sugar components from pathogens through the CRD but also proteins from pathogens or self, through sites other than the CRD (discussed in detail below). Here, we describe the detailed functions and structures of several CLRs, including orphan CLRs that have yet to be characterized, to shed light on the molecular mechanisms of the ligand recognition by the paired-receptor-type CLRs.

CD94/NKG2

The CD94/NKG2 receptors are expressed on the surfaces of a greater part of NK cells and some subsets of $CD8^+$ T cells, and belong to the CLRs. While CD94 is encoded by a single gene and has extremely low polymorphism, the NKG2 molecules have five isotypes (NKG2A, C, D, E, and F) and two splice variants, NKG2B and NKG2H, derived NKG2A and NKG2E, respectively. Five NKG2 molecules (NKG2A, B, C, E, and H) have been shown to form disulfide-linked heterodimers with CD94. CD94/NKG2A and B mediate inhibitory signaling through the ITIM of the cytosolic region in NKG2s. In contrast, CD94/NKG2C, E, and H induce activation signaling through the interaction between a Lys residue within their transmembrane region and a negatively charged residue in the ITAM-containing adaptor molecule, DAP12.

The ligand of most CD94/NKG2s is the non-classical MHCI, HLA-E (Borrego et al., 1998; Braud et al., 1998; Lee et al., 1998; Brooks et al., 1999; **Table 1**). HLA-E is expressed in all nucleated cells and has few polymorphisms, as compared to the classical MHCI. HLA-E mainly presents the peptides derived from the leader sequences of other MHCIs. Therefore, NK cells monitor the MHCI expression level through the interactions between the inhibitory NKG2A/CD94s and HLA-E. Thus, NK cells interact with healthy cells, which show normal expression levels of HLA-E, resulting in the inhibition of NK cell killing activity.

The structure of CD94/NKG2A in complex with HLA-E loaded with an HLA-G derived peptide has been determined (Kaiser et al., 2008; Petrie et al., 2008; **Figure 5A**). NKG2A and CD94 interact with the α1 and α2 helices of the peptide-binding region of HLA-E, respectively, with charge complementarity. The presented peptide is also recognized by CD94/NKG2A, while CD94 is mainly recognized HLA-E and the peptide, as compared to NKG2A. The Arg (P5) and Phe (P8) residues of the peptides contribute to binding with CD94/NKG2A (**Figure 5B**). P5Arg is conserved in the leader sequence of MHCI, and its replacement with a Lys abolished the interaction with CD94/NKG2A. Hydrophobic amino acid residues, such as Phe and Leu, are conserved in P8 of the leader sequence, and their replacement with Lys led to a dramatic reduction in the binding ability of HLA-E with CD94/NKG2. These data indicated that the Arg in P5 and the hydrophobic amino acid residue in P8 are both indispensable for the interaction with CD94/NKG2A.

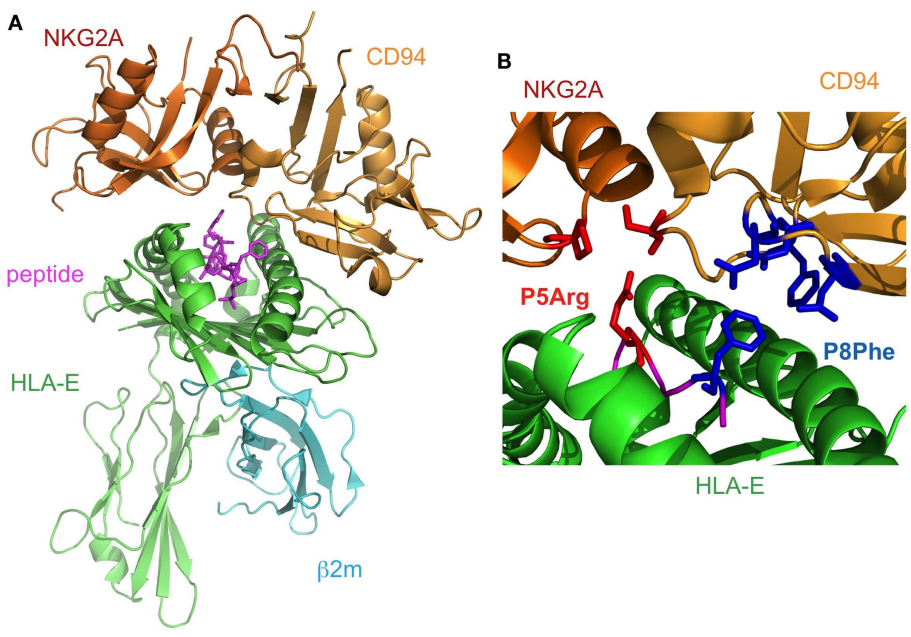


FIGURE 5 | The complex structure of NKG2A/CD94 and HLA-E. (A) The overall structure of the NKG2A/CD94/HLA-E complex (PDB ID: 3CDG). NKG2A/CD94 recognizes the α1–α2 domain of HLA-E containing a peptide. **(B)** The interface of NKG2A/CD94 and HLA-E. The residues interacting with P5Arg of the peptide are depicted by red stick models, and P8Phe is shown by a blue stick model.

The interaction between CD94/NKG2 and HLA-E is associated with a wide range of diseases, such as virus infection and cancer. For instance, HCMV utilizes the host's CD94/NKG2 system to grow. The leader sequence of the HCMV protein, UL40, is identical to that of HLA-Cw3. The CD94/NKG2A on NK cells recognizes the HLA-E associated peptide, which is derived from the leader sequence of the HCMV protein, on the infected cells, and therefore its interaction inhibits NK cells from attacking infected cells (Tomasec et al., 2000; Ulbrecht et al., 2000). In addition, in renal cell carcinoma-infiltrating NK cells, the expression level of CD94/NKG2A is relatively high. This led to the proposal that tumor cells control the expression level of CD94/NKG2A, and consequently the function of NK cells (Schleypen et al., 2003). It might be plausible that other intracellular pathogens can also utilize the CD94/NKG2A systems to escape from the host's immune system.

DENDRITIC CELL IMMUNORECEPTOR/DENDRITIC CELL IMMUNOACTIVATING RECEPTOR

Dendritic cell immunoreceptor (DCIR; CLEC4A) is one of the repressive CLRs. DCIR is expressed on the surface of various immune cells, such as dendritic cells, monocytes, macrophages, B cells and neutrophils (Bates et al., 1999). Although DCIR reportedly binds HIV-1, no physiological ligands for DCIR have been identified (Lambert et al., 2008). The ITIMs in the cytoplasmic tails of DCIR serve to recruit SHP1 or SHP2 after ligand binding (Richard et al., 2006). The Syk and Src kinases (i.e., Src, Fyn, and Hck), as well as the PKC-γ MAP kinases (i.e., Erk1/2 and p38), are reportedly involved in the subsequent signaling pathway and finally inhibit TLR8-mediated IL-12 and TNF-α production significantly (Lambert et al., 2011). However, the precise mechanism of this inhibition is still unknown.

DCIR expression on neutrophils was reportedly down-regulated by TNF-α, IL-1α, and LPS stimulation, but was not affected by anti-inflammatory stimuli, including IL-4, IL-10, and IL-13 (Bates et al., 1999; Richard et al., 2002). These results suggested that DCIR may be down-regulated during pathogen exposure and inflammation.

A recent study showed that *Dcir*$^{-/-}$ mice developed joint abnormalities, such as swelling and redness, at an early age, and these abnormalities eventually progressed to joint deformity and ankylosis (Fujikado et al., 2008). The *Dcir*$^{-/-}$ mice developed sialadenitis, which is characterized by the accumulation of lymphocytes in the interstitium and the destruction of the small duct associated with mononuclear cell infiltration. The number of activated CD4^{+} T cells, the expression of IL-4 and IL-10, and the production of IgG1 and IgG3 were increased in *Dcir*$^{-/-}$ mice. Furthermore, in *Dcir*$^{-/-}$ mice, stimulation with granulocyte-macrophage colony–stimulating factor (GM-CSF) activated the phosphorylation of STAT5 and effectively differentiated bone marrow-derived cells to dendritic cells, as compared to wild type mice. These results suggested that DCIR regulates the proliferation of dendritic cells and is involved in maintaining immune self-tolerance.

On the other hand, mouse DCIR shares substantial sequence homology (91% aminoacid identity) in the extracellular region with its activating counter member, Dendritic cell immunoactivating receptor (DCAR). DCAR is expressed similarly in tissues to DCIR, but its short cytoplasmic portion lacks a signaling motif such as an ITIM (Fujikado et al., 2008). Instead, an Arg residue is present in the transmembrane region of DCAR, which participates in the association with the FcRγ chain and finally activates immune cells. Neither the human ortholog nor the ligands for DCAR have been identified yet.

NKR-P1 (CD161)

NKR-P1 (CD161) is expressed on the surfaces of NK cells and subsets of T cells. Human NKR-P1 reportedly interacts with Lectin-like transcript-1 (LLT1, also called CLEC2D), which is expressed on many cell lines and on activated primary B cells (Aldemir et al., 2005; Rosen et al., 2005). The *NKR-P1* and *LLT1* genes are adjacent on human chromosome 12, and coordinately regulate the immune response. Rodents possess several *Nkrp1* genes for the activating (NKRP1-A, C, and F) and inhibitory (NKRP1B and G) receptors, while in contrast, there is only a single inhibitory *NKR-P1A* gene in human. LLT1 (also known as Clr-b in mouse) on target cells can inhibit NK cytotoxicity, by interacting with NKR-P1 on NK cells (Rosen et al., 2008). In rodents, NKR-P1C reportedly associates with the FcRγ chain, inducing not only cytotoxicity but also IFN-γ production. These data suggested that, although anon-self ligand has not been identified, the components of pathogens or dead cells may potentially stimulate NK cells through NKR-P1, and are eliminated by NK cells themselves and by other immune cells (Arase et al., 1997). The detailed molecular mechanisms of both the activating and inhibitory signaling pathways via NKR-P1 have not been characterized, but interestingly, the acid sphingomyelinase reportedly binds to the cytosolic region of NKR-P1 and participates in NK cell resistance to apoptosis (Pozo et al., 2006). We recently analyzed the molecular basis of the interaction between NKR-P1A and LLT1 (**Figure 6A**; Kamishikiryo et al., 2011), and proposed a model of the NKR-P1/LLT complex. The constructed model suggested that the membrane-distal head region of NKR-P1 is a new target to inhibit the NKR-P1-LLT1 interaction, potentially leading to the regulation of autoimmune and chronic inflammatory disorders.

OTHER C-TYPE LECTIN-LIKE RECEPTORS

Other CLRs have been identified and functionally investigated. Macrophage inducible C-type lectin (Mincle, also called CLEC4E) is a type II transmembrane C-type lectin receptor expressed in macrophages, dendritic cells and monocytes. Mincle was initially identified as a gene up-regulated by LPS stimulation (Matsumoto et al., 1999). The first identified ligand of Mincle was the spliceosome associated protein 130 (SAP130), the endogenous protein released from necrotic cells (Yamasaki et al., 2008). Intriguingly, Mincle also reportedly recognized malassezia species and *Mycobacterium tuberculosis* (Ishikawa et al., 2009; Yamasaki et al., 2009). While the ligand structures of the malassezia species are not known, that of *M. tuberculosis* was identified as trehalose-1,1-dimycolate (TDM). The malassezia species and *M. tuberculosis* are both recognized through the CRD of Mincle. Following the binding of either SAP130, malassezia species or TDM, Mincle associates with the ITAM-bearing FcRγ chain. This association leads

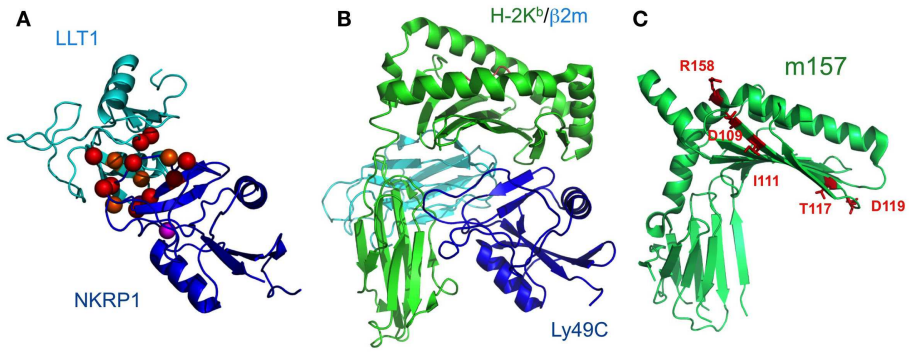


FIGURE 6 | Schematic model of LLT1 recognition by NKRP1 and comparison of the structures of the Ly49-MHCI complex and m157. (A) The model structures of LLT1 (cyan) and NKRP1 (blue) are shown by ribbon models. Residues that may contribute to the interaction with LLT1 are shown as spheres, with detrimental effects in red, and modest effects in orange. Magenta spheres indicate the pair of residues that showed detrimental effects when mutated independently, but restored the binding when mutated simultaneously (Kamishikiryo et al., 2011). **(B)** The crystal structure of Ly49C with H-2K^b (H-2K^b in green, β2m in cyan, Ly49C in blue; PDB ID: 1P4L). Ly49C interacts with the β2m subunit and the α3 domain of H-2Kb, but not with the peptide-binding region. The crystal structure of m157 (PDB ID; 2NYK). **(C)** The mutation sites that identified the Ly49H binding residues are mapped and are shown in stick style (red). The residues did not overlap with the interface of the Ly49-MHCI complex structure.

to the phosphorylation of the ITAM of the FcRγ chain and the subsequent recruitment of Syk, activating the caspase recruitment domain family, member 9 (CARD9)-mediated NF-κB signaling pathway to promote the expression of TNF and IL-6. A recent report revealed that Mincle would likely playa non-redundant role for T cell immune responses to infection by microbes (Schoenen et al., 2010).

The Dendritic Cell-Specific Intercellular adhesion molecule-3-Grabbing Non-integrin (DC-SIGN, also called CLEC4L and CD209) proteins have been extensively studied, due to their broad range recognition of pathogens and ligands from viruses to parasites, such as HIV-1, measles, dengue, SARS, *Helicobacter pylori*, *E. coli*, *Leishmania* spp., and *Schistosoma* egg (Sancho and Reis e Sousa, 2012). The binding of DC-SIGN with ligands from pathogens activates human myeloid dendritic cells through various pathways. Mannose-expressing *M. tuberculosis* and HIV-1 promote the activation of LARG and RhoA, which function as upstream activators of Raf-1 via DC-SIGN (Gringhuis et al., 2007, 2009; Hodges et al., 2007). This activation is mediated by the phosphorylation and acetylation of NF-κB subunit p65, which greatly enhances the transcriptional activity of NF-κB and results in the modulation of TLR4 signaling and the enhanced expression of IL-10, IL-12, and IL-6 (Gringhuis et al., 2007). In contrast, fucose-expressing pathogens, such as *H. pylori*, actively dissociated the KSR1–CNK–Raf-1 complex and enhanced the expression of IL-10, but down-regulated the expression of IL-12 and IL-6 in a Raf-1-independent, but LSP1-dependent, manner (Gringhuis et al., 2009). Notably, DC-SIGN cannot activate NF-κB by itself, and it modulates the p65 activity only when p65 is induced by another receptor upon the stimulation of mannose-expressing pathogens (Gringhuis et al., 2007). In summary, the signaling via DC-SIGN is tightly regulated by the characteristics of the ligands.

DC NK lectin group receptor-1 (DNGR, also called CLEC9A) was previously shown to function as a Syk-coupled C-type lectin receptor, to mediate the sensing of necrosis (Sancho et al., 2009). A recent report demonstrated that DNGR recognized exposed actin filaments from necrotic or damaged cells (Ahrens et al., 2012; Zhang et al., 2012). Mutational and crystallographic studies indicated that two exposed tryptophan residues in DNGR, which are conserved between human and mouse, are involved in the recognition of the actin filament. These residues are far from the C-type lectin domains, which function in stabilizing the structure of the protein, rather than being directly involved in the receptor-ligand interaction. So far, most of the ligands from pathogens are carbohydrate or carbohydrate-related products. However, several studies have clearly proved that CLR can interact with ligands through other regions than the C-type lectin domains. These results imply that non-carbohydrate ligands of the CLRs from pathogens will be discovered in the future.

DISCUSSION

Paired receptors are potentially dangerous, because activating receptors can disrupt homeostasis, thus threatening life. In the immune system, immune cells express such activating receptors on their cell surfaces; however, the education or licensing of these cells has been considered to require the expression of at least one inhibitory receptor to suppress inappropriate activation, at least in fully responsive mature cells. On the other hand, Arase et al. (2002) clearly showed excellent evidence for one paired receptor family, the Ly49 family. The susceptibility to mouse CMV (MCMV) depends on the mouse strain, and the protection of this virus is mediated by NK cells. Mice harboring only the inhibitory Ly49 family members, which bind to MHCIs (Held et al., 1996; Yu et al., 1996; Hanke et al., 1999) as well as the CMV MHCI homolog, m157 (Arase et al., 2002; Smith et al., 2002), cannot survive the CMV infection. However, other mice with the activating member, Ly49H, which binds to its m157 to activate the immune response, can evade CMV infection. Thus, m157 is the only known viral ligand binding to both inhibitory and activating receptors. Based on these observations, we have developed a scenario in which the activation receptor evolved from the related

inhibitory receptor in response to selective pressure imposed by the pathogen, thus providing the presence of diversified, paired receptors (**Figure 1B**). In accordance with this scenario, the activating receptor, KIR2DS2, has a relic of the ITIM sequence, which is inactivated by the direct introduction of a stop codon (Arase and Lanier, 2004). This hypothesis suggests that other activating receptors will also recognize pathogen-derived ligands. Although m157 forms a typical MHCI-fold, it neither presents a peptide nor associates with β2m (Adams et al., 2007; **Figures 6B,C**). A comparison of the crystal structures of the Ly49/H-2 complexes and m157 revealed that a different interaction interface from that of Ly49/H-2 would exist upon the binding of Ly49s to m157.

Furthermore, Barclay and Hatherley (2008) proposed an elegant counterbalance theory. Accumulating mutations on paired receptors are often unrelated to the binding regions to physiological ligands, which may support the idea that these mutations are targeted to the non-self molecules of infectious microorganisms. As described above, the ancestral paired receptors are likely inhibitory to suppress undesired immune responses, but this is essentially beneficial for microorganisms, if they can be used as not only entry receptors but also inhibitory ones to facilitate immune evasion. The KIRs are considered to have co-evolved with their ligands, the MHCIs. For example, different human populations have a reciprocal relationship between the KIR and HLA-C frequencies (Hiby et al., 2004), and the frequency of the KIR2DL3-HLA-C1 combination could be reduced in populations highly exposed to malaria, by natural selection (Hirayasu et al., 2012). These observations suggested that the paired receptors and their ligands co-evolved. Therefore, the precise understanding of the on-going activating and inhibiting balance of paired receptors can provide insight into the extent of the importance of each set of paired receptors for immune defense. In this sense, the development of small molecular-weight compounds or biopharmaceuticals targeting paired receptors can more easily and finely regulate the immune responses, as advanced therapy for either infectious diseases or tumorigenesis. Especially, we believe that future accumulating information relating the genetic, molecular, and structural bases for paired receptors will greatly contribute to the development of novel therapies with fewer side effects.

ACKNOWLEDGMENTS

We thank H. Fukuhara for helpful discussions. This work was partly supported by grants from New Energy and Industrial Technology Development Organization (NEDO), the Ministry of Education, Culture, Sports, Science, and Technology, and the Ministry of Health, Labor, and Welfare of Japan. Kimiko Kuroki was supported by a JSPS Research Fellowship for Young Scientists.

REFERENCES

Adams, E. J., Juo, Z. S., Venook, R. T., Boulanger, M. J., Arase, H., Lanier, L. L., et al. (2007). Structural elucidation of the m157 mouse cytomegalovirus ligand for Ly49 natural killer cell receptors. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10128–10133.

Adams, S., Van Der Laan, L. J., Vernon-Wilson, E., Renardel De Lavalette, C., Dopp, E. A., Dijkstra, C. D., et al. (1998). Signal-regulatory protein is selectively expressed by myeloid and neuronal cells. *J. Immunol.* 161, 1853–1859.

Ahrens, S., Zelenay, S., Sancho, D., Hanc, P., Kjaer, S., Feest, C., et al. (2012). F-actin is an evolutionarily conserved damage-associated molecular pattern recognized by DNGR-1, a receptor for dead cells. *Immunity* 36, 635–645.

Aldemir, H., Prod'homme, V., Dumaurier, M. J., Retiere, C., Poupon, G., Cazareth, J., et al. (2005). Cutting edge: lectin-like transcript 1 is a ligand for the CD161 receptor. *J. Immunol.* 175, 7791–7795.

Allen, R. L., Raine, T., Haude, A., Trowsdale, J., and Wilson, M. J. (2001). Leukocyte receptor complex-encoded immunomodulatory receptors show differing specificity for alternative HLA-B27 structures. *J. Immunol.* 167, 5543–5547.

Alter, G., Heckerman, D., Schneidewind, A., Fadda, L., Kadie, C. M., Carlson, J. M., et al. (2011). HIV-1 adaptation to NK-cell-mediated immune pressure. *Nature* 476, 96–100.

Arase, H., and Lanier, L. L. (2004). Specific recognition of virus-infected cells by paired NK receptors. *Rev. Med. Virol.* 14, 83–93.

Arase, H., Mocarski, E. S., Campbell, A. E., Hill, A. B., and Lanier, L. L. (2002). Direct recognition of cytomegalovirus by activating and inhibitory NK cell receptors. *Science* 296, 1323–1326.

Arase, N., Arase, H., Park, S. Y., Ohno, H., Ra, C., and Saito, T. (1997). Association with FcRgamma is essential for activation signal through NKR-P1 (CD161) in natural killer (NK) cells and NK1.1+ T cells. *J. Exp. Med.* 186, 1957–1963.

Atwal, J. K., Pinkston-Gosse, J., Syken, J., Stawicki, S., Wu, Y., Shatz, C., et al. (2008). PirB is a functional receptor for myelin inhibitors of axonal regeneration. *Science* 322, 967–970.

Barclay, A. N., and Brown, M. H. (2006). The SIRP family of receptors and immune regulation. *Nat. Rev. Immunol.* 6, 457–464.

Barclay, A. N., and Hatherley, D. (2008). The counterbalance theory for evolution and function of paired receptors. *Immunity* 29, 675–678.

Bates, E. E., Fournier, N., Garcia, E., Valladeau, J., Durand, I., Pin, J. J., et al. (1999). APCs express DCIR, a novel C-type lectin surface receptor containing an immunoreceptor tyrosine-based inhibitory motif. *J. Immunol.* 163, 1973–1983.

Beck, S., and Barrell, B. G. (1988). Human cytomegalovirus encodes a glycoprotein homologous to MHC class-I antigens. *Nature* 331, 269–272.

Borges, L., Hsu, M. L., Fanger, N., Kubin, M., and Cosman, D. (1997). A family of human lymphoid and myeloid Ig-like receptors, some of which bind to MHC class I molecules. *J. Immunol.* 159, 5192–5196.

Borrego, F., Ulbrecht, M., Weiss, E. H., Coligan, J. E., and Brooks, A. G. (1998). Recognition of human histocompatibility leukocyte antigen (HLA)-E complexed with HLA class I signal sequence-derived peptides by CD94/NKG2 confers protection from natural killer cell-mediated lysis. *J. Exp. Med.* 187, 813–818.

Boyington, J. C., Motyka, S. A., Schuck, P., Brooks, A. G., and Sun, P. D. (2000). Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. *Nature* 405, 537–543.

Braud, V. M., Allan, D. S., O'callaghan, C. A., Soderstrom, K., D'andrea, A., Ogg, G. S., et al. (1998). HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* 391, 795–799.

Brooke, G., Holbrook, J. D., Brown, M. H., and Barclay, A. N. (2004). Human lymphocytes interact directly with CD47 through a novel member of the signal regulatory protein (SIRP) family. *J. Immunol.* 173, 2562–2570.

Brooks, A. G., Borrego, F., Posch, P. E., Patamawenu, A., Scorzelli, C. J., Ulbrecht, M., et al. (1999). Specific recognition of HLA-E, but not classical, HLA class I molecules by soluble CD94/NKG2A and NK cells. *J. Immunol.* 162, 305–313.

Chapman, T. L., Heikeman, A. P., and Bjorkman, P. J. (1999). The inhibitory receptor LIR-1 uses a common binding interaction to recognize class I MHC molecules and the viral homolog UL18. *Immunity* 11, 603–613.

Colonna, M., Navarro, F., Bellon, T., Llano, M., Garcia, P., Samaridis, J., et al. (1997). A common inhibitory receptor for major histocompatibility complex class I molecules on human lymphoid and myelomonocytic cells. *J. Exp. Med.* 186, 1809–1818.

Cooley, S., Xiao, F., Pitt, M., Gleason, M., Mccullar, V., Bergemann, T. L., et al. (2007). A subpopulation of human peripheral blood NK cells that lacks inhibitory receptors for self-MHC is developmentally immature. *Blood* 110, 578–586.

Cosman, D., Fanger, N., Borges, L., Kubin, M., Chin, W., Peterson, L., et al. (1997). A novel immunoglobulin superfamily receptor for cellular and viral MHC class I molecules. *Immunity* 7, 273–282.

Fan, Q. R., Long, E. O., and Wiley, D. C. (2001). Crystal structure of the human natural killer cell inhibitory receptor KIR2DL1-HLA-Cw4 complex. *Nat. Immunol.* 2, 452–460.

Fournier, B., Andargachew, R., Robin, A. Z., Laur, O., Voelker, D. R., Lee, W. Y., et al. (2012). Surfactant protein D (Sp-D) binds to membrane-proximal domain (D3) of signal regulatory protein alpha (SIRPalpha), a site distant from binding domain of CD47, while also binding to analogous region on signal regulatory protein beta (SIRPbeta). *J. Biol. Chem.* 287, 19386–19398.

Fournier, N., Chalus, L., Durand, I., Garcia, E., Pin, J. J., Churakova, T., et al. (2000). FDF03, a novel inhibitory receptor of the immunoglobulin superfamily, is expressed by human dendritic and myeloid cells. *J. Immunol.* 165, 1197–1209.

Fujikado, N., Saijo, S., Yonezawa, T., Shimamori, K., Ishii, A., Sugai, S., et al. (2008). Dcir deficiency causes development of autoimmune diseases in mice due to excess expansion of dendritic cells. *Nat. Med.* 14, 176–180.

Gallez-Hawkins, G. M., Franck, A. E., Li, X., Thao, L., Oki, A., Gendzekhadze, K., et al. (2011). Expression of activating KIR2DS2 and KIR2DS4 genes after hematopoietic cell transplantation: relevance to cytomegalovirus infection. *Biol. Blood Marrow Transplant.* 17, 1662–1672.

Gardai, S. J., Xiao, Y. Q., Dickinson, M., Nick, J. A., Voelker, D. R., Greene, K. E., et al. (2003). By binding SIRPalpha or calreticulin/CD91, lung collectins act as dual function surveillance molecules to suppress or enhance inflammation. *Cell* 115, 13–23.

Gringhuis, S. I., Den Dunnen, J., Litjens, M., Van Der Vlist, M., and Geijtenbeek, T. B. (2009). Carbohydrate-specific signaling through the DC-SIGN signalosome tailors immunity to Mycobacterium tuberculosis, HIV-1 and Helicobacter pylori. *Nat. Immunol.* 10, 1081–1088.

Gringhuis, S. I., den Dunnen, J., Litjens, M., van Het Hof, B., van Kooyk, Y., Geijtenbeek, T. B. (2007). C-type lectin DC-SIGN modulates Toll-like receptor signaling via Raf-1 kinase-dependent acetylation of transcription factor NF-κB. *Immunity* 26, 605.

Hanke, T., Takizawa, H., Mcmahon, C. W., Busch, D. H., Pamer, E. G., Miller, J. D., et al. (1999). Direct assessment of MHC class I binding by seven Ly49 inhibitory NK cell receptors. *Immunity* 11, 67–77.

Hatherley, D., Graham, S. C., Turner, J., Harlos, K., Stuart, D. I., and Barclay, A. N. (2008). Paired receptor specificity explained by structures of signal regulatory proteins alone and complexed with CD47. *Mol. Cell* 31, 266–277.

Held, W., Cado, D., and Raulet, D. H. (1996). Transgenic expression of the Ly49A natural killer cell receptor confers class I major histocompatibility complex (MHC)-specific inhibition and prevents bone marrow allograft rejection. *J. Exp. Med.* 184, 2037–2041.

Hiby, S. E., Walker, J. J., O'shaughnessy, K. M., Redman, C. W., Carrington, M., Trowsdale, J., et al. (2004). Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. *J. Exp. Med.* 200, 957–965.

Hirayasu, K., Ohashi, J., Kashiwase, K., Hananantachai, H., Naka, I., Ogawa, A., et al. (2012). Significant association of KIR2DL3-HLA-C1 combination with cerebral malaria and implications for co-evolution of KIR and HLA. *PLoS Pathog.* 8:e1002565. doi:10.1371/journal.ppat.1002565

Hirayasu, K., Ohashi, J., Tanaka, H., Kashiwase, K., Ogawa, A., Takanashi, M., et al. (2008). Evidence for natural selection on leukocyte immunoglobulin-like receptors for HLA class I in Northeast Asians. *Am. J. Hum. Genet.* 82, 1075–1083.

Hodges, A., Sharrocks, K., Edelmann, M., Baban, D., Moris, A., Schwartz, O., et al. (2007). Activation of the lectin DC-SIGN induces an immature dendritic cell phenotype triggering Rho-GTPase activity required for HIV-1 replication. *Nat. Immunol.* 8, 569.

Iizuka, K., Naidenko, O. V., Plougastel, B. F., Fremont, D. H., and Yokoyama, W. M. (2003). Genetically linked C-type lectin-related ligands for the NKRP1 family of natural killer cell receptors. *Nat. Immunol.* 4, 801–807.

Ishikawa, E., Ishikawa, T., Morita, Y. S., Toyonaga, K., Yamada, H., Takeuchi, O., et al. (2009). Direct recognition of the mycobacterial glycolipid, trehalose dimycolate, by C-type lectin Mincle. *J. Exp. Med.* 206, 2879–2888.

Kaiser, B. K., Pizarro, J. C., Kerns, J., and Strong, R. K. (2008). Structural basis for NKG2A/CD94 recognition of HLA-E. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6696–6701.

Kamishikiryo, J., Fukuhara, H., Okabe, Y., Kuroki, K., and Maenaka, K. (2011). Molecular basis for LLT1 protein recognition by human CD161 protein (NKRP1A/KLRB1). *J. Biol. Chem.* 286, 23823–23830.

Kogure, A., Shiratori, I., Wang, J., Lanier, L. L., and Arase, H. (2011). PANP is a novel O-glycosylated PILRalpha ligand expressed in neural tissues. *Biochem. Biophys. Res. Commun.* 405, 428–433.

Lambert, A. A., Barabe, F., Gilbert, C., and Tremblay, M. J. (2011). DCIR-mediated enhancement of HIV-1 infection requires the ITIM-associated signal transduction pathway. *Blood* 117, 6589–6599.

Lambert, A. A., Gilbert, C., Richard, M., Beaulieu, A. D., and Tremblay, M. J. (2008). The C-type lectin surface receptor DCIR acts as a new attachment factor for HIV-1 in dendritic cells and contributes to trans- and cis-infection pathways. *Blood* 112, 1299–1307.

Lee, N., Llano, M., Carretero, M., Ishitani, A., Navarro, F., Lopez-Botet, M., et al. (1998). HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. *Proc. Natl. Acad. Sci. U.S.A.* 95, 5199–5204.

Li, H., Peng, S. L., Cui, Y., Fu, Q. X., Zhou, Y., Wang, Q. L., et al. (2010). Kinetics of interaction of HLA-B2705 with natural killer cell immunoglobulin-like receptor 3DS1. *Protein Pept. Lett.* 17, 547–554.

Ljunggren, H. G., and Karre, K. (1990). In search of the 'missing self': MHC molecules and NK cell recognition. *Immunol. Today* 11, 237–244.

Maenaka, K., Juji, T., Stuart, D. I., and Jones, E. Y. (1999). Crystal structure of the human p58 killer cell inhibitory receptor (KIR2DL3) specific for HLA-Cw3-related MHC class I. *Structure* 7, 391–398.

Matsumoto, M., Tanaka, T., Kaisho, T., Sanjo, H., Copeland, N. G., Gilbert, D. J., et al. (1999). A novel LPS-inducible C-type lectin is a transcriptional target of NF-IL6 in macrophages. *J. Immunol.* 163, 5039–5048.

Nakamura, A., Kobayashi, E., and Takai, T. (2004). Exacerbated graft-versus-host disease in Pirb-/- mice. *Nat. Immunol.* 5, 623–629.

Parham, P. (2005). MHC class I molecules and KIRs in human history, health and survival. *Nat. Rev. Immunol.* 5, 201–214.

Petrie, E. J., Clements, C. S., Lin, J., Sullivan, L. C., Johnson, D., Huyton, T., et al. (2008). CD94-NKG2A recognition of human leukocyte antigen (HLA)-E bound to an HLA class I leader sequence. *J. Exp. Med.* 205, 725–735.

Pozo, D., Vales-Gomez, M., Mavaddat, N., Williamson, S. C., Chisholm, S. E., and Reyburn, H. (2006). CD161 (human NKR-P1A) signaling in NK cells involves the activation of acid sphingomyelinase. *J. Immunol.* 176, 2397–2406.

Richard, M., Thibault, N., Veilleux, P., Gareau-Page, G., and Beaulieu, A. D. (2006). Granulocyte macrophage-colony stimulating factor reduces the affinity of SHP-2 for the ITIM of CLECSF6 in neutrophils: a new mechanism of action for SHP-2. *Mol. Immunol.* 43, 1716–1721.

Richard, M., Veilleux, P., Rouleau, M., Paquin, R., and Beaulieu, A. D. (2002). The expression pattern of the ITIM-bearing lectin CLECSF6 in neutrophils suggests a key role in the control of inflammation. *J. Leukoc. Biol.* 71, 871–880.

Rosen, D. B., Bettadapura, J., Alsharifi, M., Mathew, P. A., Warren, H. S., and Lanier, L. L. (2005). Cutting edge: lectin-like transcript-1 is a ligand for the inhibitory human NKR-P1A receptor. *J. Immunol.* 175, 7796–7799.

Rosen, D. B., Cao, W., Avery, D. T., Tangye, S. G., Liu, Y. J., Houchins, J. P., et al. (2008). Functional consequences of interactions between human NKR-P1A and its ligand LLT1 expressed on activated dendritic cells and B cells. *J. Immunol.* 180, 6508–6517.

Ryu, M., Chen, Y., Qi, J., Liu, J., Fan, Z., Nam, G., et al. (2011). LILRA3 binds both classical and non-classical HLA class I molecules but with reduced affinities compared to LILRB1/LILRB2: structural evidence. *PLoS ONE* 6:e19245. doi:10.1371/journal.pone.0019245

Sancho, D., Joffre, O. P., Keller, A. M., Rogers, N. C., Martinez, D., Hernanz-Falcon, P., et al. (2009). Identification of a dendritic cell receptor that couples sensing of necrosis to immunity. *Nature* 458, 899–903.

Sancho, D., and Reis e Sousa, C. (2012). Signaling by myeloid C-type lectin receptors in immunity and homeostasis. *Annu. Rev. Immunol.* 30, 491–529.

Satoh, T., Arii, J., Suenaga, T., Wang, J., Kogure, A., Uehori, J., et al. (2008). PILRalpha is a herpes simplex virus-1 entry coreceptor that associates with glycoprotein B. *Cell* 132, 935–944.

Schleypen, J. S., Von Geldern, M., Weiss, E. H., Kotzias, N., Rohrmann, K., Schendel, D. J., et al. (2003). Renal

cell carcinoma-infiltrating natural killer cells express differential repertoires of activating and inhibitory receptors and are inhibited by specific HLA class I allotypes. *Int. J. Cancer* 106, 905–912.

Schoenen, H., Bodendorfer, B., Hitchens, K., Manzanero, S., Werninghaus, K., Nimmerjahn, F., et al. (2010). Cutting edge: Mincle is essential for recognition and adjuvanticity of the mycobacterial cord factor and its synthetic analog trehalose-dibehenate. *J. Immunol.* 184, 2756–2760.

Shiroishi, M., Kuroki, K., Rasubala, L., Tsumoto, K., Kumagai, I., Kurimoto, E., et al. (2006). Structural basis for recognition of the nonclassical MHC molecule HLA-G by the leukocyte Ig-like receptor B2 (LILRB2/LIR2/ILT4/CD85d). *Proc. Natl. Acad. Sci. U.S.A.* 103, 16412–16417.

Shiroishi, M., Tsumoto, K., Amano, K., Shirakihara, Y., Colonna, M., Braud, V. M., et al. (2003). Human inhibitory receptors Ig-like transcript 2 (ILT2) and ILT4 compete with CD8 for MHC class I binding and bind preferentially to HLA-G. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8856–8861.

Sivori, S., Falco, M., Carlomagno, S., Romeo, E., Soldani, C., Bensussan, A., et al. (2010). A novel KIR-associated function: evidence that CpG DNA uptake and shuttling to early endosomes is mediated by KIR3DL2. *Blood* 116, 1637–1647.

Smith, H. R., Heusel, J. W., Mehta, I. K., Kim, S., Dorner, B. G., Naidenko, O. V., et al. (2002). Recognition of a virus-encoded ligand by a natural killer cell activation receptor. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8826–8831.

Stewart, C. A., Laugier-Anfossi, F., Vely, F., Saulquin, X., Riedmuller, J., Tisserant, A., et al. (2005). Recognition of peptide-MHC class I complexes by activating killer immunoglobulin-like receptors. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13224–13229.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., et al. (2010). Diversity of human copy number variation and multicopy genes. *Science* 330, 641–646.

Sun, Y., Senger, K., Baginski, T. K., Mazloom, A., Chinn, Y., Pantua, H., et al. (2012). Evolutionarily conserved paired immunoglobulin-like receptor alpha (PILRalpha) domain mediates its interaction with diverse sialylated ligands. *J. Biol. Chem.* 287, 15837–15850.

Tabata, S., Kuroki, K., Wang, J., Kajikawa, M., Shiratori, I., Kohda, D., et al. (2008). Biophysical characterization of O-glycosylated CD99 recognition by paired Ig-like type 2 receptors. *J. Biol. Chem.* 283, 8893–8901.

Thananchai, H., Makadzange, T., Maenaka, K., Kuroki, K., Peng, Y., Conlon, C., et al. (2009). Reciprocal recognition of an HLA-Cw4-restricted HIV-1 gp120 epitope by CD8+ T cells and NK cells. *AIDS* 23, 189–193.

Tomasec, P., Braud, V. M., Rickards, C., Powell, M. B., Mcsharry, B. P., Gadola, S., et al. (2000). Surface expression of HLA-E, an inhibitor of natural killer cells, enhanced by human cytomegalovirus gpUL40. *Science* 287, 1031.

Ulbrecht, M., Martinozzi, S., Grzeschik, M., Hengel, H., Ellwart, J. W., Pla, M., et al. (2000). Cutting edge: the human cytomegalovirus UL40 gene product contains a ligand for HLA-E and prevents NK cell-mediated lysis. *J. Immunol.* 164, 5019–5022.

Valiante, N. M., Uhrberg, M., Shilling, H. G., Lienert-Weidenbach, K., Arnett, K. L., D'andrea, A., et al. (1997). Functionally and structurally distinct NK cell receptor repertoires in the peripheral blood of two human donors. *Immunity* 7, 739–751.

Vivian, J. P., Duncan, R. C., Berry, R., O'connor, G. M., Reid, H. H., Beddoe, T., et al. (2011). Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B. *Nature* 479, 401–405.

Wang, J., Shiratori, I., Satoh, T., Lanier, L. L., and Arase, H. (2008). An essential role of sialylated O-linked sugar chains in the recognition of mouse CD99 by paired Ig-like type 2 receptor (PILR). *J. Immunol.* 180, 1686–1693.

Willcox, B. E., Thomas, L. M., and Bjorkman, P. J. (2003). Crystal structure of HLA-A2 bound to LIR-1, a host and viral major histocompatibility complex receptor. *Nat. Immunol.* 4, 913–919.

Wright, G. J., Cherwinski, H., Foster-Cuevas, M., Brooke, G., Puklavec, M. J., Bigler, M., et al. (2003). Characterization of the CD200 receptor family in mice and humans and their interactions with CD200. *J. Immunol.* 171, 3034–3046.

Yamasaki, S., Ishikawa, E., Sakuma, M., Hara, H., Ogata, K., and Saito, T. (2008). Mincle is an ITAM-coupled activating receptor that senses damaged cells. *Nat. Immunol.* 9, 1179–1188.

Yamasaki, S., Matsumoto, M., Takeuchi, O., Matsuzawa, T., Ishikawa, E., Sakuma, M., et al. (2009). C-type lectin Mincle is an activating receptor for pathogenic fungus, Malassezia. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1897–1902.

Yang, Z., and Bjorkman, P. J. (2008). Structure of UL18, a peptide-binding viral MHC mimic, bound to a host inhibitory receptor. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10095–10100.

Yu, Y. Y., George, T., Dorfman, J. R., Roland, J., Kumar, V., and Bennett, M. (1996). The role of Ly49A and 5E6 (Ly49C) molecules in hybrid resistance mediated by murine natural killer cells against normal T cell blasts. *Immunity* 4, 67–76.

Zelensky, A. N., and Gready, J. E. (2005). The C-type lectin-like domain superfamily. *FEBS J.* 272, 6179–6217.

Zhang, J. G., Czabotar, P. E., Policheni, A. N., Caminschi, I., Wan, S. S., Kitsoulis, S., et al. (2012). The dendritic cell receptor Clec9A binds damaged cells via exposed actin filaments. *Immunity* 36, 646–657.

Zheng, J., Umikawa, M., Cui, C., Li, J., Chen, X., Zhang, C., et al. (2012). Inhibitory receptors bind ANGPTLs and support blood stem cells and leukaemia development. *Nature* 485, 656–660.

Conflict of Interest Statement: 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.

Received: 15 August 2012; paper pending published: 04 September 2012; accepted: 06 December 2012; published online: 31 December 2012.

*Citation: Kuroki K, Furukawa A and Maenaka K (2012) Molecular recognition of paired receptors in the immune system. Front. Microbio. **3**:429. doi: 10.3389/fmicb.2012.00429*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 31 July 2012
doi: 10.3389/fmicb.2012.00278

Phylodynamic analysis of a viral infection network

Teiichiro Shiino*

Infectious Diseases Surveillance Center, National Institute of Infectious Diseases, Tokyo, Japan

Edited by:
Hironori Sato, National Institute of Infectious Diseases, Japan

Reviewed by:
Hirotaka Ode, National Hospital Organization Nagoya Medical Center, Japan
Kazushi Motomura, National Institute of Infectious Diseases, Japan

***Correspondence:**
Teiichiro Shiino, Infectious Diseases Surveillance Center, National Institute of Infectious Diseases, Toyama 1-23-1, Shinjuku-ku, Tokyo, Japan.
e-mail: tshiino@nih.go.jp

Viral infections by sexual and droplet transmission routes typically spread through a complex host-to-host contact network. Clarifying the transmission network and epidemiological parameters affecting the variations and dynamics of a specific pathogen is a major issue in the control of infectious diseases. However, conventional methods such as interview and/or classical phylogenetic analysis of viral gene sequences have inherent limitations and often fail to detect infectious clusters and transmission connections. Recent improvements in computational environments now permit the analysis of large datasets. In addition, novel analytical methods have been developed that serve to infer the evolutionary dynamics of virus genetic diversity using sample date information and sequence data. This type of framework, termed "phylodynamics," helps connect some of the missing links on viral transmission networks, which are often hard to detect by conventional methods of epidemiology. With sufficient number of sequences available, one can use this new inference method to estimate theoretical epidemiological parameters such as temporal distributions of the primary infection, fluctuation of the pathogen population size, basic reproductive number, and the mean time span of disease infectiousness. Transmission networks estimated by this framework often have the properties of a scale-free network, which are characteristic of infectious and social communication processes. Network analysis based on phylodynamics has alluded to various suggestions concerning the infection dynamics associated with a given community and/or risk behavior. In this review, I will summarize the current methods available for identifying the transmission network using phylogeny, and present an argument on the possibilities of applying the scale-free properties to these existing frameworks.

Keywords: phylodynamics, transmission network, phylogenetic inference, infection dynamics, scale-free network

In their natural habitat, various pathogen groups exist in genetically and environmentally diverse human populations. Amongst the many infectious agents, droplet- or sexually transmitted viruses spread most rapidly through complex human networks. To control these types of viral diseases, we have to learn about the behavior of pathogens in relation to their host populations. Factors that influence and determine the incidence and distribution of infectious diseases have been investigated extensively in the field of epidemiology.

LIMITATIONS OF THE CLASSICAL METHODS FOR ESTIMATING TRANSMISSION NETWORKS

Conventionally, epidemiological researchers have generally derived interpretations of the contact network by using interview or other measures available in the clinic. However, these methods have inherent issues in cases where the virus causes long-term chronic infections or short-term rapid transmissions. Diagnoses of infection cases of chronic viruses in early phase are usually made in only a small population of individuals (Pao et al., 2005; Pilcher et al., 2005). Clinical surveillance can also acquire a small number of patients as compared with a whole population of person who is infected in virus with very rapid spreading, such as pandemic influenza. Therefore, the clinic-based analyses of viral epidemiology have been restricted to low-density samples and this factor may cause a bias in the results towards under-reporting of infection networks or "clusters" (Brown et al., 1997; Lewis et al., 2008). One solution to detect these infection clusters in virus transmission networks is a phylogenetic analysis of population-based samples of viral genetic sequences. A number of studies have identified the clusters by elucidating the evolutionary relationship of human immunodeficiency virus (HIV) (Salminen et al., 1993; Brown et al., 1997; Yirrell et al., 1997), hepatitis C virus (HCV) (Aitken et al., 2004) and the influenza virus (Nelson et al., 2007; Nelson and Holmes, 2007). Nevertheless, in the phylogeny that exhibits a star-like divergence pattern, the analysis using sequences from population survey can only provide limited evidence for the cluster. Sexually-transmitted populations of HIV-1 are representative of diseases showing this type of divergence patterns. For example, a population-based phylogenetic analysis of HIV-1 in the UK (Brown et al., 1997) could only identify a limited number of clusters even though the primary infected individuals were recruited from a cohort. A similar study conducted in Quebec (Brenner et al., 2007) showed to a certain extent evidence for the cluster, however most of the findings were gathered from intravenous drug users or from the

sexual transmission patterns of a men who has sex with men (MSM). These major risk factors are still relatively unexplored and therefore poorly elucidated. These challenges are mostly due to the fact that most of the patients recruited were in the phase of chronic infection, despite the fact that in both of these research studies the recently (<6 month) seroconverted individuals were recruited from cohorts. Because a diagnosis of acute infection is usually made in only a small proportion of individuals with HIV-1 (Pilcher et al., 2005), the samples will inevitably contain viruses at chronic infection phase. Moreover, multiple introductions into target populations (Korber et al., 2000) will result in higher diversity amongst each virus group. In these situations, simple phylogenetic analyses that only employ sequences of virus at the chronic infection phase will generate inaccurate outputs due to computational bias, and skew the results thus underestimating the number of clusters (Brown et al., 1997). To resolve this supposed bias, it is necessary to obtain a larger number of sequences for analysis and/or use more sophisticated methods to infer evolutionary relationships from the sequences.

Contact tracing by interview data, which plays a key role in establishing the etiology of some infectious diseases (Klovdahl, 1985), may be difficult to carry out at the sites of epidemics. Transmission networks reconstructed by phylogenetic analysis have been considered as the standard host contact network in many studies. However, results gained by using both of the current methods have often been inconsistent due to long-term infections, a low average risk of transmission, and a relatively high rate of exposure to the virus (Wawer et al., 2005). The contact tracing method cannot effectively identify the specific instances of contact detected in the interviews associated with the infection, while the conventional sequence-based analysis cannot sufficiently confirm the results to provide quantitative descriptions concerning the transmission networks due to the above mentioned reasons.

These difficulties could potentially be overcome by acquiring an adequate number of viral sequences and also by an improved method for estimating the divergence time for each phylogenetic node. A number of recent advances in clinical and computational science have introduced the possibility of developing a novel more efficacious approach. Rapid developments in DNA sequencing technology have catalyzed the advent of medical diagnostics using viral genome sequences. In particular, genotype-based resistance tests are commonplace in antiviral therapeutics for patients infected with HIV, HBV or HCV. Progress made in computational technologies has also facilitated large-scale sequence analysis of clinical diagnostic data. With advancements in evolutionary biology, the evolutionary dynamics of a population can now be inferred from sequence data and incorporated with sampling dates. Such evolutionary dynamics information of a pathogen derived from these analyses can then be combined with epidemiological data to illustrate the influences of host transmission dynamics, immunity, and treatments against specific genetic variations of pathogen. Such a series of analyses is now referred to as "phylodynamics" (Grenfell et al., 2004; Holmes and Grenfell, 2009). Phylodynamic frameworks require sufficient sequence diversities of the sample dataset with respect to spatial as well as temporal variations. Thus, RNA viruses, which have high substitution rate, high growth rate, and a short generation time, are especially advantageous and amenable to investigation (Grenfell et al., 2004; Kühnert et al., 2011).

SEEKING TRANSMISSION NETWORKS USING PHYLODYNAMICS

Identifying transmission networks is one major issue in the epidemiological analysis of infectious diseases. In research performed on HIV sexual transmission networks, drug resistance tests accompanied by HARRT have helped to provide a sufficient number of sequences. Even in other viral diseases, the sequence datasets accepted in the framework may be available under an arrangement of surveillance system with the sequence database. The divergence time, another piece of the framework, has been estimated from the time of the most recent common ancestors (tMRCAs) for each node of phylogeny, and dating of phylogenetic tree, including tMRCA estimation, is one of the recent achievements of modern theoretical biology (Drummond et al., 2005, 2006). Currently, phylogenetic inference using Bayesian coalescent Markov Monte Carlo (Bayesian MCMC) method (Pybus et al., 2003; Drummond and Rambaut, 2007) is commonly performed in this step. It is well-known that due to their high mutation and proliferation rates, the sequence evolution of an RNA virus occurs on a time scale when any public health measures are being conducted, suggesting that a time-based phylogenetic inference, which usually requires samples taken in extremely different ages such as fossil-derived PCR sequence, is easily applicable to RNA viral sequences (Pybus et al., 2003). An excessive number of sequences would increases the probability of acquisition of the transmission networks in a phylogeny, and a relative date of infection estimated by each divergence time in a phylogeny provides the missing links for contact cases.

Quantitative description of a transmission network using phylodynamics represents a temporal and spatial profile for certain viral epidemics. For example, intra-national epidemiological study of pandemic influenza A(H1N1) in 2009 was analyzed by this framework (Shiino et al., 2010). In this study, an endemic transmission network detected in the phylogeny was regarded as a single transport case of A(H1N1) virus in Japan. The local spreading profile of 12 cases was illustrated by the date (when the case was introduced) and locality (how the virus was spread) of the infection (**Figure 1**). This type of approach is performed in a more detailed manner on sexual-HIV epidemics in the UK using partial *pol* regions from over 10,000 patients (Lewis et al., 2008; Hughes et al., 2009). A large majority of MSM persons were linked to more than one other individual and 25% engendered large transmission networks (Lewis et al., 2008), whereas only 5% of individuals generated large clusters in heterosexual transmissions (Hughes et al., 2009). On the issue of transmission intervals estimated from tMRCA, the median interval for by MSM [i.e., 13 months; (Lewis et al., 2008)] was less than half of that estimated for heterosexual individuals [i.e., 27 months; (Hughes et al., 2009)]. Thus, the phylodynamics revealed an aspect of a viral spreading pattern that was associated with a given community or risk behavior that had not been previously elucidated.

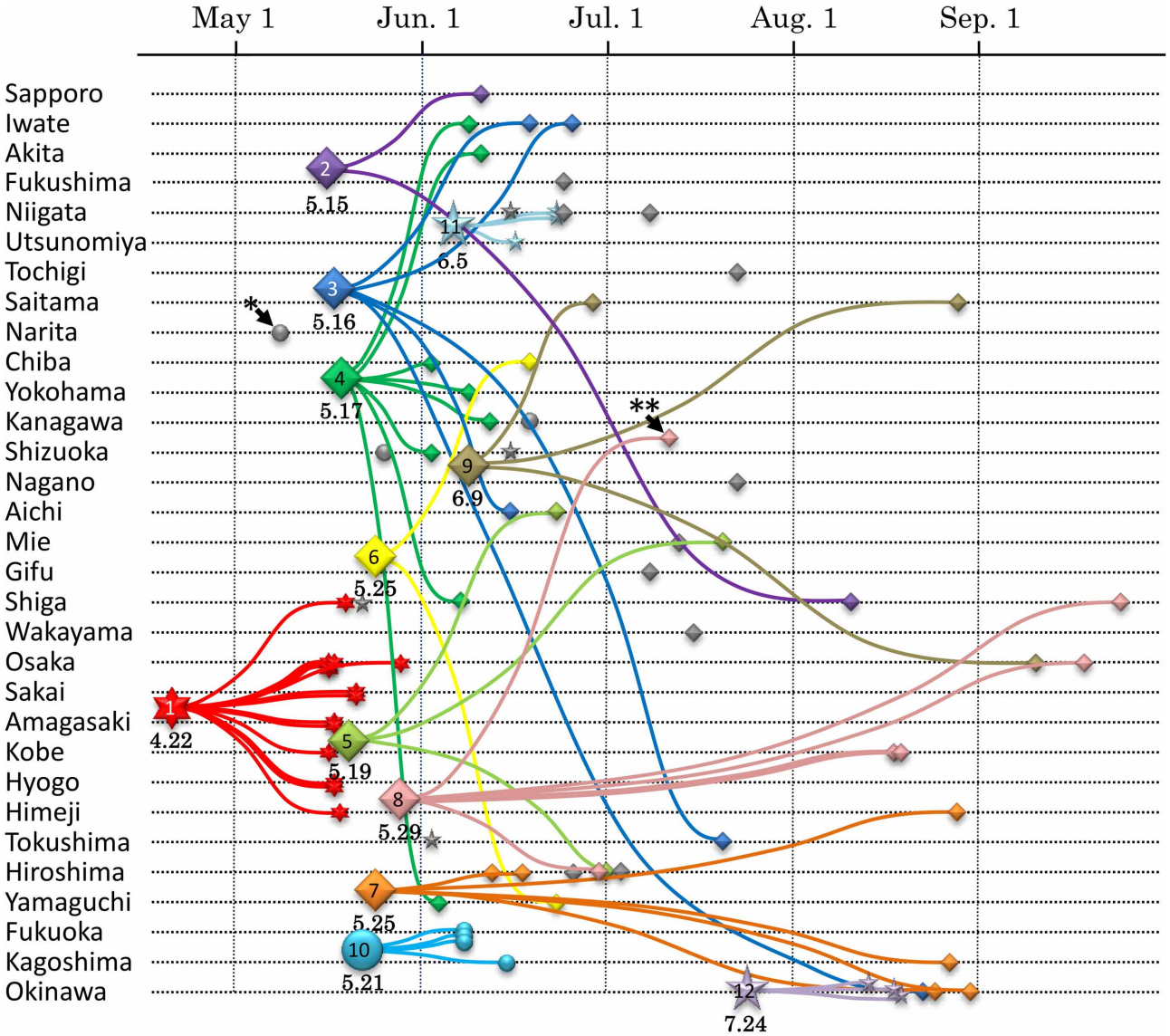


FIGURE 1 | Transmission dynamics of A(H1N1)pdm epidemic in Japan estimated by phylodynamics. The vertical and horizontal axes show geographic localities and the times of virus collection, respectively. Chronological phylogeny was inferred by BEAST v1.5.4 using the general time reversible model, taking into account site heterogeneity and invariant sites (GTR+I+G) and the logistic population model (Drummond and Rambaut, 2007). Seventy-five isolates of A(H1N1)pdm viruses from Japan are plotted with the small symbols. The MRCA of each transmission network inferred from the phylodynamics are plotted with the large symbols. The numbers below the large symbols display the tMRCA date. This figure is cited from Figure 4 in our previous report (Shiino et al., 2010).

ESTIMATING EPIDEMIOLOGICAL PARAMETERS FROM SEQUENCE DIVERSITY

Another area of interest for epidemiologists is the estimation of parameters affecting the dynamics of a particular pathogen. Predicting the population dynamics of a pathogen requires precious quantification of the key parameters in the population model. Although generally, it has been derived from enumeration of data on the disease incidence, parameters inferred by phylodynamics are also applicable in describing complex population dynamics of an RNA virus. When a sequence variation in a gene population primarily depends on the neutral mutation, the expected value of mean nucleotide diversity is proportional to $N_e\upsilon$, where N_e indicates an effective population size and υ signifies the total mutation rate per loci (Kimura, 1969). Since υ is uniform for long time-scales (although the generation time of the transmitted virus may fluctuate), the observed nucleotide diversity is thought to be a relative size of the viral population (Drummond et al., 2005). The coalescent tree analysis in a phylodynamic framework allows measurement of the nucleotide diversity for each time unit from a time-slice of the chronological phylogeny. The consequence of this slicing is the Bayesian skyline plot (BSP), which represents a piecewise graphical demonstration

of population dynamics of the virus (Drummond et al., 2005). Note that the BSP does not demonstrate population dynamics of the host individual but rather for the pathogen itself, although both are consistent in the case of a fixed number of transmitted viruses to the host. As shown in **Figure 2**, the BSP can illustrate a feature of the epidemic along with a temporal component (Rambaut et al., 2008). Moreover, the BSP is useful in analyzing the intra-host virus struggle against an immune response (Bernini et al., 2011). Additionally, better precise estimates of the parameters are now capable of improved assessment using the phylodynamic framework. The population growth rate (r) can be estimated by the Bayesian MCMC inference (Drummond and Rambaut, 2007) as well as the maximum likelihood phylogeny with branch length correction for the sampling date (Pybus and Rambaut, 2002). The mean time of infectiousness (D) is dependent on a function of distribution of the generation time periods that elapsed between transmission processes [$w(t)$] (Grassly and Fraser, 2008), although this can also be inferred from the phylogenies, it is difficult to determine one estimates due to the various properties of viral infections (i.e., fluctuating viral load, and the wide range of transmission probability with respect to risk behavior) (Sherlock, 1993; Chevaliez and Pawlotsky, 2007; Romano et al., 2010). When we have obtained r and D, we can approximately estimate the basic reproductive number (R_0) using the relation $R_0 = 1 + rD$ (Pybus et al., 2001). Estimating r and R_0 in HCV-infected individuals revealed that subtype 1b, which is found chiefly amongst elderly individuals with a history of blood transfusions, spreads slower than the compared to other subtypes (Romano et al., 2010).

RANDOM GRAPH ANALYSIS OF VIRAL TRANSMISSION DYNAMICS AND PHYLODYNAMICS

Since Watts and Strogatz (1998), Barabási and Albert (1999), introduced the "small-world" and "scale-free" network models (**Figure 3**) to random graph theorem, respectively, it have been elucidated that the scale-free and small-world properties are observed in many human-intervened communicative networks such as the internet and infectious diseases. Pastor-Satorras and Vespignani (2001) demonstrated that computer viruses rampant on the internet spread through a scale-free network, which can drive the viruses to spread even when infection probabilities are negligibly small. This prediction can be applied to not only computer systems but also human pathogenic viruses. Scale-free properties are observed in transmission networks reconstructed by the phylodynamic framework, e.g., connectivities of the individuals in the network often follow a power law (**Figure 4**). Such features have been found in the phylodynamic network of HIV both in MSM (Lewis et al., 2008) and heterosexual populations

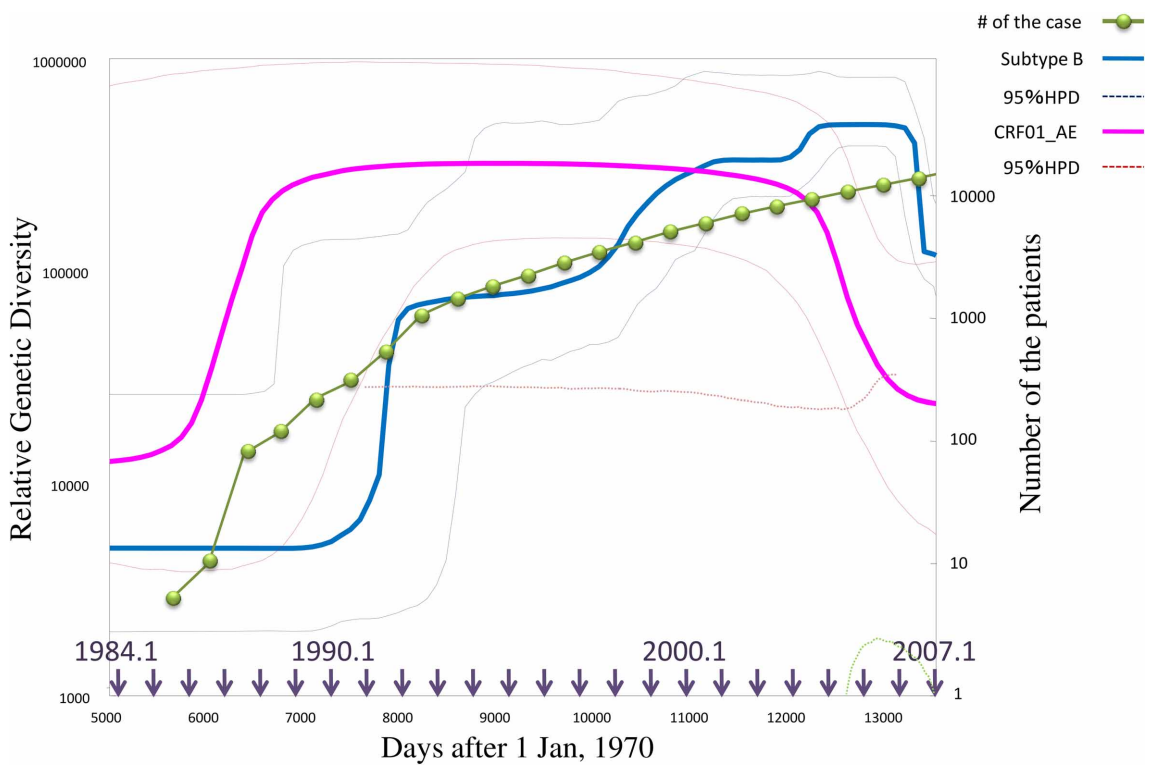


FIGURE 2 | Population dynamics inferred by Bayesian skyline plot analysis of HIV-1 epidemics in Japan. This graph represents the median estimate of the genetic diversities, i.e., relative population sizes, of HIV-1 Subtype B (blue line) and CRF01_AE (pink line) in Japan, with 95% posterior intervals (dotted line) estimated by BSP. The green line shows the number of HIV-1 infection cases in Japan. The horizontal axis indicates days after the first day of 1970 with the arrows showing approximate date. BSP is performed using BEAST v.1.5.4 with GTR+I+G (Drummond and Rambaut, 2007), 50 million states of MCMC runs, and burn-in of 4 million states. The sequence dataset was provided by the research program of Japanese Drug Resistance HIV-1 Surveillance Network Group (Gatanaga et al., 2007; Hattori et al., 2010) making use of over 2000 protease-RT sequences from Japan.

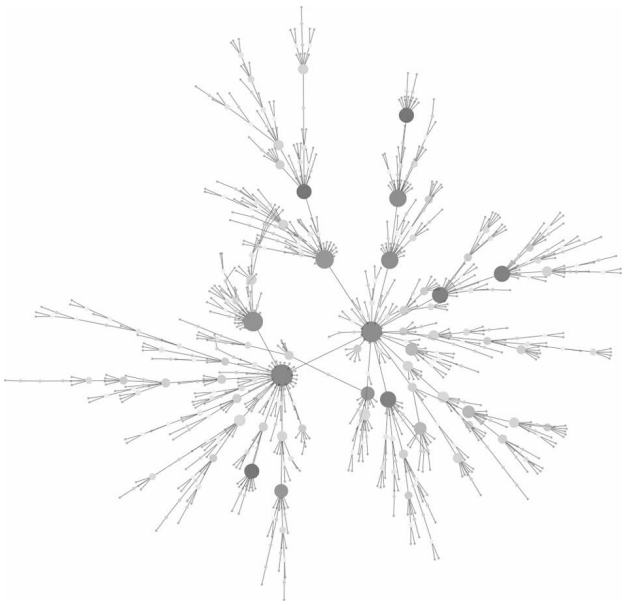


FIGURE 3 | Example of a scale-free network. The graph consisting of 1000 nodes was generated with the Barabási and Albert model (Barabási and Albert, 1999). Large and deeper-colored nodes show higher connectivity. Note the majority of nodes have few connections. The graph was generated by igraph 0.6 using ba-model and visualized in Cytoscape 2.5.

(Hughes et al., 2009) and in HCV groups (Romano et al., 2010). Moreover, the transmission network for the HIV epidemic among MSM in the UK was recently reanalyzed using fine distribution model with the preferential attachment process. Observed distribution specifically fitted to the Waring distribution at all time-depths of the phylodynamic clusters (Leigh Brown et al., 2011). These findings give a significant message for preparation of a public health measures; Lloyd et al. stated in their perspective in Science (Lloyd and May, 2001) that "the study highlights the potential importance of studies on communication and other network, especially those with scale-free and small world properties, for those seeking to manage epidemics within human and other animal population." Under the scale-free and small-world condition, an infection will spread regardless of its transmissibility and a control program targeted at highly connectable individuals (i.e., super-spreaders) is important to curb the epidemic (Keeling and Eames, 2005). A phylodynamic framework would help to decide a target population for the treatment and allow a decrease in the cost of treatment.

IDENTIFICATION OF THE TRANSMISSION NETWORK USING THE SCALE-FREE FEATURE

From previous discussions, it is evident that phylodynamics can engender important knowledge about the epidemiology of infectious diseases. The problem is that no accurate and consistent process for deciding the transmission network on the phylogeny is present in this framework. Before conducting Bayesian MCMC

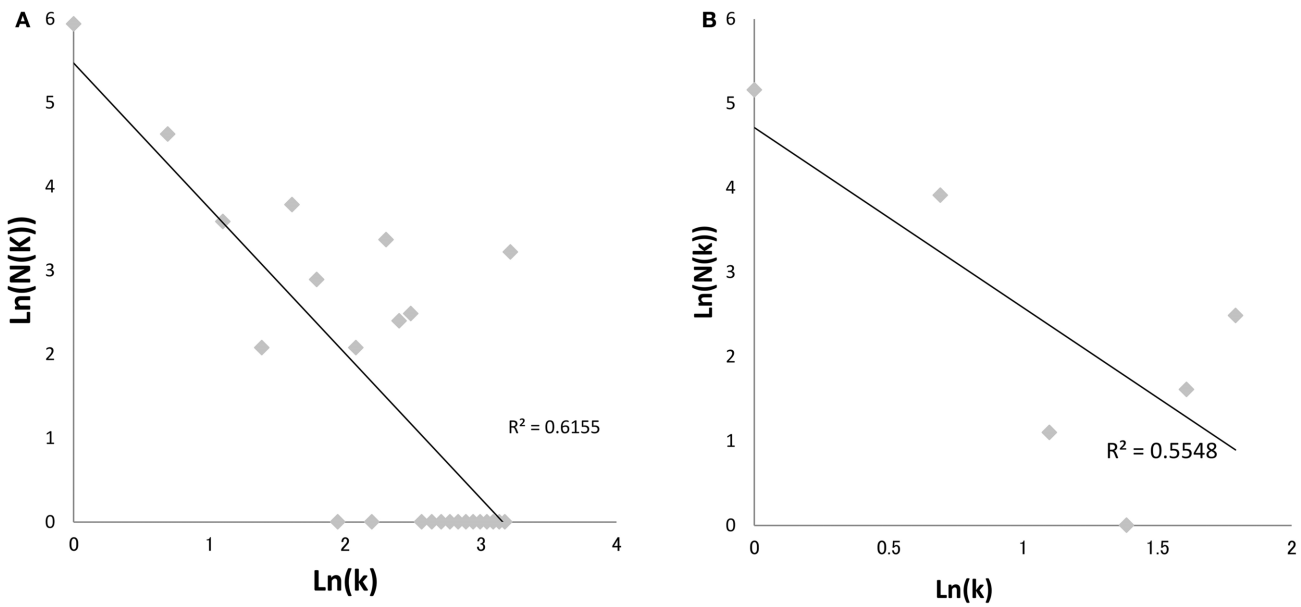


FIGURE 4 | Power-law plot of HIV-1 transmission network in Japan. The graphs depict log-log plots of the numbers of patients with k contacts [$N(k)$] against the number of contacts (k) in HIV-1 clusters of Japan subtype B **(A)** and CRF01_AE **(B)**. Regression line was calculated using the least-square method. The adjusted coefficients of determination (R^2) were 0.5548 ($p = 0.04$) and 0.6155 ($p < 0.01$), respectively. The Neighbor-Joining trees were reconstructed using MEGA version 5.0.5 (Tamura et al., 2011), and >80% bootstrap probabilities adhering to the criteria in a previous report (Shiino et al., 2010) were adopted in the significant clades. The statistic analyses were calculated using R version 2.10.0. The sequence dataset was provided by the research program of Japanese Drug Resistance HIV-1 Surveillance Network Group (Gatanaga et al., 2007; Hattori et al., 2010) using over 2000 protease-RT sequences from Japan.

analysis, which is a core process in phylodynamics, one has to identify transmission networks from monophyletic groups (clade) inferred by the conventional phylogenetic tree analysis. It is difficult to evaluate the credibility of the clade in the pathogen phylogeny, especially in star-like trees observed in chronic (e.g., HIV-1) and/or pandemic (e.g., novel influenza) infections. While it is common to use the bootstrap probability method for verifying the reliability of the clade, this is a dubious index in the case of viral sequence analyses since it may be different from the probability distribution of error when it is applied to the viral sequences that widely fluctuate their base substitution rate along with the site or host environment. Although the posterior probability of nodes calculated in the Bayesian tree inference was sometimes used as phylogenetic support for each clade (Lewis et al., 2008), this method has a computational issue, as the MCMC search with large datasets requires a huge resource of computational power. In the case of imported infectious disease, robustness of the cladding is also examined by supplying a large number of closely related reference sequences (Hughes et al., 2009; Shiino et al., 2010). However, this type of method depends on whether these reference sequences are freely available or otherwise. Consequently, here I wish to propose a method considered to be more effective in determining the cutoff value of bootstrap probability for transmission network reconstruction. As mentioned above, assuming that nearly all viruses transmitted along with the scale-free network, degree distribution of the number of members in valid networks estimated from the clades in a phylogeny should follow the power law. Therefore, if the relationship between the results of each bootstrap value of the observed phylogenetic cluster and fitting of the degree distribution to power law plot is investigated, a bootstrap probability for selecting the network to be adopted may be clear. **Figure 5** showed the relationship between bootstrap probability of the neighbor-joining tree and the coefficient of determination (R^2) in linear regression of log-log plots of the member distribution of the significant clades, using 1882 sequences of the pol domain of HIV-1 subtype B in Japan collected by the Japanese Drug Resistance HIV-1 Surveillance Network Group (Hattori et al., 2010). The R^2 with regard to the power law fitting shows highest values at bootstrap probabilities between 76 and 82%, and decreased in both higher and lower the probabilities. On the other hand, the number of infected persons included in the significant clades decreased consistently as the bootstrap probability increased. This result suggests that the optimal bootstrap probabilities for the scale-free property, which is approximately 80% in this case, is present in a viral sequence data.

CONCLUSION

In order to manage and control infectious diseases, it is important for epidemiologists to monitor and comprehensively analyze

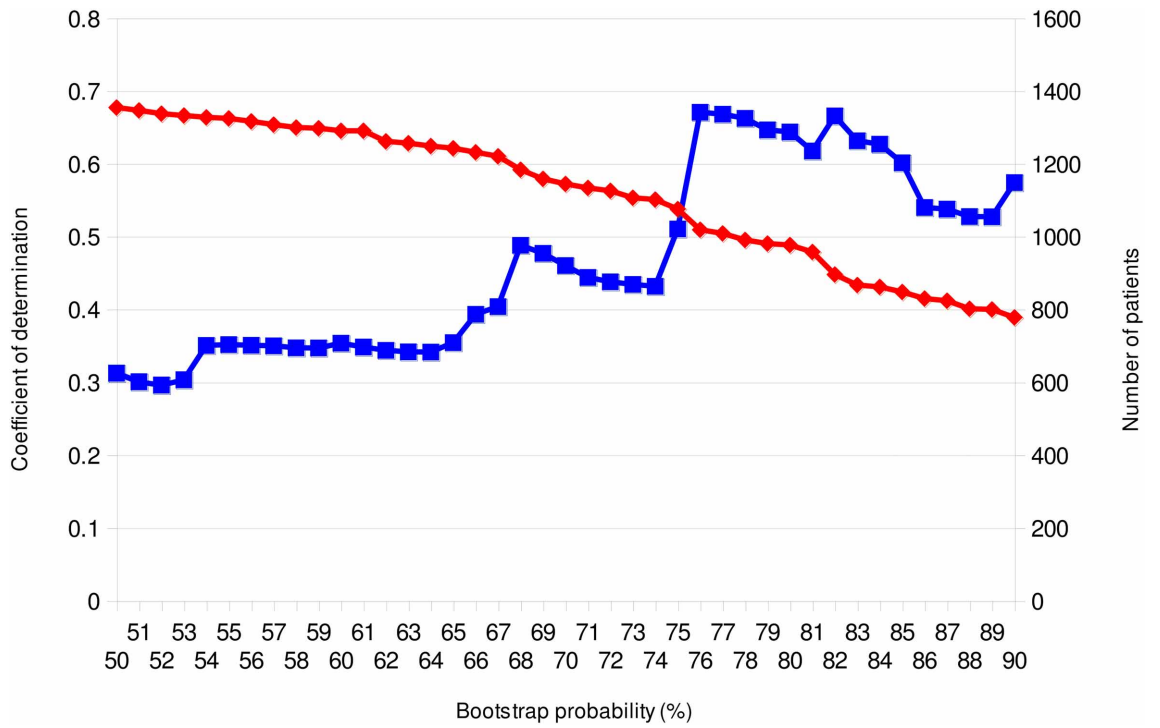


FIGURE 5 | Bootstrap-scanning on power-law fitting of clusters in the Neighbor-Joining tree. The adjusted coefficients of linear regression determination of log-log plots and the number of individuals in the clusters are plotted on the bootstrap cut-off probability for the clade distinction. The blue line with square symbols and the red line with lozenge symbols show the coefficients of determination and the numbers of individuals included among the clusters, respectively. The phylogenies were reconstructed using MEGA version 5.0.5 (Tamura et al., 2011). MEGA output was analyzed using a combination of PERL and R scripts. This analysis was performed using 1882 sequences of the protease-RT regions of HIV-1 Subtype B in Japan; the data was provided by the research program of Japanese Drug Resistance HIV-1 Surveillance Network Group (Gatanaga et al., 2007; Hattori et al., 2010).

representative epidemiological indices. The phylodynamic framework proposed here can serve as a powerful tool for handling such data by integrating related areas of epidemiology such as population dynamics, genetics and molecular evolutionary research. Additionally, this framework is relevant with respect to aspects of pathogen evolution against immunological responses, drug administration and pathological development. Remaining challenges pivotal in the eventual success of the framework include necessary improvements in data sampling for both disease incidence and pathogenic sequences. At present, these types of data are only collected periodically. In addition, public sequence databases such as GenBank should begin to stringently record the collection date, location, and the accompanying clinical information. Achieving these objectives will allow the phylodynamic framework to future contribute to successful disease control.

ACKNOWLEDGMENTS

I am very grateful to each member of the Japanese Drug Resistance HIV-1 Surveillance Network Group, especially to Dr. Wataru Sugiura and Dr. Junko Hattori for providing of data. I am also grateful to Dr. Hironori Sato for helpful discussions.

REFERENCES

Aitken, C. K., McCaw, R. F., Bowden, D. S., Tracy, S. L., Kelsall, J. G., Higgs, P. G., Kerger, M. J., Nguyen, H., and Crofts, J. N. (2004). Molecular epidemiology of hepatitis C virus in a social network of injection drug users. *J. Infect. Dis.* 190, 1586–1595.

Barabási, A. L., and Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509–512.

Bernini, F., Ebranati, E., De Maddalena, C., Shkjezi, R., Milazzo, L., Lo Presti, A., Ciccozzi, M., Galli, M., and Zehender, G. (2011). Within-host dynamics of the hepatitis C virus quasispecies population in HIV-1/HCV coinfected patients. *PLoS ONE* 6:e16551. doi: 10.1371/journal.pone.0016551

Brenner, B. G., Roger, M., Routy, J. P., Moisi, D., Ntemgwa, M., Matte, C., Baril, J. G., Thomas, R., Rouleau, D., Bruneau, J., Leblanc, R., Legault, M., Tremblay, C., Charest, H., and Wainberg, M. A. (2007). High rates of forward transmission events after acute/early HIV-1 infection. *J. Infect. Dis.* 195, 951–959.

Brown, A. J., Lobidel, D., Wade, C. M., Rebus, S., Phillips, A. N., Brettle, R. P., France, A. J., Leen, C. S., McMenamin, J., McMillan, A., Maw, R. D., Mulcahy, F., Robertson, J. R., Sankar, K. N., Scott, G., Wyld, R., and Peutherer, J. F. (1997). The molecular epidemiology of human immunodeficiency virus type 1 in six cities in Britain and Ireland. *Virology* 235, 166–177.

Chevaliez, S., and Pawlotsky, J. M. (2007). Hepatitis C virus: virology, diagnosis and management of antiviral therapy. *World J. Gastroenterol.* 13, 2461–2466.

Drummond, A. J., Ho, S. Y., 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., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214.

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.

Gatanaga, H., Ibe, S., Matsuda, M., Yoshida, S., Asagi, T., Kondo, M., Sadamasu, K., Tsukada, H., Masakane, A., Mori, H., Takata, N., Minami, R., Tateyama, M., Koike, T., Itoh, T., Imai, M., Nagashima, M., Gejyo, F., Ueda, M., Hamaguchi, M., Kojima, Y., Shirasaka, T., Kimura, A., Yamamoto, M., Fujita, J., Oka, S., and Sugiura, W. (2007). Drug-resistant HIV-1 prevalence in patients newly diagnosed with HIV/AIDS in Japan. *Antiviral Res.* 75, 75–82.

Grassly, N. C., and Fraser, C. (2008). Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* 6, 477–487.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 327–332.

Hattori, J., Shiino, T., Gatanaga, H., Yoshida, S., Watanabe, D., Minami, R., Sadamasu, K., Kondo, M., Mori, H., Ueda, M., Tateyama, M., Ueda, A., Kato, S., Ito, T., Oie, M., Takata, N., Hayashida, T., Nagashima, M., Matsuda, M., Ibe, S., Ota, Y., Sasaki, S., Ishigatsubo, Y., Tanabe, Y., Koga, I., Kojima, Y., Yamamoto, M., Fujita, J., Yokomaku, Y., Koike, T., Shirasaka, T., Oka, S., and Sugiura, W. (2010). Trends in transmitted drug-resistant HIV-1 and demographic characteristics of newly diagnosed patients: nationwide surveillance from 2003 to 2008 in Japan. *Antiviral Res.* 88, 72–79.

Holmes, E. C., and Grenfell, B. T. (2009). Discovering the phylodynamics of RNA viruses. *PLoS Comput. Biol.* 5:e1000505. doi: 10.1371/journal.pcbi.1000505

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

Keeling, M. J., and Eames, K. T. (2005). Networks and epidemic models. *J. R. Soc. Interface* 2, 295–307.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 893–903.

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. *Soc. Sci. Med.* 21, 1203–1216.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science* 288, 1789–1796.

Kühnert, D., Wu, C. H., and Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infect. Genet. Evol.* 11, 1825–1841.

Leigh Brown, A. J., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., Dunn, D. T.; on behalf of the UK HIV Drug Resistance Collaboration. (2011). Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* 204, 1463–1469.

Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A., and Leigh Brown, A. J. (2008). Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 5:e50. doi: 10.1371/journal.pmed.0050050

Lloyd, A. L., and May, R. M. (2001). Epidemiology. How viruses spread among computers and people. *Science* 292, 1316–1317.

Nelson, M. I., and Holmes, E. C. (2007). The evolution of epidemic influenza. *Nat. Rev. Genet.* 8, 196–205.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., and Holmes, E. C. (2007). Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog.* 3:1220–1228. doi: 10.1371/journal.ppat.0030131

Pao, D., Fisher, M., Hué, S., Dean, G., Murphy, G., Cane, P. A., Sabin, C. A., and Pillay, D. (2005). Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS* 19, 85–90.

Pastor-Satorras, R., and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86, 3200–3203.

Pilcher, C. D., Fiscus, S. A., Nguyen, T. Q., Foust, E., Wolf, L., Williams, D., Ashby, R., O'Dowd, J. O., McPherson, J. T., Stalzer, B., Hightow, L., Miller, W. C., Eron, J. J., Cohen, M. S., and Leone, P. A. (2005). Detection of acute infections during HIV testing in North Carolina. *N. Engl. J. Med.* 352, 1873–1883.

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.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H., and Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol.* 20, 381–387.

Pybus, O. G., and Rambaut, A. (2002). GENIE: estimating demographic history from molecular phylogenies. *Bioinformatics* 18, 1404–1405.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological

dynamics of human influenza A virus. *Nature* 453, 615–619.

Romano, C. M., de Carvalho-Mello, I. M., Jamal, L. F., de Melo, F. L., Iamarino, A., Motoki, M., Pinho, J. R., Holmes, E. C., and de Andrade Zanotto, P. M. (2010). Social networks shape the transmission dynamics of hepatitis C virus. *PLoS ONE* 5:e11170. doi: 10.1371/journal.pone.0011170

Salminen, M., Nykänen, A., Brummer-Korvenkontio, H., Kantanen, M. L., Liitsola, K., and Leinikki, P. (1993). Molecular epidemiology of HIV-1 based on phylogenetic analysis of *in vivo* gag p7/p9 direct sequences. *Virology* 195, 185–194.

Sherlock, S. (1993). European livers. *Lancet* 342, 1127–1128.

Shiino, T., Okabe, N., Yasui, Y., Sunagawa, T., Ujike, M., Obuchi, M., Kishida, N., Xu, H., Takashita, E., Anraku, A., Ito, R., Doi, T., Ejima, M., Sugawara, H., Horikawa, H., Yamazaki, S., Kato, Y., Oguchi, A., Fujita, N., Odagiri, T., Tashiro, M., and Watanabe, H. (2010). Molecular evolutionary analysis of the influenza A(H1N1)pdm, May-September 2009, temporal and spatial spreading profile of the viruses in Japan. *PLoS ONE* 5:e11057. doi: 10.1371/journal.pone.0011057

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5, molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.

Wawer, M. J., Gray, R. H., Sewankambo, N. K., Serwadda, D., Li, X., Laeyendecker, O., Kiwanuka, N., Kigozi, G., Kiddugavu, M., Lutalo, T., Nalugoda, F., Wabwire-Mangen, F., Meehan, M. P., and Quinn, T. C. (2005). Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J. Infect. Dis.* 191, 1403–1409.

Yirrell, D. L., Robertson, P., Goldberg, D. J., McMenamin, J., Cameron, S., and Leigh Brown, A. J. (1997). Molecular investigation into outbreak of HIV in a Scottish prison. *BMJ* 314, 1446–1450.

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

Received: 26 June 2012; paper pending published: 07 July 2012; accepted: 17 July 2012; published online: 31 July 2012.

Citation: Shiino T (2012) Phylodynamic analysis of a viral infection network. Front. Microbio. ***3****:278. doi: 10.3389/fmicb.2012.00278*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

frontiers in MICROBIOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 21 May 2012
doi: 10.3389/fmicb.2012.00178

Estimating the risk of re-emergence after stopping polio vaccination

Akira Sasaki [1,2] ****, Yoshihiro Haraguchi*** [3] ***and Hiromu Yoshida*** [4]

[1] *Department of Evolutionary Studies of Biosystems, The Graduate University for Advanced Studies, Hayama, Kanagawa, Japan*
[2] *Evolution and Ecology Program, International Institute for Applied Systems Analysis, Laxenburg, Austria*
[3] *Department of Biology, Faculty of Science, Kyushu University Graduate Schools, Fukuoka, Japan*
[4] *Department of Virology 2, National Institute of Infectious Diseases, Musashimurayama, Tokyo, Japan*

Edited by:
Hiroyuki Toh, National Institute of Advanced Industrial Science and Technology, Japan

Reviewed by:
Hironori Sato, National Institute of Infectious Diseases, Japan
Alun Lloyd, North Carolina State University, USA

****Correspondence:***
Akira Sasaki, Department of Evolutionary Studies of Biosystems, The Graduate Universityu for Advanced Studies, Hayama, Kanagawa 240-0193, Japan.
e-mail: sasaki_akira@soken.ac.jp

Live vaccination against polio has effectively prevented outbreaks in most developed countries for more than 40 years, and there remain only a few countries where outbreaks of poliomyelitis by the wild strain still threaten the community. It is expected that worldwide eradication will be eventually achieved through careful surveillance and a well-managed immunization program. The present paper argues, however, that based on a simple stochastic model the risk of outbreak by a vaccine-derived strain after the cessation of vaccination is quite high, even if many years have passed since the last confirmed case. As vaccinated hosts are natural reservoirs for virulent poliovirus, the source of the risk is the vaccination itself, employed to prevent the outbreaks. The crisis after stopping vaccination will emerge when the following two conditions are met: the susceptible host density exceeds the threshold for epidemics and the vaccinated host density remains large enough to ensure the occurrence of virulent mutants in the population. Our estimates for transmission, recovery, and mutation rates, show that the probability of an outbreak of vaccine-derived virulent viruses easily exceeds 90%. Moreover, if a small fraction of hosts have a longer infectious period, as observed in individuals with innate immunodeficiency, the risk of an outbreak rises significantly. Under such conditions, successful global eradication of polio is restricted to a certain range of parameters even if inactivated polio vaccine (IPV) is extensively used after the termination of live vaccination.

Keywords: vaccine-derived strain, live vaccination, risk of re-emergence, silent circulation, poliovirus, branching process, demographic stochasticity, epidemiological dynamics

INTRODUCTION

The World Health Organization (WHO) has a target to interrupt wild poliovirus transmission throughout the world by 2013 (World Health Organization, 2010). The number of patients with poliomyelitis by wild-type poliovirus infection has decreased drastically due to a program using live oral polio vaccine (OPV). Immunity by OPV is defensible against excreted viruses because the major antigenic sites on the viral genome are relatively conserved between serotypes during replication (Minor, 1992). However, nucleotide substitutions responsible for increased neurovirulence frequently occur during replication in the human gut (Poyry et al., 1988; Dunn et al., 1990; Abraham et al., 1993; Kew et al., 1998; Matsuura et al., 2000; Shulman et al., 2000). It has been reported since the 1960s that the vaccine-derived strain excreted from humans can exhibit pathogenicity (Benyesh-Melnick et al., 1967; Marker Test Subcommittee. The Japan Live Poliovaccine Research Commission, 1967). This suggests the possibility that vaccine-derived viruses could cause a poliomyelitis outbreak in a susceptible population after the cessation of an OPV program (Wood et al., 2000). The objective of this study was to estimate the risk of outbreak of vaccine-derived strains after stopping OPV. While the number of attenuated virus carriers, the source of neurovirulent viruses, would decline after the discontinuation of OPV, the number of susceptible hosts would increase and may finally exceed the threshold for an outbreak. Therefore, successful eradication depends on which of these processes is faster. We calculated the probability of successful global eradication, that is, the probability that the last carrier will be recovered before the population could experience an outbreak.

It will be shown below that the mean excretion period from an infected individual is one of the key factors that determine whether or not eradication fails. Except for immunodeficient individuals, virus is excreted from humans for ∼1–3 months after OPV administration to a susceptible host (Alexander et al., 1997). Excreted viruses are often virulent. For example, Yoshida et al. (2000) showed that type 3 vaccine-derived polioviruses isolated from an environment in Japan had high neurovirulence. These strains were isolated from river or sewage waters ∼3 months after routine OPV administration, showing that vaccine-derived strains could circulate in the human community. Other studies showed silent circulation of vaccine-derived strains occurred in the human community (Zdrazilek et al., 1982; Miyamura et al., 1992).

To avoid risks such as contact infection or vaccine-associated paralysis (VAP), inactivated polio vaccine (IPV) has been used in several countries (Murdin et al., 1996). The USA switched its

immunization strategy from OPV to IPV in 2000 (American Academy of Pediatrics Committee on Infectious Diseases, 1999). As IPV-immunized hosts can be infected by polioviruses and excrete infectious virus, IPV is less effective than OPV in preventing infection, though numbers of excreted viruses are greatly reduced (Fine and Carneiro, 1999). Our study also investigated whether switching to IPV after the cessation of OPV effectively reduced outbreak risk.

The Pan American Health Organization (PAHO) reported a poliomyelitis outbreak by a type 1 vaccine-derived strain in Haiti and the Dominican Republic in July 2000 (Centers for Disease Control and Prevention, 2000). In the Latin American region, poliomyelitis caused by a wild strain was last reported in Peru in 1991, and eradication of poliomyelitis was declared in 1994. The recent outbreak in Haiti and the Dominican Republic could be ascribed to the decreased rate of OPV coverage and the spread of a neurovirulent vaccine-derived strain.

The polio eradication program plans to stop administering OPV after disappearance of the wild strain. If vaccine-derived strains remain when herd immunity falls below the epidemic threshold, outbreak by these strains could occur. In this paper, we study the probability of disease re-emergence caused by a vaccine-derived strain using a simple mathematical model. Epidemiological and genetic parameters, such as transmission rate, mean excretion period, mutation rate from attenuated to neurovirulent strains, are varied around estimated values (Gelfand et al., 1959; Benyesh-Melnick et al., 1967; Dunn et al., 1990; Fine and Carneiro, 1999), and dependence on the probability of eradication detailed. In assessing the risk we assumed the following:

1. That the excretion period of vaccine-derived neurovirulent viruses can be longer than that of the attenuated viruses used in live immunization. Likewise, the transmission rates of vaccine-derived strain can be greater than that of the attenuated strain. When hosts recover from infection by either viral strain, the degree of immunity is as effective as that raised by OPV immunization.
2. That infection by either the vaccine-derived or attenuated poliovirus can occur in IPV-immunized hosts. However, the number of secondary transmissions from a previously IPV-immunized host is smaller than that from a susceptible host, and the mean excretion period is shorter in an IPV-immunized host than in a susceptible host.
3. That when re-infection occurs in an individual immunized by OPV, excretion from the re-infection is ignored because the amount of virus excretion is negligibly small (Abraham et al., 1993).
4. That antigenic drift does not occur. The focus of the study is on the risk of outbreak by a neurovirulent vaccine-derived strain with unchanged antigenic properties.
5. That a constant fraction (e.g., 70%) of hosts is efficiently immunized (seroconverted) before OPV is stopped, and that the population at that time is in endemic equilibrium under constant OPV coverage.

We first examine the risk of outbreak after OPV cessation (in the absence of an alternate program); second, we evaluate the effect of host heterogeneity on excretion duration; and third, we examine outbreak risk where extensive IPV immunization follows OPV cessation.

Mathematical modeling is a powerful tool in the understanding of epidemiological dynamics (Anderson and May, 1991). Previous models of polio eradication have considered neither the re-infection by vaccine-derived strains of IPV-immunized hosts nor mutation giving rise to neurovirulent strains (Eichner and Hadeler, 1995; Eichner and Dietz, 1996). Our model allows for the mutation of attenuated strains to virulent strains while replicating in the human gut (Poyry et al., 1988; Dunn et al., 1990; Abraham et al., 1993; Kew et al., 1998; Matsuura et al., 2000; Shulman et al., 2000), and also allows both strains to infect IPV-immunized hosts. The probability for the success of global eradication is then calculated based on the stochastic model of epidemiological dynamics.

MATERIAL AND METHODS

We attempted to determine the risk of virulent poliovirus outbreaks after stopping live vaccination. Time $t = 0$ represents the point at which immunization by live-poliovirus vaccine (OPV) is stopped. With a sufficiently high rate of immunization, the great majority of the population at time would be OPV-immunized hosts, which neither the attenuated (Sabin) nor virulent strain could infect. We first examined the risk where no alternative program followed OPV cessation. The effect of extensive administration of inactivated vaccine (IPV) following OPV discontinuation will be discussed later.

DETERMINISTIC EPIDEMIOLOGICAL DYNAMICS

The number of carriers of attenuated virus would decline after the end of a live vaccination program. Poliovirus is considered to have been eradicated when the last carrier had recovered. However, while the number of carriers declines, the number of hosts immunized by the live vaccine declines also. When the number of susceptible hosts exceeds a certain threshold, the way is opened for the spread of a virulent poliovirus. Thus, the risk of outbreak critically depends on the speed at which carrier numbers, as the source of virulent mutant virus, decrease, and the speed at which susceptible hosts increase. Therefore, we need to keep track of the changes over time of the following demographic variables: the fraction of susceptible hosts (x), hosts infected with or carrying attenuated virus (y), virulent virus infected hosts (v), and recovered and immune hosts (z), with $x + y + v + z = 1$. The population size K is kept constant over time. A virulent virus strain can emerge through mutation in attenuated virus carriers. The probability of successful eradication, or conversely, the probability of an outbreak by a virulent virus, can be evaluated by constructing a stochastic process for the change in the number of infected hosts. To construct the stochastic process, we first derive the corresponding deterministic dynamics.

Deterministic dynamics before the cessation of OPV

Under the immunization of OPV to newborns the dynamics for x, y, v, and z are

$$\frac{dx}{dt} = -(\beta_a y + \beta_v v)x - ux + u(1-p), \tag{1a}$$

$$\frac{dy}{dt} = \beta_a xy - (u+\gamma_a)y - \mu y + up, \tag{1b}$$

$$\frac{dv}{dt} = \beta_v xv - (u+\gamma_v)v + \mu y, \tag{1c}$$

$$\frac{dz}{dt} = \gamma_a y + \gamma_v v - uz, \tag{1d}$$

where t denotes the time variable in units of weeks, p is the immunization fraction to newborns (the fraction to be immunized times the seroconversion rate), u denotes both the natural mortality and the birth rate of the host where we assume that host population is at demographic equilibrium so that the numbers of births and deaths are balanced, β_a and β_v are the transmission rates of attenuated and virulent virus, respectively, $1/\gamma_a$ and $1/\gamma_v$ are the mean durations of attenuated and virulent virus infection, respectively, and μ is the mutation rate from attenuated to virulent virus (**Figure 1**). As the numbers of births and deaths are balanced [$d(x+y+v+z)/dt = 0$ follows from Eq. (1)], the total population is kept constant (K), and we can focus on the changes in the fraction of each class. As $z(t) = 1 - x(t) - y(t) - v(t)$, we omit Eq. 1d from the analysis. If $\mu = 0$, the condition for virulent or wild polio virus being wiped out from the population is that the immunization fraction p is smaller than the threshold p_c:

$$p > p_c = \left(1 - \frac{1}{R_v}\right)\left(1 - \frac{R_a}{R_v}\right), \tag{2}$$

where $R_v = \beta_v/(u+\gamma_v)$ and $R_a = \beta_a/(u+\gamma_a)$ are the basic reproductive ratios of virulent and attenuated viruses (see, for example, Nowak and May, 2000). The threshold immunization fraction p_c necessary for the eradication of virulent viruses is lower than that without circulation of attenuated viruses ($\tilde{p}_c = 1 - 1/R_v$). Thus silent circulation of attenuated virus can significantly increase the efficiency of vaccination. With non-zero mutation rate $\mu > 0$, both the attenuated and the virulent virus can be maintained in the population. The fractions of susceptible host $\hat{x}$, attenuated virus infected hosts $\hat{y}$, virulent virus infected hosts $\hat{v}$ (and recovered and immune hosts $\hat{z} = 1 - \hat{x} - \hat{y} - \hat{v}$) at endemic equilibrium of dynamics (1) are defined as

$$\hat{y} = \frac{u}{(u+\gamma_a)}\frac{p}{\left\{(1-R_a\hat{x}) + \tilde{\mu}\right\}}, \tag{3a}$$

$$\hat{v} = \frac{u}{(u+\gamma_v)}\frac{p}{\left\{(1-R_a\hat{x}) + \tilde{\mu}\right\}}\frac{\tilde{\mu}}{(1-R_v\hat{x})}, \tag{3b}$$

where $\tilde{\mu} = \mu/(u+\gamma_a)$ and $\hat{x}$ being defined as a positive root of

$$R_a R_v \hat{x}^3 - (R_a + R_v + R_a R_v + \tilde{\mu} R_v)\hat{x}^2 + \left[(1+\tilde{\mu}) + R_a + (1-p+\tilde{\mu})R_v\right]\hat{x} - (1-p)(1+\tilde{\mu}) = 0. \tag{3c}$$

Figure 2 shows how the equilibrium numbers defined above depend on the immunization fraction p and the mutation rate μ, together with the mean number of virulent virus infections per week, $\beta_v \hat{x}\hat{v}$, under immunization.

As we will see later, the success or failure of global eradication after the cessation of OPV critically depends on the equilibrium densities of susceptible, attenuated virus infected, and virulent virus infected hosts at the time of stopping OPV illustrated above. Their parameter dependences are best described if there was no significant difference in transmission rates and recovery rates between attenuated and virulent polio strains, such that we can assume $\beta = \beta_a = \beta_v$, and $\gamma = \gamma_a = \gamma_v$. This is an important special case that is also partly supported from the data (see later). Substituting $\beta_a = \beta_v = \beta$ and $\gamma_a = \gamma_v = \gamma$ into Eqs 3a–3c then yields the equilibrium factions under OPV immunization in symmetric case:

$$\hat{x} = \left[R_0 + 1 - \sqrt{(R_0-1)^2 + 4pR_0}\right]/2R_0, \tag{4}$$

and

$$\hat{y} = \frac{u}{(u+\gamma)}\frac{p}{\left\{(1-R_0\hat{x}) + \tilde{\mu}\right\}},$$

$$\hat{v} = \frac{u}{(u+\gamma)}\frac{p}{\left\{(1-R_0\hat{x}) + \tilde{\mu}\right\}}\frac{\tilde{\mu}}{(1-R_0\hat{x})}, \tag{5}$$

where $R_0 = \beta/(u+\gamma)$ is the basic reproductive ratio of both strains. If R_0 is sufficiently large ($R_0 >> 1$), the equilibrium fractions are approximated as

$$\hat{x} \approx \frac{1-p}{R_0},$$

$$\hat{y} \approx \frac{u}{u+\gamma}\frac{p}{p+\tilde{\mu}}, \tag{6}$$

$$\hat{v} \approx \frac{u}{u+\gamma}\frac{\tilde{\mu}}{p+\tilde{\mu}},$$

which describe well how the equilibrium densities change with the immunization fraction p and mutation rate $\mu = (u+\gamma)\tilde{\mu}$ in the right panels of **Figure 2** (for $\beta_a = \beta_v$).

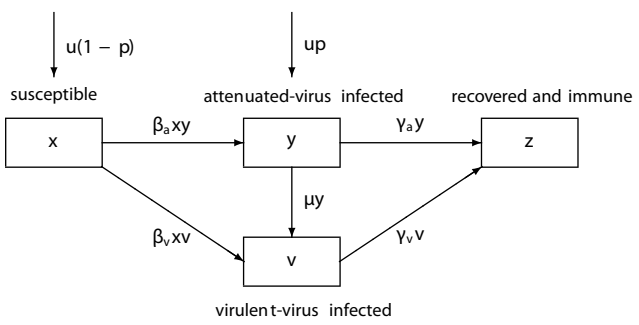


FIGURE 1 | The schematic diagram of the epidemiological dynamics. β_a and β_v: the transmission rate of attenuated and virulent virus, γ_a and γ_v: the recovery rate of attenuated and virulent virus, μ: the mutation rate from attenuated to virulent virus, u: the host birth rate (=death rate), p: the fraction of newborns immunized by OPV. The flows by natural host mortality are omitted.

Deterministic dynamics after the cessation of OPV

The epidemiological dynamics for x, y and v after stopping OPV are

$$dx/dt = -(\beta_a y + \beta_v v)x - ux + u,$$

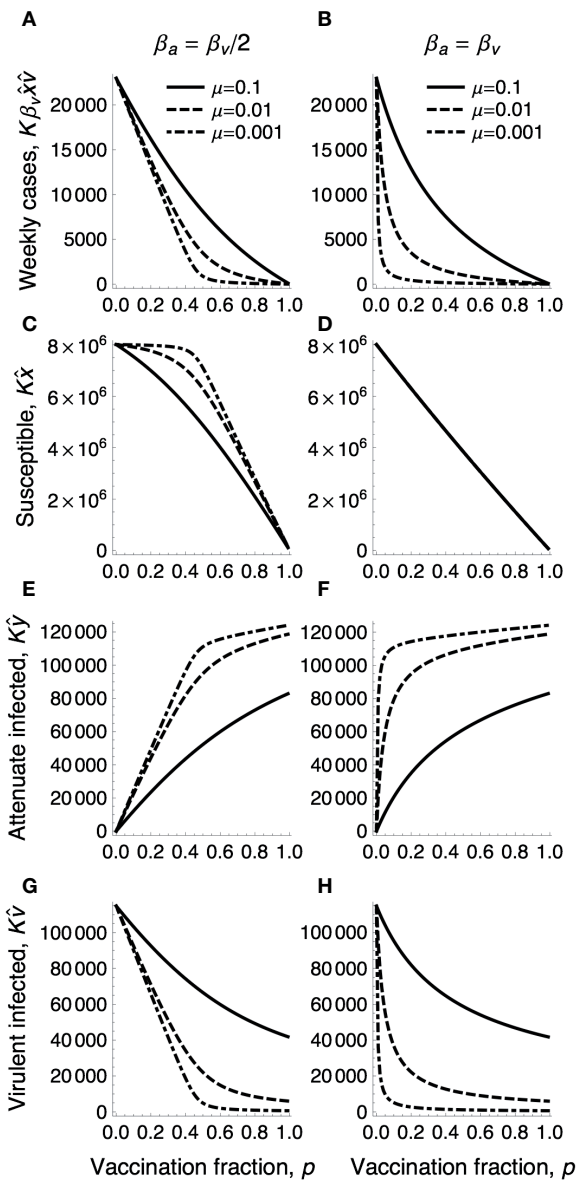


FIGURE 2 | The densities in endemic equilibrium under the immunization fraction *p*. The number, $K\beta_v\hat{x}\hat{v}$ of hosts newly infected by virulent virus in a week (top row), the equilibrium number $K\hat{x}$ of susceptible hosts (second row), that $K\hat{y}$ of attenuated virus infected hosts (third row), and that $K\hat{v}$ of virulent virus infected hosts (bottom row) are plotted as a function of immunization fraction p for varying mutation rates μ for the emergence of virulent virus from an attenuated virus (sold: $\mu = 0.1$, dashed: $\mu = 0.01$, dot dashed: $\mu = 0.001$). The population size K is 100 million, $\beta_v = 2.5$ is the transmission rate of virulent virus. **(A,C,E,G)** The transmission rate of attenuated virus is half of that of virulent virus: $\beta_a = 1.25$. **(B,D,F,H)** β_a is the same as β_v. Other parameters are $\gamma_a = \gamma_v = 0.25$, $u = 0.00025$.

$$dy/dt = \beta_a xy - (u + \gamma_a)y - \mu y, \tag{7}$$
$$dv/dt = \beta_v xv - (u + \gamma_v)v + \mu y,$$

where t is now the number of weeks after OPV is stopped (**Figure 3**). We assume that the population was in endemic equilibrium at time $t = 0$ under a constant fraction p of newborns immunized by OPV. As before, if we can assume that the transmission rates and recovery rates of attenuated and virulent polio strains are the same: $\beta = \beta_a = \beta_v$ and $\gamma = \gamma_a = \gamma_v$, the dynamics can be described by only two variables: x (the fraction of susceptible hosts) and $w = y + v$ (the fraction of hosts infected by either attenuated or virulent virus),

$$dx/dt = -\beta xw - ux + u, \tag{8a}$$
$$dw/dt = \beta xw - (u + \gamma)w. \tag{8b}$$

The susceptible density increases with time, while the densities of attenuated or virulent virus infected hosts decrease with time as long as $t > t_c$, where t_c is the time at which the susceptible density hits the epidemiological threshold: $x(t_c) = (u + \gamma)/\beta$

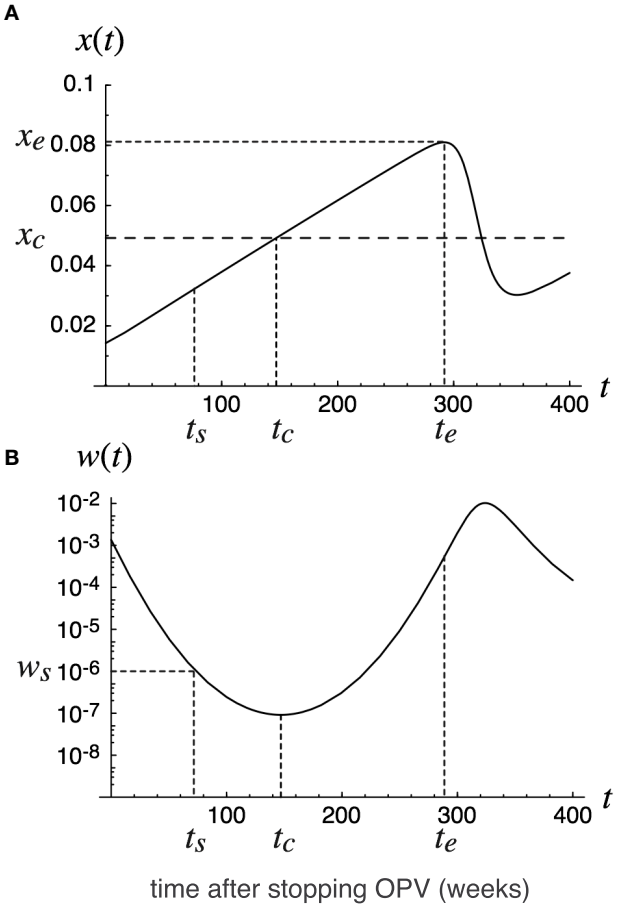


FIGURE 3 | Deterministic trajectory after stopping OPV. Deterministic trajectory of epidemiological dynamics (8) in the text. The fraction $x(t)$ of susceptibles **(A)** and the fraction $w(t)$ of infecteds **(B)** are plotted as functions of the time $t = 0$ since the cessation of OPV. The dotted line indicates the threshold host density for outbreak: $x_c = (u + \gamma)/\beta$. The initial fractions x_0 and w_0 at time $t = 0$ are assumed to be in endemic equilibrium under OPV immunization to a constant fraction, p, of newborns. The time $t = t_c$ at which the fraction of infecteds is minimized in deterministic trajectory is indicated, together with time $t = t_s$ and $t = t_e$ defined for the calculation of the global eradication probability (Eq. 11). Parameters are: $p = 0.7$, $\beta = 3.7$, $\gamma = 0.18$, $u = 0.00025$.

(see **Figure 3**). The poliovirus infected density then starts increasing again. The question we ask in the following is whether the poliovirus goes to extinction around the time $t = t_c$ where its density approaches the minimum. In the following we derive the global eradication probability of poliovirus by analyzing the stochastic analog of dynamics (7) for $\beta_a < \beta_v$ or $\gamma_a > \gamma_v$, and that of the dynamics (8) for the special case of $\beta_a = \beta_v$ and $\gamma_a = \gamma_v$.

PROBABILITY OF SUCCESSFUL ERADICATION

We then examine the probability of poliovirus eventually being lost from a population without causing an outbreak. To calculate extinction probabilities, we consider discrete time dynamics corresponding to (8) with weeks as time units. We assume that the number of secondary infections from a virulent virus infected host per week follows the Poisson distribution with mean $\beta K x(t)$, where K is the total population size and $x(t)$ is the fraction of susceptible hosts defined as the solution to (8). The probability that the progeny of a virulent virus strain found in an infected host at time t eventually goes to extinction by chance before causing an outbreak is defined as $q(t)$. We also define $1 - q(t)$ as the marginal risk of outbreak at time t, which is the probability that an infected host present at time t harbors the viruses whose progeny will cause outbreaks in the future. If $\beta_a = \beta_v = \beta$ and $\gamma_a = \gamma_v \gamma$, the extinction probability $q(t)$ then satisfies the recursive equation

$$q(t) = \left[(1-\delta)q(t+1) + \delta\right] \exp\left[-\beta K x(t)(1 - q(t+1))\right], \quad (9)$$

where $\delta = u + \gamma$ (see Appendix for the derivation). The extinction probability $q(t)$ for arbitrary time t can be determined by solving (9), with $x(t)$ obtained from (5) and (8). The boundary condition for the recursion (9) is chosen at the time at which the fraction x of susceptibles first approaches a local maximum x_e at $t = t_e$ (such x_e and t_e exist because susceptible hosts as unvaccinated newborns should first be boosted after stopping live vaccination until x exceeds the epidemic threshold x_c – see **Figure 3**):

$$q_e = \left[(1-\delta)q_e + \delta\right] \exp\left[-\beta K x_e (1 - q_e)\right], \quad (10)$$

where $q_e = q(t_e)$ is the extinction probability at $t = t_e$. In deriving (10), we used the approximation $q(t_e) \approx q(t_e + 1)$, as the change in $x(t)$ is negligibly small around its maximum x_e.

The probability of eventual eradication can then be calculated as follows. We choose a reference time point $t = t_s$ before the deterministic trajectory for w reaches its minimum (see **Figure 3**), at which the number of infected hosts $Kw_s = Kw(t_s)$ was large enough so that eradication before that time point could be ignored, but small enough so that competition between different viral lines could be ignored. According to extensive Monte Carlo simulations we found that the stochastic loss of the infecteds may occur only after their expected number falls below 100 or less. Noting this and the fact that the competition between viral strains can be ignored when $Kw_s/K << 1$, we chose $Kw_s = 100$. The probability of eventual extinction is then

$$P_{\text{ext}} = q(t_s)^{Kw_s}, \quad (11)$$

i.e., poliovirus eventually goes to extinction without causing outbreaks if and only if all progenies of the viruses present at $t = t_s$ go to extinction. Note that if the total population is subdivided into mutually isolated communities (e.g., 100 cities each with one million population), then the probability that none of the cities experiences the outbreak is given by (11) with $K = 100\times$ one million.

We conducted extensive Monte Carlo simulations of the fully stochastic process to check the accuracy of formula (11). For the Monte Carlo simulations, week by week changes in numbers of susceptibles, attenuated virus infecteds, and virulent virus infecteds in population of size K were followed. The changes between weeks caused by infection, recovery, mutation, and host mortality were generated by binomial pseudo-random numbers with the rates given by the dynamics (7). As shown below, the formula (11) for the probability of eventual eradication agreed quite well with that observed in the Monte Carlo simulations for 1000 independent runs.

EPIDEMIOLOGICAL PARAMETERS

The probability of global eradication depends on epidemiological, host demographic, and genetic parameters. Thus, estimates of the recovery rate γ, the transmission rate β, and the mutation rate μ are critical. All parameters used in the model were scaled in units of weeks.

Recovery rate γ, or the reciprocal of the mean excretion period.

The mean excretion duration after challenge with 6 logs of Sabin type 1 virus has been estimated to be 20.4 days for hosts not previously immunized, 12.3 days for previously IPV-immunized hosts, and 4.6 days for previously OPV-immunized hosts (Fine and Carneiro, 1999). Thus, the mean infectious period of a type 1 primary infection is about 3 weeks. While type 2 poliovirus showed a similar excretion period to type 1, type 3 has a significantly longer excretion period (Vaccine Administration Subcommittee. The Japan Live Poliovaccine Research Commission, 1966). Mean excretion periods are estimated as 20.5, 20.6, and 38.6 days for types 1, 2, and 3, respectively, for TOPV (trivalent oral polio vaccination; Gelfand et al., 1959). Regarding the risk of re-emergence, type 3 poliovirus would be the most likely agent to persist and circulate longest after stopping OPV, and hence cause outbreaks. Therefore we adopted the excretion period for type 3 in assessing outbreak risk. Thus, we varied the recovery rate around $\gamma_a = 0.18$/week, corresponding to 5.5 weeks as the mean excretion period. We assume that the recovery rates are similar between attenuated (γ_a) and virulent (γ_v) polio infections, and set as $\gamma_v = 0.18$. Indeed, durations of excretion of attenuated type 1 polioviruses showed no significant difference from that of wild polioviruses (compare **Figures 2** and **3** of Alexander et al., 1997). A constant recovery rate assumed here implies that the infectious period has the long tail in an exponential distribution. The effect of tail in the infectious period will be examined later.

Transmission rate β, or the mean number of secondary infections.

While the probability of within-family infection was estimated to be 0.5 per case (Benyesh-Melnick et al., 1967), we also needed to evaluate the mean transmission rate to other members of the community. The mean transmission rate was estimated from the basic reproductive rate: $R_0 = \beta/(u+\gamma) \approx \beta/\gamma$. The basic reproductive ratio of wild polioviruses in England and Wales during

the pre-vaccination period has been estimated to be $R_0 = 10$–12 (Anderson and May, 1991). More recent estimates have been $R_0 = 10$–15 in countries with poor sanitation and hygiene, and R_0 less than 10 in countries with good sanitation and hygiene (Fine and Carneiro, 1999). If we assume $\gamma = 0.18$, this gives estimates of $\beta = 1.8$–2.7/week in developing countries. Much higher R_0's of more than 20 have been reported by studies of poliomyelitis outbreaks over the past 20 years (Patriarca et al., 1997). Because of this large variance in the estimated β, we varied the value rather widely, from 2 to 6, to evaluate eradication probability.

Mutation rate μ from the attenuated to the virulent virus

Oral polio vaccine produced from Sabin 1 to 3 strains is a highly attenuated vaccine. It is known that virulent mutants appear after replication in the human gut after OPV given. Such virulent strains have caused outbreaks in populations with low OPV coverage in Haiti, the Dominican Republic, and Egypt, accumulating mutations through human to human transmission (Centers for Disease Control and Prevention, 2000, 2001). Several nucleotide mutations responsible for attenuation have ever been reported (Plotkin et al., 2008). Of them, the critical and unstable attenuating mutations in 5′-UTR (A480G in Sabin 1, G481A in Sabin 2, and C472U in Sabin 3) appear initially during viral replication. Dunn et al. (1990) reported that at least one viral serotype excreted from a susceptible individual immunized by OPV had mutated completely in 5′-UTR within 28 days. The average contents of revertants (virulent forms) from OPV recipients were 28–40% in type 1, 97% in type 2, and 67% in type 3 at 3 weeks after the most recent dose (Laassri et al., 2006). Similar estimates were reported by Minor et al. (2005) and Martinez et al. (2004). Thus, the mutation rate from attenuated to virulent viruses appeared to be high, in the order no smaller than $\mu = 0.1$/week.

RESULTS

Before proceeding to specific parameter dependences, it should be noted that the time at which the fraction of susceptible hosts exceeds the threshold for epidemics is crucial in understanding the problem. The number of virulent virus infected hosts increases if the fraction of susceptible hosts is larger than the threshold $x_c = (u+\gamma)/\beta$, which is the reciprocal of the basic reproductive rate $R_0 = \beta/(u+\gamma)$, and decreases when x is smaller than x_c. During the initial period, when the fraction of OPV-vaccinated individuals is large, the fraction of susceptibles is less than the threshold x_c, so that the risk of an outbreak is negligible, even though considerable numbers of virulent mutants are being generated at each time step. The number of virus carriers decreases during the period from the cessation of OPV to time t_c at which the susceptible density exceeds the threshold x_c. If the number of carriers becomes zero around t_c, polio will be globally eradicated. However, if virus survives this "endangered" period around t_c, the infected density increases again and a future outbreak becomes certain. The following formula (derived in Appendix) provides an approximate time t_c and minimum infected fraction w_c as a function of epidemiological parameters:

$$t_c \approx Lp/R_0, \quad (R_0 >> 1), \tag{12a}$$

$$Kw_c \approx K\frac{D}{L}exp\left[-\frac{p^2}{2R_0}\frac{L}{D}\right], \quad (R_0 >> 1,\ L >> D), \tag{12b}$$

where $D = 1/\gamma$ is the mean duration of infection, $L = 1/u$ the life expectancy of the host, and $R_0 = \beta/(u+\gamma)$ the basic reproductive ratio. There is a high probability of global eradication if Kw_c is sufficiently smaller than 1; whereas, there is a high-risk of re-emergence if Kw_c is greater than 10. Although assessment of outbreak risk should be based on the probability of global viral extinction as discussed below, the above approximate formula gives insights into the likelihood of re-emergence and parameter dependence on eradication probability. It also gives an accurate estimate of the critical time t_c at which either global eradication occurs or an outbreak starts.

PATHS TO EXTINCTION AND PATHS TO OUTBREAK

Figure 3 shows deterministic changes in fraction x of susceptibles and fraction $w = y + v$ of poliovirus carrying hosts after cessation of live vaccination. The fraction of susceptibles exceeded the epidemiological threshold x_c around time $t = t_c(=150)$ weeks after live vaccination discontinuation. When the fraction of susceptibles exceeds the epidemiological threshold, the fraction of infecteds is at its minimum. The public health objective is to make the number of infecteds zero around time $t = t_c$. **Figure 4** illustrates sample paths for the stochastic process corresponding to the deterministic trajectory in **Figure 3**. In this example, 61 out of 100 independent

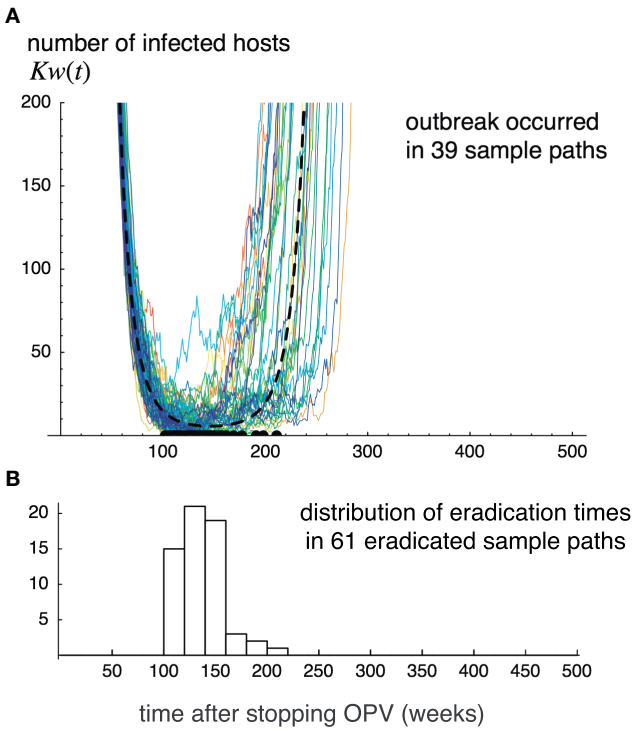


FIGURE 4 | Sample paths for the number of infecteds observed in Monte Carlo simulations. Sample paths for the number of infecteds observed in Monte Carlo simulations of the stochastic process corresponding to dynamics (7). One hundred independent runs are illustrated by thin lines. Thick broken lines indicate the deterministic trajectory **(A)**. The histogram shows the distribution for the times at which viruses went to extinction **(B)**. 38 out of 100 runs never go to extinction, and cause outbreaks. The parameters are the same as in **Figure 3**, and $K = 10^8$.

runs led to the global eradication of poliovirus (i.e., the number of infected hosts hit the absorbing boundary at zero). However, in the remaining runs, poliovirus escaped extinction around $t = t_c$, increased again, leading to an outbreak by a virulent strain. The probability of successful eradication is thus 61% by the parameter set used in **Figure 4**.

PARAMETER DEPENDENCE

Figure 5 illustrates how the probability of the failure of global eradication $P_{\text{fail}} = 1 - P_{\text{ext}}$ depends on each parameter, which we discuss in turn below. We set the following values as "standards," and varied each of the parameters to see its effect. The fraction of immunized newborns before $t = 0$: $p = 0.7$; transmission rate of virulent virus: $\beta_v = 3.7$, that of attenuated virus: either $\beta_a = \beta_v$ or $\beta_a = \beta_v/2$; recovery rate: $\gamma = 0.18$ (in both viruses); mutation rate from attenuated to virulent viruses: $\mu = 0.1$; natural host mortality: $u = 0.00025$ (all measured in units of weeks), and total population: $K = 100$ million. With the chosen values of β, u, and γ, the basic reproductive rate of polioviruses was $R_0 = 20$. In **Figure 5**, lines indicate the eradication probability calculated from Eqs 8–11 for $\beta_a = \beta_v$, the dots indicate the observed eradication probability for 1000 independent runs of the stochastic process corresponding to the deterministic model (7) for $\beta_a = \beta_v$, and the crosses indicate that for $\beta_a = \beta_v/2$. We first discuss the results for $\beta_a = \beta_v$ in Section "The Immunization Fraction p Before Stopping OPV, The Recovery Rate γ, The Transmission Rate β, The Mutation Rate μ From the Attenuated to Virulent Viruses, and The Total Population Size K" below, and discuss the effect of a lower transmission rate of attenuated virus in 3.2.6.

The immunization fraction p before stopping OPV

The effect of fraction p of OPV-immunized newborns before stopping the live vaccination is illustrated in **Figure 5A**. While the probability of failing eradication is low when p is sufficiently high, it rises drastically around $p = 0.7$ when p is decreased. For example, if the immunization fraction is 60% or less before OPV is stopped, future outbreak by virulent poliovirus is almost certain. There are two reasons why a lower p before stopping OPV enhances the risk of future outbreaks: first, it shortens the time for the susceptible host density to reach the epidemiological threshold, and second, it increases the initial infected density w_0, thereby keeping the minimum density from extinction.

The recovery rate γ

The success of global eradication greatly depends on the recovery rate, or its reciprocal, the mean infectious period (**Figure 5B**). The higher the recovery rate, the more rapidly the number of poliovirus carriers decreases after supply by OPV is stopped. It is then possible to make the expected number of infecteds negligibly small when the susceptible fraction exceeds the epidemiological threshold. Conversely, by having a longer infectious period (a lower recovery rate), viruses safely persist over the endangered period around $t = t_c$. In examples shown in **Figure 5B**, infectious periods

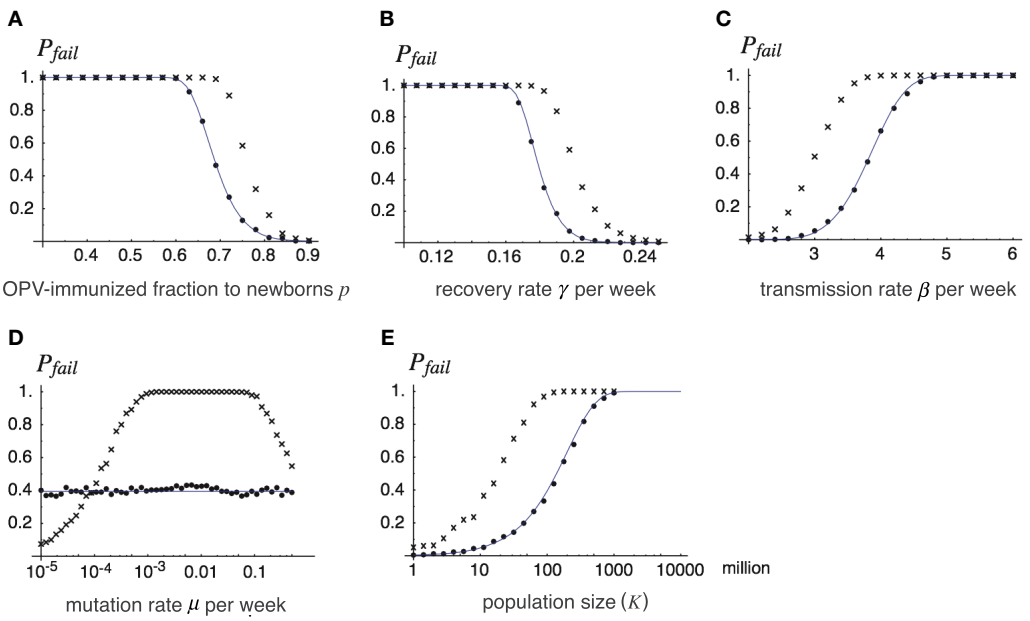


FIGURE 5 | The probability of the failure of global eradication as a function of epidemiological and genetic parameters. Each panel shows how the probability of failing the global eradication $P_{\text{fail}} = 1 - P_{\text{ext}}$ depends on a chosen parameter, where lines show analytical results drawn from P_{ext} defined in Eq. 11, and dots show Monte Carlo simulation results. Except for the varying parameter in each panel, the parameters are fixed as $p = 0.7$, $\beta = 3.7$ ($\beta_v = \beta_a = \beta$ for dots and lines, and $\beta_v = \beta$, $\beta_a = \beta/2$ for cross-hatched), $\gamma = \gamma_v = \gamma_a = 0.18$, $m = 0.1$, $K = 10^8$, and $u = 0.00025$. Varying parameters are: **(A)** fraction p of OPV immunization before its cessation, **(B)** recovery rate γ, **(C)** transmission rate β, **(D)** mutation rate μ, **(E)** total population size K. Lines: the probability of failure obtained from formula (11) in the text (for $\beta_v = \beta_a = \beta$), dots: the proportion of failing eradications in 1000 independent runs of the Monte Carlo simulation for $\beta_v = \beta_a = \beta$, and cross-hatched: that for $\beta_v = \beta$, $\beta_a = \beta/2$.

of 7 weeks or longer are disastrous for eradication. In reality, the infectious period varies between hosts, such that in hosts with innate immunodeficiency the infectious period can be typically longer than 1 year (Hara et al., 1981; Kew et al., 1998). Even a tiny fraction of such hosts significantly increases the risk of virulent virus outbreaks, as we show later.

The transmission rate β

The effect of increasing the transmission rate (**Figure 5C**) is parallel to decreasing the recovery rate described above, and both can be regarded as having the effect of increasing R_0. However, decreasing the recovery rate affects eradication probability more sensitively than increasing the transmission rate, as the former contributes to slowing the decay rate for the number of virus carriers as well as increasing R_0 (see also Eq. 12).

The mutation rate μ *from the attenuated to virulent viruses*

The eradication probability is insensitive to the mutation rate from attenuated to virulent viruses for the case of $\beta_v = \beta_a$ (**Figure 5D**). If viruses persist during the period around $t = t_c$, it does not matter which type survived as eventually the virulent virus increases its relative frequency in the viral population (if $\beta_v = \beta_a$). Quite different results follow when the attenuated virus has a lower transmission rate than the virulent virus (the crosses), where the probability of failing eradication is maximized for an intermediate mutation rate.

The total population size K

This has an obvious dependence on the risk of outbreaks. The larger the population size, the larger the probability that viruses are not lost during the endangered period, and hence, the larger the risk of outbreaks. In the example shown in **Figure 5E**, a population of 10 million individuals has a more than 90% of chance for successful eradication, but communities of 100 and 1000 million have only 50% and less than 5% chances, respectively, using the same epidemiological parameters.

The transmission rate β_a *of attenuated virus smaller than that* β_v *of virulent virus*

In each panel of **Figure 5**, the probability of failing global eradication when the transmission rate β_a of attenuated virus is half of that of virulent virus β_v is plotted as the cross-hatches. In all cases except for the dependence of mutation rate, a lower transmission rate of attenuated viruses *increases* the risk of virulent virus outbreak after the cessation of OPV. This rather counter-intuitive results follow from the fact that silent circulation of attenuated viruses under live vaccination helps increasing the efficiency of immunization, as we have seen in the comparison between the threshold immunization fractions with and without silent circulation [see (2)], and the equilibrium densities for $\beta_a < \beta_v$ (left panels of **Figure 2**) and for $\beta_a = \beta_v$ (right panels). Decreasing the transmission rate of attenuated virus increases the density of susceptibles in the equilibrium population under vaccination, thus shortening the time until the susceptible density hits the epidemiological threshold after the cessation of OPV (compare **Figure 2C** with **Figure 2D**).

TAIL OF INFECTIOUS PERIOD

A constant recovery rate assumed in the previous sections implies that the infectious period is exponentially distributed. One may suspect that an outbreak of vaccine-derived viruses a few years after the cessation of OPV might be the artifact caused by this long tail in the infectious period. We found, however, that the long tail in the infectious period is not necessary for this to happen – it is the silent circulation of avirulent polio viruses in the population, commonly observed in nature and occurring in our model as well, that is responsible for the outbreak that occurs long after the cessation of OPV. To show this, we conducted numerical simulations in which we assume that the host recovers exactly 4 weeks after the infection, i.e., the distribution of infectious period has no tail at all. The infected hosts nevertheless persist in the population far longer than 4 weeks (the infectious period of an individual) after stopping OPV, which allows the outbreak of vaccine-derived strain to occur a few years after the cessation (**Figure 6**).

MARGINAL RISK OF OUTBREAK

Figure 7 illustrates change over time in the marginal risk of viruses found at time *t*. Marginal risk is defined as $1 - q(t)$ – the probability that an infected host present at time t harbors viruses whose progeny will cause a future outbreak. Marginal risk is negligibly small just after $t = 0$, and rapidly increases with t near $t = t_c$. In the parameters used in **Figure 7**, the rate of increase in probability is the highest around $t = 150$ when the susceptible host density exceeds the threshold (see **Figure 3**). However, the marginal risk of viruses before this point is by no means negligible as there is notable probability that progenies of viruses found during $t = 100$–150 would later cause an outbreak.

EFFECT OF A HIGH-RISK GROUP

We here examine the case where a small fraction r of hosts has a recovery rate, γ', much lower than γ for other hosts. In the simulation shown in **Figure 8**, the recovery rate of most individuals was $\gamma = 0.2$. Using this value, successful eradication is certain (other parameters: transmission rate, $\beta = 2.5$; natural mortality, $u = 0.00025$; immunization fraction before stopping OPV, $p = 0.7$; total population, $K = 100$ million). When we assume only 0.01% of newborns have a 10-times longer infectious period than other members, i.e., $\gamma' = 0.1\gamma$, due to innate (World Health Organization, 1989; Fine and Carneiro, 1999), or acquired immunodeficiency, the probability of failure in global eradication rises to 79% (**Figure 8**). Thus even a tiny fraction of high-risk group drastically makes the global eradication difficult.

EFFECTIVENESS OF IPV

What if extensive IPV immunization follows the cessation of OPV? We assume in this case that all newborns are immunized by inactive vaccine before eventual eradication. The probability of global eradication is then evaluated in the light of the results obtained so far by replacing the transmission rates and recovery rates with values for previously IPV-immunized hosts instead of the values for susceptible hosts. IPV cannot prevent infection by either attenuated or virulent viruses, although it can reduce disease severity, and fewer viruses are excreted from IPV-immunized hosts than from

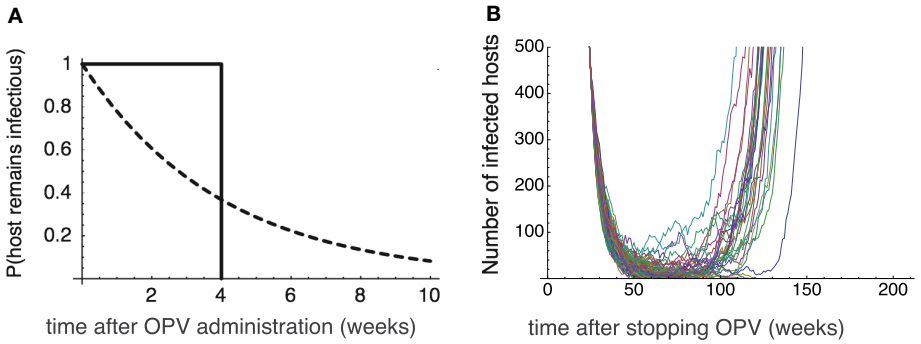


FIGURE 6 | The effect of tail in the infectious period. (A) The probability that the host remains infectious after it is infected at time 0. Dotted curve: the exponential distribution assumed in the previous sections with a constant recovery rate $\gamma = 0.25$/week. Solid curve: the truncated distribution in which all the hosts recovers exactly 4 weeks after the infection. **(B)** The Monte Carlo simulation results assuming the truncated distribution of the infectious period. The time change in the number of virus infected hosts since OPV is stopped. Twenty-six out of 100 runs never go to extinction, and cause outbreaks. The emergence of virulent virus occurs after 50–60 weeks after the secession of OPV. The parameters are $\beta_a = 2.5$, $\beta_v = 5$, $u = 0.00025$, $p = 0.6$, $\mu = 0.1$, and $K = 10^8$. The "mean" infectious period is 4 weeks.

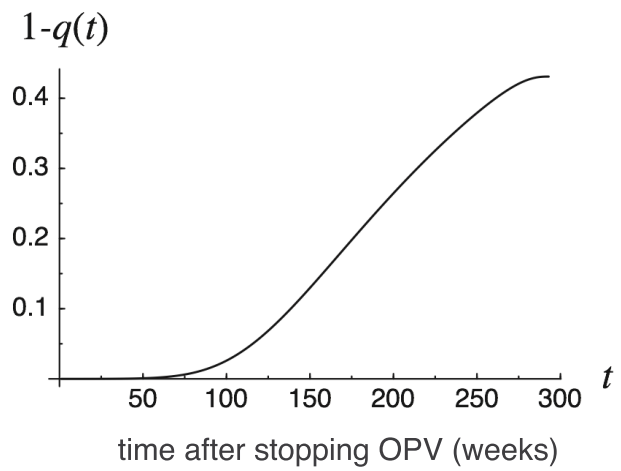


FIGURE 7 | Marginal risk $1 - q(t)$ of outbreaks as a function of time t since OPV cessation. The marginal risk $1 - q(t)$ is defined as the probability that an infected host present at time t harbors viruses whose progeny will cause outbreaks in the future. $p = 0.7$, $\beta = 3.7$, $\gamma = 0.18$, $u = 0.00025$, $K = 10^8$.

unvaccinated hosts (Henry et al., 1966). IPV vaccination would therefore reduce the transmission rate and increase the global eradication probability (see **Figure 5C**). Also, IPV immunization reduces the infectious period, again increasing the probability of successful eradication (**Figure 5B**). However, these considerations assume that *all* hosts are IPV-immunized after the cessation of OPV. The actual amount of risk reduction by IPV depends on coverage, vaccine efficiency, and host heterogeneity in the excretion period.

DISCUSSION

The PAHO and WPRO (Regional Office for the Western Pacific) declared the eradication of poliomyelitis in 1994 and 2000, respectively. Nevertheless, an outbreak of poliomyelitis caused by a type 1 vaccine-derived strain was reported in Haiti and the Dominican Republic in 2000 (Centers for Disease Control and Prevention, 2000), and an outbreak by a type 2 vaccine-derived strain has been reported in Egypt (Centers for Disease Control and Prevention, 2001), in Nigeria (Wassilak et al., 2011). It is assumed that both cases were due to the low rate of vaccine coverage. Although OPV or IPV immunization have been effective in controlling the transmission of wild-type strains, cases of re-emergence by wild-type strains have been reported in several countries (Patriarca et al., 1997) in which inadequate vaccine potency or a high rate of unimmunized individuals led to low herd immunity in the population.

According to a review by Patriarca et al. (1991) rates of seroconversion by OPV approached 100% for each serotype in industrialized countries, but were ∼70% for types 1 and 3 in developing countries. Many studies have demonstrated that interference by enteroviruses in human gut and other factors in OPV administration affect the seroconversion rate (Triki et al., 1997). Thus, even if OPV coverage is as high as 90%, the immunized fraction p in our model becomes 62%, under the 70% seroconversion rate observed in developing countries. This should invoke serious concern if we recall that the reduction in immunization fraction p before cessation of OPV drastically increases the risk of outbreak, as shown in **Figure 5A**).

Our results have specifically shown that a herd immunity level of less than 60% before the cessation of OPV led to the failure of poliovirus eradication under typical epidemiological parameters adopted in this paper. This suggests that maintaining more than 90% OPV coverage is not enough to ensure successful eradication, and that every effort should be made to increase the seroconversion rate in developing countries. Another important parameter affecting the probability of eradication is the recovery rate γ estimated from the mean infectious period. Most data concerning virus excretion rates available from field studies were for the type 1 vaccine strain (Alexander et al., 1997), while much less information is available for types 2 and 3. As type 2 and particularly type 3 have longer excretion periods than type 1, these strains are

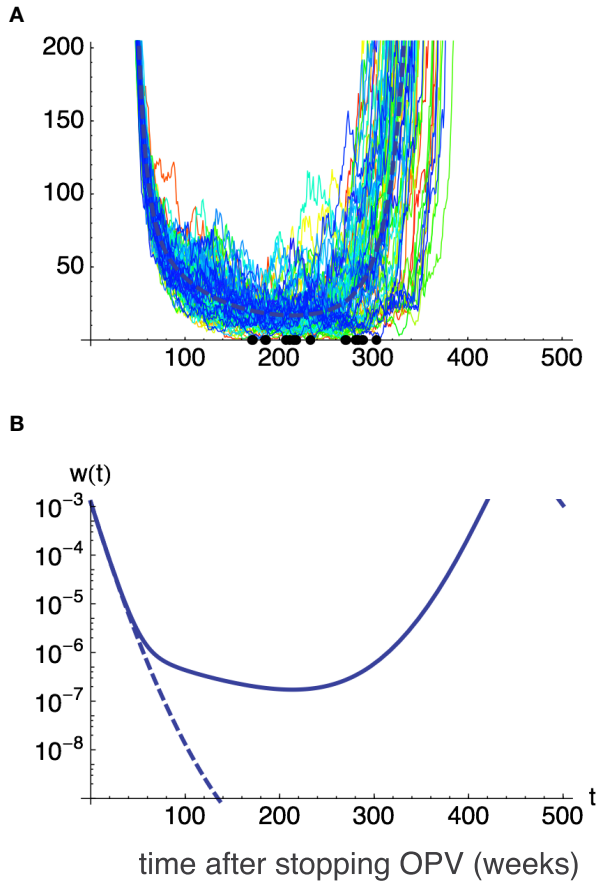


FIGURE 8 | The effect of a high-risk group on global eradication. One in 10,000 (0.01%) of hosts are assumed to be born having a longer excretion period (lower recovery rate γ') when infected by virus. Remaining hosts have the recovery rate γ. $\gamma' = 0.1\gamma$ and $\gamma = 0.2$ is assumed and values $p = 0.7$, $\beta = 2.5$, $\mu = 0.1$, $u = 0.00025$, $K = 10^8$ are used for other parameters. Without the high-risk group, i.e., when all hosts have the recovery rate $\gamma = 0.2$, global eradication is certain. However, with the addition of a fraction 0.01% of high-risk group in the population, eradication fails in 79 out of 100 independent runs, allowing the outbreak of virulent virus. **(A)** Sample paths for the number of infecteds for 100 independent runs. Thick broken lines show the deterministic trajectory. **(B)** Deterministic trajectories for the fraction of infected $w(t)$ when all hosts have the recovery rate γ (broken line), and when 0.01% of hosts have lower recovery rate γ' (solid line).

more likely to persist after cessation of OPV and be the causative agents of outbreaks. In assessing risk, we varied the recovery rate in the range $\gamma = 0.1 - 0.25$, based on estimates for the excretion period of type 3 poliovirus, which appears to have the longest excretion period. Whether this overestimates the risk will eventually be settled by more accurate estimations of excretion periods. However, there may not be enough time to allow the necessary studies, and action may need to taken now assuming the worst possible scenario.

We have shown that even when the mean infectious period is far below the fatal level for eradication failure (e.g., less than 7 weeks in the example shown in **Figure 5B**), the presence of a tiny fraction of immunodeficient individuals greatly increases the risk of disease re-emergence. This was because the primary immunodeficient group acts as a long-term viral reservoir, allowing the virus to persist through the endangered period around t_c (which comes typically 150–200 weeks after the cessation of OPV). At present, no evidence exists whether secondary immunodeficient groups, such as HIV infected patients, could act as a long-term reservoir of poliovirus, but it is possible. Monitoring virus excretion from such high-risk groups would become critically important.

Another factor that drastically increases the risk of polio outbreak after the cessation of OPV is lower transmission rate β_a of attenuated viruses than that β_v of vaccine-derived virulent viruses, as we have shown in **Figure 5** where the results for $\beta_a = \beta_v/2$ is compared with the case $\beta_a = \beta_v$. If we further reduces the transmission rate of attenuated viruses to $\beta_a = \beta_v/4$, the risk of outbreak rises up still more (not shown). This rather unexpected and hazardous dependency comes from the fact that silent circulation of attenuated viruses under vaccination is beneficial in increasing the efficiency of herd immunity. The more is the transmission rate of attenuated viruses, the less is the fraction of hosts that remain susceptible under a fixed vaccination rate. Reducing the transmission rate of attenuated viruses thus increases the susceptible density under vaccination, and hence shortens the time until the susceptible density hits the epidemiological threshold after the cessation of OPV.

Transmission rates (β) can be estimated from R_0, which in turn have been estimated from the mean host age at infection (Anderson and May, 1982; Patriarca et al., 1997; Fine and Carneiro, 1999). Such surveys indicate that R_0 of vaccine-derived poliovirus lies in the range 5–25, depending on the hygiene levels of the region. This is well above the threshold $R_0 = 1$ that allows circulation in susceptible hosts. Eradication probability can be increased by reducing the transmission rate, i.e., by preventing vaccine-derived viruses from circulating in the population as much as possible. Public health attempts to reduce contact with infectious individuals becomes important in reducing the transmission rate β. At the same time, monitoring the circulation of shed virus in the healthy human population and environment becomes even more important after the last round of OPV.

Many studies have shown that immunity by IPV cannot prevent re-infection by poliovirus (Murdin et al., 1996). However, IPV immunization reduces mean excretion duration by 40% compared to unimmunized cases, thus increasing the recovery rate γ by 67% (Henry et al., 1966). IPV also reduces the transmission rate because the number of excreted viruses per unit time also declines. As a result of the increased γ and decreased β, the probability of eradication is higher if IPV immunization follows the cessation of OPV than if no program follows it. Although eradication cannot be achieved without OPV, IPV should be considered, together with its high seroconversion rate, as the primary follow-up strategy after OPV cessation to prevent the secondary transmission of vaccine-derived virus (Ghendon and Robertson, 1994; Sutter et al., 2000).

Neither escape-mutation by antigenic drift (Nowak and May, 1991; Nowak et al., 1991; Sasaki, 1994; Haraguchi and Sasaki, 1997; Sasaki and Haraguchi, 2000) nor the emergence of vaccine-resistant strains (Anderson and May, 1991; McLean,

1995) is considered in this paper, though, in our analysis of IPV immunization, both attenuated and virulent viruses can be regarded as IPV-resistant strains. The presence of multiple serotypes in the viral population complicates the eradication strategy (Lipsitch, 1997). The reason we have ignored such factors in this model of polio eradication is the observation that nucleotide divergence within the VP1 region, which includes the antigenic site, is less than 1.4% in vaccine strains, enabling the protection by OPV or IPV immunization (Matsuura et al., 2000). In a study using a monoclonal antibody toward a vaccine strain, substitutions in the VP1 region did affect neutralization (Wiegers et al., 1989). However, these vaccine-derived strains could still be neutralized by polyclonal antiserum (Matsuura et al., 2000), or be prevented under well-maintained herd immunity (Iwai et al., 2008).

Our model suggests that susceptible host density exceeds the threshold around the time $t_c \approx Lp/R_0$ after the cessation of OPV (e.g., $t_c = 140$ weeks when life expectancy $L = 1/u = 4000$ weeks, immunization fraction $p = 0.7$ and basic reproductive ratio $R_0 = 20$). During the dangerous period around t_c, additional surveillance systems other than normal AFP (acute flaccid paralysis) surveillance should be organized to reduce the risk of re-emergence:

1. Seroepidemiological surveillance of the seroconversion rate within a population. For communities with low seroconversion rates, additional immunization by IPV should be offered. Herd immunity should be maintained at a level over 80% seroconversion.
2. Surveillance of the environment and of shed virus from the source of infection. Upon poliovirus isolation, immunization by IPV is to be administrated to the risk area.
3. Public health administration. A hygiene control program (hand washing practice, use of disposal diapers, etc.) would contribute to the reduction in transmission rate β, preventing the virus from circulating.
4. Monitoring of high-risk groups such as immunodeficient individuals. It is very difficult to use IPV globally due to economic reasons and other administrative difficulties. IPV immunization in restricted regions and in at-risk communities, together with good surveillance systems and hygiene control programs, would be more practical tactics to globally extinguish vaccine-derived viruses.

ACKNOWLEDGMENTS

This work was supported in part by The Graduate University for Advanced Studies (Sokendai).

REFERENCES

Abraham, R., Minor, P., Dunn, G., Modlin, J., and Ogra, P. (1993). Shedding of virulent poliovirus revertants during immunization with oral poliovirus vaccine after prior immunization with inactivated polio vaccine. *J. Infect. Dis.* 168, 1105–1109.

Alexander, J. P. Jr., Gary, H. E. Jr., and Pallansch, M. A. (1997). Duration of poliovirus excretion and its implications for acute flaccid paralysis surveillance: a review of the literature. *J. Infect. Dis.* 175(Suppl. 1), S176–S182.

American Academy of Pediatrics Committee on Infectious Diseases. (1999). Poliomyelitis prevention: revised recommendations for use of inactivated and live oral poliovirus vaccines. *Pediatrics* 103, 171–172.

Anderson, R. M., and May, R. M. (1982). Directly transmitted infections diseases: control by vaccination. *Science* 215, 1053–1060.

Anderson, R. M., and May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control.* Oxford: Oxford University Press.

Benyesh-Melnick, M., Melnick, J. L., Rawls, W. E., Wimberly, I., Oro, J. B., Ben-Porath, E., and Rennick, V. (1967). Studies of the immunogenicity, communicability and genetic stability of oral poliovaccine administered during the winter. *Am. J. Epidemiol.* 86, 12–136.

Centers for Disease Control and Prevention. (2000). Public health dispatch: Outbreak of poliomyelitis – dominican Republic and Haiti, 2000. *JAMA* 49, 1094–1103.

Centers for Disease Control and Prevention. (2001). Circulation of a type 2 vaccine- derived poliovirus – Egypt, 1982-1993. *MMWR Morb. Mortal. Wkly. Rep.* 50, 41–42, 51.

Dunn, G., Begg, N. T., Cammack, N., and Minor, P. D. (1990). Virus excretion and mutation by infants following primary vaccination with live oral poliovaccine from two sources. *J. Med. Virol.* 32, 92–95.

Eichner, M., and Dietz, K. (1996). Eradication of poliomyelitis: when can one be sure that polio virus transmission has been terminated? *Am. J. Epidemiol.* 143, 816–822.

Eichner, M., and Hadeler, K. P. (1995). Deterministic models for the eradication of poliomyelitis: vaccination with the inactivated (IPV) and attenuated (OPV) polio virus vaccine. *Math. Biosci.* 127, 149–166.

Fine, P. E. M., and Carneiro, I. A. M. (1999). Transmissibility and persistence of oral polio vaccine virus: Implications for the global poliomyelitis eradication initiative. *Am. J. Epidemiol.* 150, 1001–1021.

Gelfand, H. M., Potash, L., LeBlanc, D. R., and Fox, J. P. (1959). "Revised preliminary report on the Louisiana observation of the natural spread within families of living vaccine strains of poliovirus," in *Live Poliovirus Vaccines, Volume Scientific Publication No. 44*, ed. C. H. Stuart-Harris (Washington, DC: Pan American Sanitary Bereau), 203–217.

Ghendon, Y., and Robertson, S. E. (1994). Interrupting the transmission of wild polioviruses with vaccines: immunological considerations. *Bull. World Health Organ.* 72, 973–983.

Hara, M., Saito, Y., Komatsu, T., Kodama, H., Abo, W., Chiba, S., and Nakao, T. (1981). Antigenic analysis of polioviruses isolated from a child with a gammaglobulinemia and paralytic poliomyelitis after Sabin vaccine administration. *Microbiol. Immunol.* 25, 905–913.

Haraguchi, Y., and Sasaki, A. (1997). Evolutionary pattern of intra-host pathogen anti- genic drift: Effect of cross-reactivity in immune response. *Philos. Trans. R. Soc. Lond. B Boil. Sci.* 352, 11–20.

Henry, J. L., Jaikaran, E. S., Davies, J. R., Tomlinson, A. J., Mason, P. J., Barnes, J. M., and Beale, A. J. (1966). A study of poliovaccination in infancy: excretion following challenge with live virus by children given killed or living poliovaccine. *J. Hyg. (Lond.)* 64, 105–120.

Iwai, M., Takizawa, T., Nakayama, T., Matsuura, K., Yoshida, H., Hasegawa, S., Obara, M., Horimoto, E., Kurata, T., and Horie, H. (2008). Evaluation of a two-dose administration of live oral poliovirus vaccine for wild and virulent vaccine-derived poliovirus type 1, 2, 3 strains in Japan. *Scand. J. Infect. Dis.* 40, 247–253.

Kew, O. M., Sutter, R. W., Nottay, B. K., McDonough, M. J., Prevots, D. R., Quick, L., and Pallansch, M. A. (1998). Prolonged replication of a type 1 vaccine-derived poliovirus in an immunodeficient patient. *J. Clin. Microbiol.* 36, 2893–2899.

Laassri, M., Lottenbach, K., Belshe, R., Rennels, M., Plotkin, S., and Chumakov, K. (2006). Analysis of reversions in the 5'-untranslated region of attenuated poliovirus after sequential administration of inactivated and oral poliovirus vaccines. *J. Infect. Dis.* 193, 1344–1349.

Lipsitch, M. (1997). Vaccination against colonizing bacteria with multiple serotypes. *Proc. Natl. Acad. Sci. U.S.A.* 94, 6571–6576.

Marker Test Subcommittee. The Japan Live Poliovaccine Research Commission. (1967). Evaluation of Sabin live poliovirus vaccine in Japan. IV. Maker tests on poliovirus strains recovered from vaccinees and their contacts. *Jpn. J. Med. Sci. Biol.* 20, 167–173.

Martinez, C., Old, M. O., Kwock, D. K., Khan, S. S., Garcia, J. J., Chan, C. S., and Webster, R., Falkovitz-Halpern, M. S., and Maldonado, Y. A. (2004). Shedding of Sabin poliovirus type 3 containing the nucleotide 472 uracil-to-cytosine point mutation after administration of oral poliovirus vaccine. *J. Infect. Dis.* 190, 409–416.

Matsuura, K., Ishikura, M., Yoshida, H., Nakayama, T., Hasegawa, S., Ando, S., Horie, H., Miyamura, T., and Kitamura, T. . (2000). Assessment of poliovirus eradication in Japan: Genomic analysis of the polioviruses isolated from the river water and the sewage in Toyama prefecture. *Appl. Environ. Microbiol.* 66, 5087–5091.

McLean, A. R. (1995). Vaccination, evolution and changes in the efficacy of vaccines - a theoretical framework. *Proc. Biol. Sci.* 261, 389–393.

Minor, P. D. (1992). The molecular biology of poliovaccines. *J. Gen. Virol.* 73, 3065–3077.

Minor, P. D., Dunn, G., Ramsay, M. E., and Brown, D. (2005). Effect of different immunisation schedules on the excretion and reversion of oral poliovaccine strains. *J. Med. Virol.* 75, 153–160.

Miyamura, K., Yamashita, K., Yamadera, S., Kato, N., Akatsuka, M., Hara, M., Inouye, S., and Yamazaki, S. (1992). Poliovirus surveillance: isolation of polioviruses in Japan, 1980-1991. A report of the National Epidemiological Surveillance of Infectious Agents in Japan. *Jpn. J. Med. Sci. Biol.* 45, 203–214.

Murdin, A. D., Barreto, L., and Plotkin, S. (1996). Inactivated poliovirus vaccine: past and present experience. *Vaccine* 14, 735–746.

Nowak, M. A., Anderson, R. M., McLean, A. R., Wolfs, T. F. W., Goudsmit, J., and May, R. M. (1991). Antigenic diversity thresholds and the development of AIDS. *Science* 254, 963–969.

Nowak, M. A., and May, R. M. (1991). Mathematical biology of HIV infection: antigenic variation and diversity threshold. *Math. Biosci.* 106, 1–21.

Nowak, M. A., and May, R. M. (2000). *Viral Dynamics.* Oxford: Oxford University Press.

Patriarca, P. A., Sutter, R. W., and Oostvogel, P. M. (1997). Outbreaks of paralytic poliomyelitis, 1976-1995. *J. Infect. Dis.* 175(Suppl. 1), S165–172.

Patriarca, P. A., Wright, P. F., and John, T. J. (1991). Factors affecting the immunogenicity of oral poliovirus vaccine in developing countries: review. *Rev. Infect. Dis.* 13, 926–939.

Plotkin, S., Orenstein, W., and Offit, P. (2008). *Vaccine,* 5th Edn. Saunders: Elsevier, 647–650.

Poyry, T., Stenvik, M., and Hovi, T. (1988). Viruses in sewage waters during and after a poliomyelitis outbreak and subsequent nationwide oral poliovirus vaccination campaign in Finland. *Appl. Environ. Microbiol.* 54, 371–374.

Sasaki, A. (1994). Evolution of antigen drift/switching - continuously evading pathogens. *J. Theor. Biol.* 168, 291–308.

Sasaki, A., and Haraguchi, Y. (2000). Antigenic drift of viruses within a host: a finite site model with demographic stochasticity. *J. Mol. Evol.* 51, 245–255.

Shulman, L. M., Manor, Y., Handsher, R., Delpeyroux, F., McDonough, M. J., Halmut, T., Silberstein, I., Alfandari, J., Quay, J., Fisher, T., Robinov, J., Kew, O. M., Crainic, R., and Mendelson, E. (2000). Molecular and antigenic characterization of a highly evolved derivative of the type 2 oral poliovaccine strain isolated from sewage in Israel. *J. Clin. Microbiol.* 38, 3729–3734.

Sutter, R. W., Suleiman, A., Malankar, P., Al-Khusaiby, S., Mehta, F., Clements, G. B., Pallansch, M. A., and Robertson, S. E. (2000). Trial of a supplemental dose of four poliovirus vaccines. *New Engl. J. Med.* 343, 767–773.

Triki, H., Abdallah, M. V., Ben Aissa, R., Bouratbine, A., Ben Ali Kacem, M., Bouraoui, S., Koubaa, C., Zouari, S., Mohsni, E., Crainic, R., and Dellagi, K. (1997). Influence of host related factors on the antibody response to trivalent oral polio vaccine in Tunisian infants. *Vaccine* 15, 1123–1129.

Vaccine Administration Subcommittee. The Japan Live Poliovaccine Research Commission. (1966). Evaluation of Sabin live poliovirus vaccine in Japan. II. Clinical, virologic and immunologic effects of vaccine in children. *Jpn. J. Med. Sci. Biol.* 19, 277–291.

Wassilak, S., Pate, M. A., Wannemuehler, K., Jenks, J., Burns, C., Chenoweth, P., Abanida, E. A., Adu, F., Baba, M., Gasasira, A., Iber, J., Mkanda, P., Williams, A. J., Shaw, J., Pallansch, M., and Kew, O. (2011). Outbreak of type 2 vaccine-derived poliovirus in Nigeria: emergence and widespread circulation in an underimmunized population. *J. Infect. Dis.* 203, 898–909.

Wiegers, K., Uhlig, H., and Dernick, R. (1989). N-AgIB of poliovirus type 1: a discontinuous epitope formed by two loops of VP1 comprising residues 96-104 and 141-152. *Virology* 170, 583–586.

Wood, D. J., Sutter, R. W., and Dowdle, W. R. (2000). Stopping poliovirus vaccination after eradication: issues and challenges. *Bull. World Health Orang.* 78, 347–357.

World Health Organization. (1989). Report of a WHO sponsored meeting. Primary immunodeficiency diseases. *Immunodefic. Rev.* 1, 173–205.

World Health Organization. (2010). *Global Polio Eradication Initiative Strategic Plan 2010–2012.* Available at: http://www.polioeradication.org/Portals/0/Document/StrategicPlan/StratPlan2010_2012_ENG.pdf

Yoshida, H., Horie, H., Matsuura, K., and Miyamura, T. (2000). Characterisation of vaccine-derived polioviruses isolated from sewage and river water in Japan. *Lancet* 356, 1461–1463.

Zdrazilek, J., Drasnar, M., Hlavova, H., Jadrnickova, E., Jandasek, L., Kasova, V., Koza, J., Matyasova, I., Uvizl, M., Valihrach, J., and Weigen-dova, J. (1982). Presence of polioviruses and other enteral viruses in sewage: a survey in the Czech Socialistic Republic 1969-1976. *J. Hyg. Epidemiol. Microbiol. Immunol.* 26, 1–14.

Conflict of Interest Statement: 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.

Received: 25 January 2012; paper pending published: 20 February 2012; accepted: 23 April 2012; published online: 21 May 2012.

Citation: Sasaki A, Haraguchi Y and Yoshida H (2012) Estimating the risk of re-emergence after stopping polio vaccination. Front. Microbio. ***3****:178. doi: 10.3389/fmicb.2012.00178*

This article was submitted to Frontiers in Virology, a specialty of Frontiers in Microbiology.

APPENDIX

DERIVATION OF EQ. 9

Here we derive Eq. 9 in the text. This is derived by noting that there may be i infected hosts in the next time step either if an infected host gives rise to $i-1$ secondary infections and itself remains infected, or if it gives rise to i secondary infections and itself dies or recovers. Thus

$$\begin{aligned} q(t) &= (1-\delta)\sum_{i=1}^{\infty}\frac{\lambda(t)^{i-1}}{(i-1)!}e^{-\lambda(t)}q(t+1)^i \\ &\quad + \delta\sum_{i=0}^{\infty}\frac{\lambda(t)^{i}}{i!}e^{-\lambda(t)}q(t+1)^i \\ &= \left[(1-\delta)q(t+1)+\delta\right]e^{-\lambda(t)(1-q(t+1))} \\ &\quad \times\sum_{j=0}^{\infty}\frac{\left\{\lambda(t)q(t+1)\right\}^j}{j!}e^{-\lambda(t)q(t+1)} \\ &= \left[(1-\delta)q(t+1)+\delta\right]e^{-\lambda(t)(1-q(t+1))} \end{aligned} \tag{A1}$$

with $\lambda(t)=\beta Kx(t)$, which then leads to (9) in the text.

APPROXIMATE TIME AND NUMBER OF INFECTEDS AT THE MINIMUM POINT

It is useful to obtain an explicit formula for the minimum number of infecteds and the time at which this number reaches its minimum in the deterministic trajectory. This clarifies the parameter dependence on the risk of re-emergence. We found the following approximation useful. We ignore the first term in the right hand of (8a), because it remains very small during the time interval from $t=0$ to $t=t_c$, to give

$$x(t) = 1-(1-x_0)e^{-ut}, \tag{A2}$$

(see, for example, Anderson and May, 1991). Integrating (8b) we have

$$w(t) = w_0 exp\left[\int_0^t\left[\beta x(s)-(u+\gamma)\right]ds\right]. \tag{A3}$$

Clearly $w(t)$ attains the local minimum when $t=t_c$ where $\beta x(t)=u+\gamma$. Letting

$$a=\frac{\beta-(u+\gamma)}{u}=k(R_0-1), \quad b=\frac{\beta(1-x_0)}{u}=kR_0(1-x_0), \tag{A4}$$

with $k=(u+\gamma)/u$ and $R_0=\beta/(u+\gamma)$, we therefore have

$$t_c \approx \frac{1}{u}log\left[\frac{b}{a}\right] = Llog\left[\frac{R_0(1-x_0)}{R_0-1}\right], \tag{A5a}$$

$$w_c \approx w_0\left(\frac{b}{a}\right)^a e^{a-b} = w_0\left(\frac{R_0(1-x_0)}{R_0-1}\right)^{k(R_0-1)} exp\left[R_0x_0-1\right], \tag{A5b}$$

where $L=1/u$ is the life expectancy, and $R_0=\beta/(u+\gamma)$ the basic reproductive rate. We expect a high probability of eradication if Kw_c is sufficiently smaller than 1, and show significant risk of re-emergence if it is 10 or more. The deviation of w_c from the true minimum is small in logarithmic scale, though it is as large as 50% in normal scale. However, for the purpose of quickly checking the likelihood of successful eradication, this formula is useful. If we assume that x_0 and w_0 take the values at the endemic equilibrium with the vaccination rate p (Eq. 5 in the text), we obtain the asymptotic formula for large R_0:

$$t_c \approx Lp/R_0, \quad (R_0 >> 1), \tag{A6a}$$

$$Kw_c \approx K\frac{D}{L}exp\left[-\frac{p^2}{2R_0}\frac{L}{D}\right], \quad (R_0 >> 1,\ L >> D), \tag{A6b}$$

where $D=1/\gamma$ is the mean duration of infection.