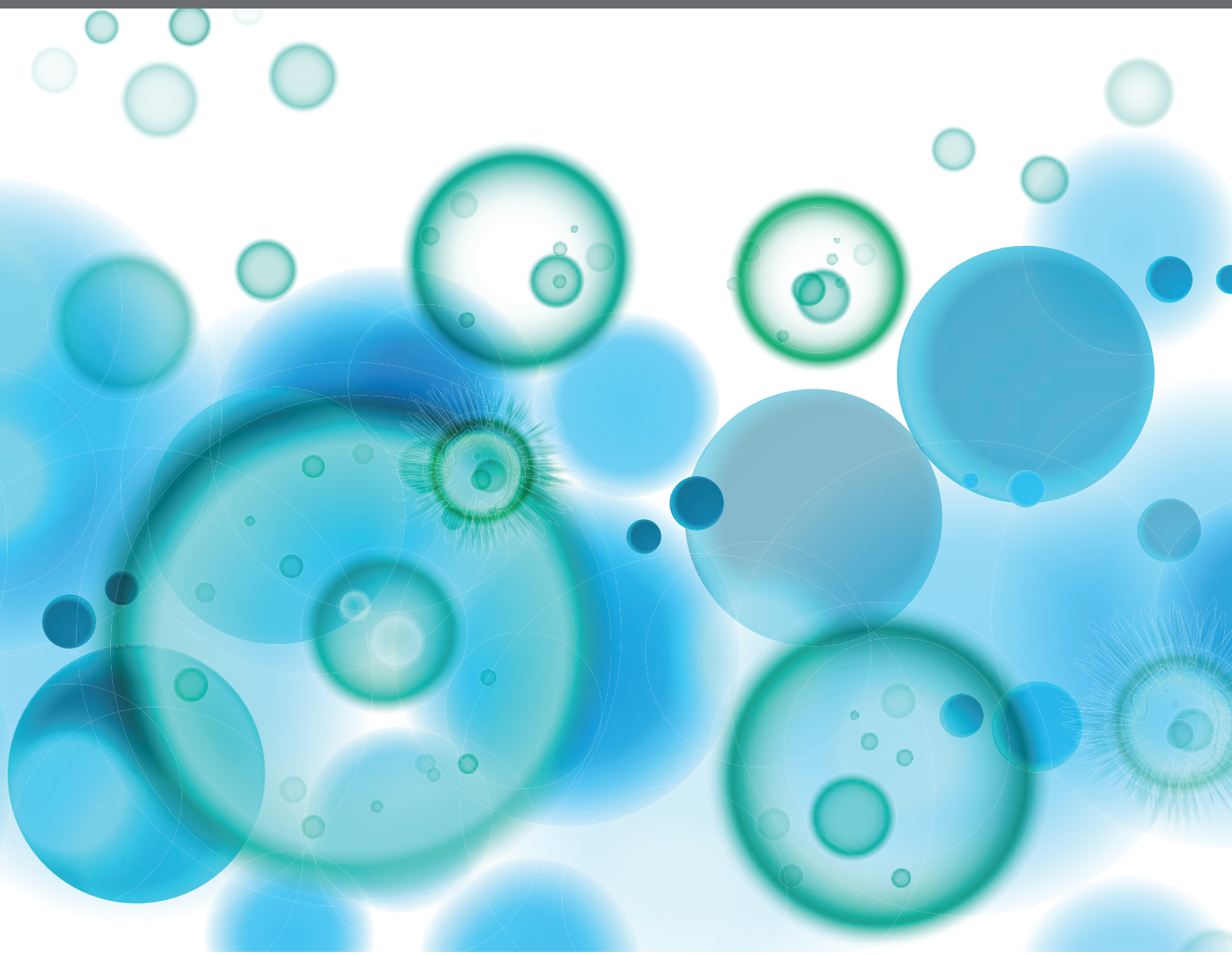
NEXT-GENERATION SEQUENCING OF HUMAN ANTIBODY REPERTOIRES FOR EXPLORING B-CELL LANDSCAPE, ANTIBODY DISCOVERY AND VACCINE DEVELOPMENT

EDITED BY: Jacob Glanville, Prabakaran Ponraj and Gregory C. Ippolito
PUBLISHED IN: Frontiers in Immunology



frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88963-951-9
DOI 10.3389/978-2-88963-951-9

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

NEXT-GENERATION SEQUENCING OF HUMAN ANTIBODY REPERTOIRES FOR EXPLORING B-CELL LANDSCAPE, ANTIBODY DISCOVERY AND VACCINE DEVELOPMENT

Topic Editors:
Jacob Glanville, Distributed Bio, United States
Prabakaran Ponraj, Sanofi (United States), United States
Gregory C. Ippolito, University of Texas at Austin, United States

Citation: Glanville, J., Ponraj, P., Ippolito, G. C., eds. (2020). Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-951-9

Table of Contents

frontiers in Immunology

EDITORIAL
published: 30 June 2020
doi: 10.3389/fimmu.2020.01344

Editorial: Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development

Ponraj Prabakaran[1]*, *Jacob Glanville*[2] and *Gregory C. Ippolito*[3]

[1] Biologics Research, Sanofi, Framingham, MA, United States, [2] Distributed Bio, South San Francisco, CA, United States, [3] Department of Molecular Biosciences, University of Texas at Austin, Austin, TX, United States

Keywords: B-cell receptor repertoire, next generation sequencing, antibodyome, immunoglobulin, B-cell clonotypes, vaccination, immunogenetics, immunoinformatics

OPEN ACCESS

Edited by:
Harry W. Schroeder, University of Alabama at Birmingham, United States

Reviewed by:
Paolo Casali, University of Texas Health Science Center San Antonio, United States
Nichol E. Holodick, Homer Stryker M.D. School of Medicine, Western Michigan University, United States

***Correspondence:**
Ponraj Prabakaran
prabakaran.ponraj@sanofi.com

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *09 March 2020*
Accepted: *27 May 2020*
Published: *30 June 2020*

Citation:
Prabakaran P, Glanville J and Ippolito GC (2020) Editorial: Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development. Front. Immunol. 11:1344. doi: 10.3389/fimmu.2020.01344

Editorial on the Research Topic

Next-Generation Sequencing of Human Antibody Repertoires for Exploring B-cell Landscape, Antibody Discovery and Vaccine Development

The next-generation sequencing (NGS) analysis of human antibody repertoires has enabled a heightened appreciation and comprehensive characterization of the B-cell receptor (BCR) landscape at an unprecedented resolution (1–4). This advance has expanded our insights and lent itself to numerous applications, including the following: NGS coupled with bioinformatics has enhanced phage biopanning of complex antibody libraries and facilitated the antibody discovery process (5); NGS analysis when coupled with large-scale computational structural modeling has revealed sequence and structural correlates between naive and antigen-experienced antibody repertoires (6); and in recent years, NGS-aided study of the antibodyome of HIV-1-infected individuals has increased our understanding of antibody responses and aided the design of antibody lineage-based immunogens that could, in principle, activate naïve precursor B cells to give rise to broadly-reactive neutralizing clones (7, 8). Thus, generally speaking, NGS of human antibody repertoires holds great promise for antibody discovery (9) and vaccine development (10, 11). This editorial introduces 17 high-quality research papers published in the Research Topic which summarize recent developments and applications within the context of NGS analysis of human antibody repertoires, through a combination of Original Research, Methodology, and Review articles.

The topic contains seven Original Research articles. These articles span a wide variety of topics. These studies illustrate means by which to harness the power of NGS for antibody discovery, B-cell immunogenetics, and the evolution of affinity maturation, as well as the investigation of antibody lineages in HIV-1/SIV infections. Hong et al. used cord blood samples from 10 newborn babies and peripheral blood from 33 healthy adults to perform an in-depth analysis of human neonatal and adult IgM heavy chain repertoires. Their comparative study revealed unexpectedly high levels of similarity between the neonatal and adult repertoires although antibody repertoire of healthy adults was more diverse than that of neonates. These results are helpful in understanding the antibody development and diversity in newborn babies and adults. Kirik et al. used NGS to analyze human bone marrow B cells to elucidate how different mutational paths are traversed by antibody lineages stemming from different germline gene origins both in

terms of amino acid substitutions, insertions, and deletions. Specifically, they identified germline gene-specific mutational patterns as found in selected and non-selected repertoires. These findings provide a framework for understanding patterns of evolution of antibodies arising from specific, defined germline genes. D'Angelo et al. showed that many different heavy-chain complementarity determining regions 3 (CDR-H3s) could be identified within a target-specific antibody population after *in vitro* selection by using a data set of 32,138 CDR-H3 sequences derived previously from the yeast display sorting and analysis of CDK2-specific antibodies. One of the remarkable observations demonstrated numerous rearranged heavy chains, derived from 19 different germline IGHV genes, were found to contain the same CDR-H3. Their main finding concludes that the same CDR-H3 can be generated by many different rearrangements, but that specific target binding is achieved by only certain unique V-D-J rearrangements and V_L pairing. Jeliazkov et al. determined the structural flexibility of the CDR-H3 loops, using previously published algorithms, for thousands of homology models of antibodies derived from the NGS data to find if affinity maturation reduces their conformational flexibility or not. They also used a total of 922 antibody crystal structures from the Protein Data Bank (12) and performed temperature factor analysis and molecular dynamic simulation to assess the flexibility. By using different computational approaches, they came with a conclusion that there is no significant difference between antibody CDR-H3 loop flexibility in repertoires of naïve and mature antibodies. However, they also noted inconsistent results across those methods for some antibodies. They concluded that further experimental methods, for example, hydrogen deuterium exchange mass spectrometry and more accurate modeling or structure determination of antibodies would resolve the inconsistencies. VanDuijn et al. profiled the immune repertoire of rats after immunization with purified antigens using NGS and proteomics. The data obtained from different analysis methods and experimental platforms demonstrate that the immunoglobulin repertoires of immunized animals have overlapping and converging features; however, the quantitative differences between the immune repertoires obtained using proteomic and NGS methods that might relate to differences between the biological niches could not be correlated in this study. With further improvement on the proteomic and NGS immune profiling approaches, their method may enable more interesting applications in biotechnology and clinical diagnostics. Then, He et al. and Han et al. combined the biopanning of scFv phage-displayed antibody libraries and 900 bp long-reads, enabling V_H/V_L paired NGS analysis. He et al. identified broadly neutralizing antibody intermediates from a HIV-1 patient, particularly PGT124 sub-lineage, possessing an invariable CDR-H3 loop and multiple library-derived intermediates, which might serve as a promising template for B-cell lineage vaccine design targeting. Han et al. also showed how they used long-read NGS combined with scFv phage display libraries for identifying SIV gp140-specific antibodies and analyzing their clonotypes and lineages correlating to neutralization activity.

Technical landscape for NGS analysis of human antibodies has changed tremendously and will continue toward the improvement of methods, immunoinformatics and data analysis tools. In this respect, we have four exciting articles devoted to methods/protocols. Hemadou et al. successfully developed, using the PacBio RS II system, and generated long reads (>800 bp) covering full length scFvs following *in vivo* panning in an animal model of atherosclerosis. They tested its performance by tracking and analysis of known, identical and related scFv-phage clone P3. Rosenfeld et al. and Vergani et al. present on a topic of bulk B-cells which provides a way for computationally assessing B-cell clone sizes and a library preparation method for NGS to capture an exhaustive full-length repertoire for nearly every sampled B-cell to be sequenced respectively. Rosenfeld et al. used three different measures of B cell clone size: copy numbers, instances and unique sequences, and then showed how these measures can be used to rank clones, analyze their diversity, and study their distribution within and between individuals. Overall, this method showed how different clone size measures can be used to study the clonal landscape in bulk B cell immune repertoire profiling data. On the other hand, the methodology as adopted by Vergani et al. serves as a useful protocol for Ig-seq where every IGHV-D-J rearrangement in the starting B-cell populations can be detected. Finally, advancements in NGS and error corrections have enabled antibody repertoire sequencing with single mutation precision but still compromising with sequencing accuracy. This opens the possibility for undocumented novel germline alleles. To address on this important issue, Wendel et al. present a method that can be quickly and easily applied to any antibody repertoire data set to mitigate the effects of germline mismatches on SHM patterns.

Next, we provide five excellent reviews in the Research Topic, starting with a review by Chaudhary and Wesemann, which provides a sound introduction to practical steps involved in the process of immune repertoire profiling including sample preparation, platforms available for NGS, sequencing data processing and annotations, and fundamental measurable features of the immune repertoire such as V/D/J gene-segment frequencies, CDR-H3 diversity and physicochemical properties, and immunoglobulin somatic hypermutation (SHM). They also highlight additional analyses using the NGS-derived repertoire data: isotype analysis, which offers insights into the effector biology mediated by heavy chain constant regions, such as complement fixation or binding to Fc receptors; clonal lineage analysis, which is used to trace clonal evolution of HIV-1 broadly neutralizing antibodies; and B-cell network analysis that can link mature antibody sequences to their germline precursor sequences. Extrapolation of these procedures for analyzing paired V_H:V_L repertoires was also discussed. The readers attracted to this review article will likely appreciate the detailed description of statistical tools and their features that can be used for analysis and interpretation of NGS big data sets, along with a comprehensive list of software tools available for sequence error correction, annotation, and evaluation of B cell repertoires. This is followed by a review in which Miho et al. discuss four computational strategies: (i) measuring immune repertoire diversity, (ii) clustering and network approaches to

resolve the sequence similarity architecture, (iii) phylogenetic methods to retrace antigen-driven evolution, and (iv) machine learning methods to dissect naïve and antigen-driven repertoire convergence. Furthermore, they summarize outstanding questions in computational immunology and propose new directions for systems immunology by possibly linking NGS-based potential metrics with computational discovery of immunotherapeutics, vaccines, and immunodiagnostics. These two reviews are followed by a mini-review article by Rouet et al., which specifically addresses the strategies for NGS of phage- and other antibody-display libraries, and list NGS platforms and analysis tools. This review also touches briefly on bioinformatic tools and applications to design validation with analyses of naïve antibody libraries, affinity maturation and epitope mapping with specific examples from literature. After these three reviews, our Research Topic addresses a challenging question of how B-cell receptor repertoire sequencing can potentially be enriched when coupled with structural antibody data, as described in the review by Kovaltsuk et al.. This review covers the basic principles about structural architecture of IgG, repertoire sequencing technologies and antibody structural properties. Further, they highlight on computational approaches and tools that leverage antibody structure information and provide a generalized workflow of antibody modeling. Overall, the authors illustrate how these two data types—NGS DNA sequences (i.e., BCR-seq) and atomic structures, that can enrich one another and yield potential for advancing our knowledge of the immune system and improving antibody engineering and developability. Along this line of work, Mishra and Mariuzza review the structural basis of antibody affinity maturation from NGS data. Interestingly, they looked at the studies of antibody affinity maturation prior to and after NGS. They further emphasized how important the NGS is for the reconstruction of antibody clonal lineages in immune responses to viral pathogens, such as HIV-1. They discussed in detail about various mechanisms of paratope preorganization, rigidification, reorientation, and indels as described for many antibodies. Overall, this review provides a more holistic perspective to structural basis of antibody affinity maturation from the point of next-generation sequencing.

To finish this topic, we aptly include a perspective article on reproducibility and reuse of adaptive immune receptor repertoire data. We are delighted to have included an excellent contribution from the Adaptive Immune Receptor Repertoire (AIRR) community (Breden et al.), which provides an overview of the founding principles and presents the progress it has made to develop and promote standards and recommendations for best practices and data-sharing protocols. In conclusion, NGS combined with innovative single-B-cell technologies has the potential to yield millions of native human antibody sequences and some of them that could match with therapeutic antibodies (13, 14). This suggests a possible implication for data mining in the NGS repositories for discovering therapeutic antibody candidates in future. Also, large-scale NGS analysis of individual antibodyome will lead to improved insights into overall diversity of the human antibody repertoire and B cell immunogenetics (15–17).

AUTHOR CONTRIBUTIONS

PP wrote the manuscript. All authors contributed to this work and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The editors thank all reviewers for their time and constructive feedback on submitted manuscripts. This Research Topic would not have been be possible without the support of the Frontiers in Immunology editorial team. We thank Prof. Thomas L. Rothstein for his helpful comments and support. PP thanks Dr. Partha Chowdhury and Dr. Maria Wendt for their support and encouragement. GI wishes to acknowledge his grant support during this period, including NIH grants AI135682 and AI119368, The William and Ella Owens Medical Research Foundation, and the PATH Malaria Vaccine Initiative.

REFERENCES

1. Dimitrov DS. Therapeutic antibodies, vaccines and antibodyomes. *MAbs.* (2010) 2:347–56. doi: 10.4161/mabs.2.3.11779
2. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol.* (2014) 32:158–68. doi: 10.1038/nbt.2782
3. Robinson WH. Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nat Rev Rheumatol.* (2015) 11:171–82. doi: 10.1038/nrrheum.2014.220
4. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci USA.* (2009) 106:20216–21. doi: 10.1073/pnas.0909775106
5. Vaisman-Mentesh A, Wine Y. Monitoring phage biopanning by next-generation sequencing. *Methods Mol Biol.* (2018) 1701:463–73. doi: 10.1007/978-1-4939-7447-4_26
6. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci USA.* (2016) 113:E2636–45. doi: 10.1073/pnas.1525510113
7. Jardine JG, Kulp DW, Havenar-Daughton C, Sarkar A, Briney B, Sok D, et al. HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science.* (2016) 351:1458–63. doi: 10.1126/science.aad9195
8. Havenar-Daughton C, Sarkar A, Kulp DW, Toy L, Hu X, Deresa I, et al. The human naive B cell repertoire contains distinct subclasses for a germline-targeting HIV-1 vaccine immunogen. *Sci Transl Med.* (2018) 10:448. doi: 10.1126/scitranslmed.aat0381
9. Naso MF, Lu J, Panavas T. Deep sequencing approaches to antibody discovery. *Curr Drug Discov Technol.* (2014) 11:85–95. doi: 10.2174/15701638113106660040
10. Kwong PD, Chuang GY, DeKosky BJ, Gindin T, Georgiev IS, Lemmin T, et al. Antibodyomics: bioinformatics technologies for understanding B-cell immunity to HIV-1. *Immunol Rev.* (2017) 275:108–28. doi: 10.1111/imr.12480

11. Prabakaran P, Zhu Z, Chen W, Gong R, Feng Y, Streaker E, et al. Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. *Front Microbiol.* (2012) 3:277. doi: 10.3389/fmicb.2012.00277
12. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* (2019) 47:D464–74. doi: 10.1093/nar/gky1004
13. Krawczyk K, Raybould MIJ, Kovaltsuk A, Deane CM. Looking for therapeutic antibodies in next-generation sequencing repositories. *MAbs.* (2019) 11:1197–205. doi: 10.1080/19420862.2019.1633884
14. Ponraj P. Next-generation sequencing may challenge antibody patent claims. *Nature.* (2018) 557:166. doi: 10.1038/d41586-018-05065-5
15. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature.* (2019) 566:393–7. doi: 10.1038/s41586-019-0879-y
16. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, et al. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature.* (2019) 566:398–402. doi: 10.1038/s41586-019-0934-8
17. Prabakaran P, Chowdhury PS. Landscape of non-canonical cysteines in human V_H repertoire revealed by immunogenetic analysis. *Cell Rep.* (2020) 30. doi: 10.1016/j.celrep.2020.107831

Conflict of Interest: PP is an employee of Sanofi Genzyme. JG is an employee and CEO of Distributed Bio.

The remaining 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.

The handling editor declared a past co-authorship with one of the authors GI.

Copyright © 2020 Prabakaran, Glanville and Ippolito. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 24 August 2017
doi: 10.3389/fimmu.2017.01025

Hidden Lineage Complexity of Glycan-Dependent HIV-1 Broadly Neutralizing Antibodies Uncovered by Digital Panning and Native-Like gp140 Trimer

Linling He[1], Xiaohe Lin[1], Natalia de Val[2,3,4], Karen L. Saye-Francisco[1,4], Colin J. Mann[1], Ryan Augst[1], Charles D. Morris[1], Parisa Azadnia[1], Bin Zhou[5], Devin Sok[1,3,4], Gabriel Ozorowski[2,3,4], Andrew B. Ward[2,3,4], Dennis R. Burton[1,3,4,6] and Jiang Zhu[1,2,4]**

[1] Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA, United States, [2] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, United States, [3] International AIDS Vaccine Initiative Neutralizing Antibody Center and the Collaboration for AIDS Vaccine Discovery, The Scripps Research Institute, La Jolla, CA, United States, [4] Scripps Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery, The Scripps Research Institute, La Jolla, CA, United States, [5] Department of Chemistry, The Scripps Research Institute, La Jolla, CA, United States, [6] Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, Cambridge, MA, United States

OPEN ACCESS

Edited by:
Prabakaran Ponraj,
Intrexon, United States

Reviewed by:
Sanjay Kumar Phogat,
Sanofi Pasteur, United States
Bin Su,
Beijing You'an Hospital and Capital Medical University, China
Neil S Greenspan,
Case Western Reserve University, United States

***Correspondence:**
Linling He
linling@scripps.edu;
Jiang Zhu
jiang@scripps.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: 23 June 2017
Accepted: 08 August 2017
Published: 24 August 2017

Citation:
He L, Lin X, de Val N, Saye-Francisco KL, Mann CJ, Augst R, Morris CD, Azadnia P, Zhou B, Sok D, Ozorowski G, Ward AB, Burton DR and Zhu J (2017) Hidden Lineage Complexity of Glycan-Dependent HIV-1 Broadly Neutralizing Antibodies Uncovered by Digital Panning and Native-Like gp140 Trimer. Front. Immunol. 8:1025. doi: 10.3389/fimmu.2017.01025

Germline precursors and intermediates of broadly neutralizing antibodies (bNAbs) are essential to the understanding of humoral response to HIV-1 infection and B-cell lineage vaccine design. Using a native-like gp140 trimer probe, we examined antibody libraries constructed from donor-17, the source of glycan-dependent PGT121-class bNAbs recognizing the N332 supersite on the HIV-1 envelope glycoprotein. To facilitate this analysis, a digital panning method was devised that combines biopanning of phage-displayed antibody libraries, 900 bp long-read next-generation sequencing, and heavy/light (H/L)-paired antibodyomics. In addition to single-chain variable fragments resembling the wild-type bNAbs, digital panning identified variants of PGT124 (a member of the PGT121 class) with a unique insertion in the heavy chain complementarity-determining region 1, as well as intermediates of PGT124 exhibiting notable affinity for the native-like trimer and broad HIV-1 neutralization. In a competition assay, these bNAb intermediates could effectively compete with mouse sera induced by a scaffolded BG505 gp140.681 trimer for the N332 supersite. Our study thus reveals previously unrecognized lineage complexity of the PGT121-class bNAbs and provides an array of library-derived bNAb intermediates for evaluation of immunogens containing the N332 supersite. Digital panning may prove to be a valuable tool in future studies of bNAb diversity and lineage development.

Keywords: antibody phage display, B-cell lineage development, broadly neutralizing antibodies, HIV-1 vaccine design, native-like trimer, next-generation sequencing

INTRODUCTION

Broadly neutralizing antibodies (bNAbs) isolated from a small fraction of infected individuals have provided valuable insights into the humoral response against HIV-1 (1–4). It has been proposed that bNAbs with structurally defined antigen interactions can be used as templates for designing immunogens capable of eliciting similar antibody responses upon vaccination (5–7). Considering the

extensive somatic hypermutation (SHM) and unusual sequence features of bNAbs, such as long complementarity-determining region (CDR) loops, an in-depth understanding of their ontogeny is imperative to designing sequential immunogens for guided antibody maturation (7). To this end, next-generation sequencing (NGS) has been utilized to explore details of the antibody repertoire and lineage development for bNAbs of vaccine interest (8–19). However, with the exception of a few cases (16, 17, 19–22), studies regarding early bNAb development continue to be restrained by limited sample availability and low frequency of lineage intermediates in memory B cells. Nonetheless, germline-reverted precursors and inferred lineage intermediates have been derived for several bNAbs targeting the CD4-binding site (CD4bs) and the N332 supersite near the base of variable loop 3 (V3) to facilitate immunogen design and *in vivo* evaluation (18, 23–31).

The HIV-1 envelope glycoprotein (Env) is covered with a dense layer of glycans. While this "glycan shield" poses barriers for bNAbs to access epitopes such as the CD4bs (17), it harbors key neutralizing antibody targets (32, 33) including the trimeric apex and the N332 supersite, of which the latter is a high-mannose patch centered around the N332 glycan (34, 35). The N332-dependent bNAb classes represented by PGT121, PGT128, and PGT135 have been extensively studied, showing an inherent promiscuity of the N332 supersite (36–42). NGS has revealed sequence diversity within the PGT121 and PGT135 families (9, 15, 18), with putative intermediates inferred for the former by phylogenetic analysis (18). The PGT121 class consists of bNAbs PGT121-123, PGT124/10-1074 (36), and PGT133-134, with PGT121 and 10-1074 demonstrating therapeutic potential in macaques and humans, respectively (43, 44). Structures of the PGT121-class bNAbs and their intermediates in complex with a modified gp120 core and BG505 SOSIP.664 gp140 trimer have placed PGT124 (and 10-1074 (36)) on a distinct evolutionary branch mainly focusing on the N332 glycan, whereas PGT121-123 recruited multiple glycans during lineage maturation (41, 42, 45). Recently, Steichen et al. designed a series of SOSIP.664 trimers with optimized binding to the PGT121 precursor and inferred intermediates (31), which induced bNAb-like responses in immunoglobulin (Ig) knock-in mice (23). This study provided a proof of concept for the B-cell lineage vaccine design targeting the N332 supersite (7). Furthermore, a ferritin nanoparticle displaying 24 copies of an N332 epitope-scaffold was reported to elicit a consistent N332-specific antibody response in BALB/c mice cross-reactive with a native-like trimer, which itself failed to induce such antibody response in immunization (46). Interestingly, once the membrane-proximal external region (MPER) and a C-terminal scaffold were included in this trimer construct, a robust antibody response to the apex was observed, indicating enhanced immune recognition of the glycan shield (46). Notwithstanding, several critical issues remain elusive in vaccine design targeting the N332 supersite: (i) whether native precursors and intermediates of the PGT121 class can be found within the donor repertoire; (ii) the minimal level of SHM required for the PGT121-class bNAbs to recognize an unmutated trimer; and (iii) whether a trimer immunogen with an intact glycan shield is capable of eliciting an N332-specific antibody response in animal immunization. A careful assessment of these issues is pertinent to future vaccine design efforts targeting the N332 glycan supersite.

In this study, we examined samples from donor-17, the source of PGT121-class bNAbs, to search for bNAb precursors and lineage intermediates. To facilitate this analysis, we developed a digital panning method by combining 900 bp NGS and H/L-paired antibodyomics with phage display of single-chain variable fragments (scFvs) to probe the donor repertoire. A biotinylated Avi-tagged BG505 gp140 trimer containing an optimized heptad repeat 1 (HR1) bend (47) was utilized as antigenic bait for antibody library screening. Although scFvs identified from this library may not possess authentic heavy and light chains, they nonetheless provide a glimpse into Env recognition by the diverse antibody repertoire that gave rise to the PGT121 lineage. Digital panning of a diverse scFv library identified bNAb-like clones with increased affinity for the native-like trimer, and PGT124 variants with a unique 2-aa insertion in the heavy chain complementarity-determining region 1 (HCDR1) loop. Similar sequences were also found in the donor antibody repertoire by deep sequencing, albeit with low frequency. A focused scFv library was then constructed using the PGT121 class-specific primers and subjected to digital panning against the trimer probe, revealing heavy and light chain (HC and LC) intermediates of the PGT121 class that differed notably from the inferred sequences (18). All library-derived antibody clones were assessed for antigen binding and HIV-1 neutralization. The utility of selected bNAb intermediates was further demonstrated using mouse sera from a previous trimer immunization (46). Serum analysis indicated that a scaffolded BG505 gp140.681 trimer induced consistent antibody responses to the apex and the N332 supersite, the latter of which could be blocked effectively by mature PGT124 and partially by a near-germline HC intermediate. Collectively, our study uncovers previously unrecognized lineage complexity of the PGT121-class bNAbs and presents a set of functional intermediates potentially valuable for the rational design and evaluation of HIV-1 immunogens containing the N332 supersite.

MATERIALS AND METHODS

Human Specimen

Peripheral blood mononuclear cells (PBMCs) from an HIV-1 infected donor (donor-17) (34) of the Protocol G cohort were used for scFv library construction and antibody repertoire sequencing.

HIV-1 Panning Antigens

The Avi-tagged BG505 gp140 trimer containing a redesigned HR1 bend (47) and the clade-C (ZM109) V1V2 nanoparticle (48) were transiently expressed in HEK293 F cells and in *N*-acetylglucosaminyltransferase I-negative (GnTI$^{-/-}$) HEK293 S cells (Life Technologies), respectively. In brief, HEK293 F/S cells were thawed and incubated with FreeStyle™ 293 Expression Medium (Life Technologies) in a Shaker incubator at 37°C, with 120 rpm and 8% CO_2. When cells reached a density of 2.0×10^6/ml, expression medium was added to reduce cell density to 1.0×10^6/ml for transfection with polyethylenimine (PEI-MAX) (Polysciences). For 1-l transfection of the gp140 trimer, 800 μg of plasmid, 300 μg of furin plasmid, and 300 μg of pAdVAntage were mixed in 25 ml of Opti-MEM transfection medium (Life

Technologies) and added to 25 ml of Opti-MEM with 5 ml of PEI-MAX (1.0 mg/ml). For 1-l transfection of the V1V2 nanoparticle, 900 μg of plasmid was added to 25 ml of Opti-MEM and then mixed with 5 ml of PEI-MAX in 25 ml of Opti-MEM. After incubation for 30 min, the DNA-PEI-MAX complex was added to the cells. Culture supernatants were harvested 5 days after transfection, clarified by centrifugation at 1,800 rpm for 22 min, and filtered using 0.45 μm filters (Millipore). A *Galanthus nivalis* lectin (GNL) column (Vector Labs) was used to extract HIV-1 antigens from the supernatants and eluted with PBS containing 500 mM NaCl and 1 M methyl-α-D-mannopyranoside. For the Avi-tagged gp140 trimer, biotinylation was performed using the BirA biotin-protein ligase standard reaction kit (BirA-500) following the manufacturer's instructions (Avidity). The gp140 trimer and the V1V2 nanoparticle were then purified using size-exclusion chromatography (SEC) on a HiLoad 16/600 Superdex 200 PG column and a Superose 6 10/300 GL column (GE Healthcare), respectively.

Negative-Stain Electron Microscopy (EM)

The biotinylated Avi-tagged BG505 gp140 trimer, termed gp140.664.R1-Avi-Biot, was analyzed by negative-stain EM using a previously published protocol (47). Briefly, images were acquired with a Tietz 4 k × 4 k TemCam-F416 CMOS camera using a nominal defocus of 1,000 nm and the Leginon package (49) with an electron dose of ~29 e^{-}/Å^2. For image data processing, the Appion software package (50) was used to pick up particles and to make a stack. 2D classes were obtained using iterative multivariate statistical analysis (MSA)/multireference alignment (MRA) (51). To assess the quality of the trimers (native-like closed and open, or non-native), the reference-free 2D class averages were examined by eye using the same metrics as previously described (47). The 3D reconstruction of the gp140.664.R1-Avi-Biot trimer was obtained from the refinement of 17,783 particles using EMAN2 (52). The crystal structure of the BG505 SOSIP trimer (PDB ID: 4TVP) was fitted into the EM density and refined by using the UCSF Chimera "Fit in map" function (53).

Biolayer Interferometry (BLI)

Antibody-binding kinetics of the biotinylated gp140 trimer was assessed using an Octet RED96 instrument (fortéBio) as previously described (47). All assays were performed with agitation set to 1,000 rpm in fortéBio 1× kinetic buffer. The final volume for all the solutions was 200 μl/well. Assays were performed at 30°C in solid black 96-well plates (Geiger Bio-One). 5 μg/ml of protein in 1× kinetic buffer was used to load the HIV-1 antibody on the surface of anti-human Fc Capture Biosensors (AHC) for 300 s. Typical capture levels were between 0.5 and 1 nm and variability within a row of eight tips did not exceed 0.1 nm. A 60-s biosensor baseline step was applied prior to the analysis of the association of the antibody on the biosensor to the trimer in solution for 200 s. A twofold concentration gradient of trimer starting at a maximum of 200 nM was used in a titration series of six. The dissociation of the interaction was followed for 300 s. Correction of baseline drift was performed by subtracting the mean value of shifts recorded for a sensor loaded with antibody but not incubated with trimer and for a sensor without antibody but incubated with trimer. Octet data were processed by fortéBio's data acquisition software v.8.1. For apex-directed bNAbs, experimental data were fitted with the binding equations describing a 1:1 interaction, whereas for other bNAbs, the binding equations describing a 2:1 interaction were utilized to obtain the optimal fitting. K_D values were determined using the estimated response at equilibrium for each trimer concentration rather than the k_{on} and k_{off} values.

Antibody Phage Display

The construction of scFv libraries was performed using a protocol modified from a previously described method (54). Briefly, total RNA was extracted from ~10 million PBMCs for single-stranded cDNA synthesis using the SuperScript™ III system (Life Technologies) with random hexamer and oligo(dT)12–18 primers. Antibody HC and LC variable regions were obtained from a primary polymerase chain reaction (PCR) with mixed HuJ reverse primers and separate HuV forward primers, including a set of forward primers designed to capture PGT121-class light chains containing framework region 1 (FR1) deletions (Table S1 in Supplementary Material). To generate HC-LC fragments, overlap PCR was performed in 25 × 50 μl reactions (10 cycles) with 50 ng of gel-purified HC and 50 ng of gel-purified LC (λ only or equal amounts of κ and λ LC) without primer. To obtain full-length scFv inserts, PCR was performed in 50 × 50 μl reactions (15 cycles) with S*fi*I-F and S*fi*I-R primers (Table S1 in Supplementary Material) using 100 ng of gel-purified HC-LC as template. The resulting scFv inserts and the phagemid vector, pAdL™-20c (Antibody Design Labs), were digested with S*fi*I and gel-purified. The scFv DNA (256 ng) was ligated into the phagemid vector (400 ng) using the T4 Ligase Kit (New England BioLabs) in 25 × 40 μl reactions at 16°C overnight. Purified phagemids were electroporated into competent TG1 cells (Lucigen) with the MicroPulser™ system (Bio-Rad). Specifically, 1 μl of phagemids and 25 μl of competent cells were placed into a 0.1-cm cuvette for electroporation using the pre-set program at 1.8 kV. The transformed bacteria were spread on 2YT agar plates supplied with 100 μg/ml carbenicillin and 2% (w/v) glucose, which were incubated at 37°C overnight. Bacteria were then scraped from the plates for phage culture with the helper phage, CM13 (Antibody Design Labs), and biopanning. To facilitate NGS analysis, 3 ml of the transformed bacteria was grown in 50 ml 2YT-Carb-Glu medium at 37°C for 2 h, with the plasmids extracted using the Plasmid Midi Kit (Qiagen). Prior to biopanning, HIV-1 antigens were conjugated to magnetic beads following the manufacturer's instructions (Invitrogen), with Dynabeads™ M-280 Streptavidin beads used for biotinylated gp140 trimers and Dynabeads™ M-270 Epoxy beads used for V1V2 nanoparticles. Four biopanning cycles were performed, with 6–8 wash steps in each cycle to remove phage that did not recognize the antigen. Plasmids were extracted from 3 ml of the bacteria after each cycle for subsequent NGS analysis of scFv libraries.

Next-Generation Sequencing

Next-generation sequencing was performed on the Ion Torrent Personal Genome Machine (PGM) and S5 systems. The scFv-coding regions were amplified from the plasmid stock using PCR with fp1-S*fi*I-F and A-S*fi*I-[Barcode]-R primers (Table S1 in Supplementary Material). Of note, the forward primer (fp1-S*fi*I-F) contained a PGM full-length P1 (fP1) adaptor, whereas the

reverse primer (A-S*fi*I-[Barcode]-R) contained a PGM A adaptor and an Ion Xpress™ barcode (Life Technologies) to differentiate each scFv library. A total of 25 PCR cycles were performed and the PCR products with an expected length of 800–900 bp were gel-purified (Qiagen). The procedure used for PGM sequencing has been described previously (20). Briefly, the libraries were quantitated using Qubit® 2.0 Fluorometer with Qubit® dsDNA HS Assay Kit. The dilution factor required for PGM template preparation was determined such that the final concentration was 50 pM. Template preparation was performed with the Ion PGM Template IA 500 Kit. Long-read sequencing was performed on the Ion Torrent PGM sequencer with the Ion PGM™ Hi-Q™ Sequencing Kit using an Ion 314 v2 chip for a total of 1,500 nucleotide flows. Raw sequencing data were processed without the 3′-end trimming in base calling to obtain full-length scFvs. The donor-17 HC library was generated using a 5′-RACE PCR protocol as previously described (17). Template preparation and (Ion 530) chip loading were performed on Ion Chef using the Ion 530 Ext Kit, followed by sequencing on the Ion S5 system with default settings. The raw NGS data can be found in the NCBI Sequence Read Archive with the accession number SRP105512.

Bioinformatics Analysis of scFv Libraries

The human *Antibodyomics* 1.0 pipeline (8, 15, 17, 55) has been adapted to analyze the 900 bp sequencing data of scFv libraries. Following the housekeeping step (assigning a unique index to each scFv) and light-chain germline gene assignment, each scFv sequence was divided into HC and LC by matching a 15-aa $(G_4S)_3$ linker connecting the two chains. The HC and LC datasets were then processed separately by the chain-specific antibodyomics pipelines. Finally, full-length HC and LC from the same scFv were re-matched according to their indices in the sequenced scFv library (scFv indices) and deposited into correlated HC and LC databases for more in-depth analysis.

The *Antibodyomics* 1.0 pipeline consists of the following five steps (8, 15, 17): (1) data reformatting and cleaning; (2) germline gene assignment; (3) sequencing error correction; (4) calculation of sequence identity to a set of known antibodies; and (5) determination of CDR3 sequences and variable domain boundaries. In this study, a number of changes have been made to improve the pipeline accuracy and efficiency. In step 2, the original pipeline assigned variable (V), diversity (D), and joining (J) genes sequentially for heavy chains, whereas the modified pipeline first assigned V and J genes and then determined an appropriate D gene for the region between the V and J gene segments. In step 3, BLASTn (56) was replaced with LALIGN (57) to generate pairwise local alignment, as BLASTn often outputs a partial alignment lacking the N-terminus when sequencing errors occur in this region. By contrast, LALIGN can generate a more complete alignment with the assigned germline V gene, thus enabling error correction for the N-terminus and proper translation to the protein sequence. Of note, two consecutive deletions or insertions separated by a single nucleotide may produce a merged gap in alignment due to a lower gap penalty. Such homopolymer errors can now be detected and corrected by the modified pipeline. In step 5, CLUSTALW2 (58) was replaced by MUSCLE (59) to generate multiple sequence alignment, with a reduction of computational time by threefold to fourfold. The modified pipeline (version 2.0) has been validated using the 454 sequencing data reported for donor-17 (18). The *Antibodyomics* 2.0 pipeline can be obtained upon request to the authors.

H/L-Paired, CDR3-Based Clustering Analysis

A novel method has been devised to determine non-redundant scFv clones and their frequencies within a converged phage library to facilitate antibody selection. This method is hierarchical, as it identifies the unique HCDR3 lineages first, and subsequently all the unique LCDR3 lineages associated with each HCDR3 lineage, resulting in a list of scFv clones each characterized by a distinct HCDR3-LCDR3 pair. In the first step, the unique HCDR3 lineages can be identified as follows: (1) HCDR3 loops are extracted from HC sequences and clustered into "groups" using CD-HIT (60) with an identity cutoff of 95% and a criterion of ≤1 mismatch within the core alignment; (2) a consensus HCDR3 is derived for each HCDR3 group based on the multiple sequence alignment by MUSCLE; and (3) HCDR3 groups are merged into unique HCDR3 "lineages" by BLASTClust (56) with a sequence identity of 95% or greater covering 95% of the sequence length. For each HCDR3 lineage, a similar procedure can be applied to the matching LCs to determine the unique LCDR3 lineages. The resulting scFv clones will be ranked by their frequencies (the number of scFvs possessing a pair of unique HCDR3 and LCDR3), with the HC and LC sequence files pertaining to each scFv clone provided as the output. For large scFv clonal families (>500 members), which potentially represent high-affinity binders in the enriched library, an additional clustering analysis is performed to derive high-quality consensus sequences for antibody synthesis and functional validation. In brief, the H/LCDR3 sequences divisible by 3 will be clustered into groups using CD-HIT with an identity cutoff of 95%, no mismatch within the core alignment and identical sequence lengths. The largest group is then manually inspected and subjected to the consensus calculation by MUSCLE using no more than 4,000 randomly selected sequences. For small scFv clonal families (≤500 members), all sequences in the dataset are used to derive consensus HC and LC. The program used to perform H/L-paired, CDR3-based clustering analyses can be obtained upon request to the authors.

Enzyme-Linked Immunosorbent Assay (ELISA)

Each well of a Costar™ 96-well assay plate (Corning) was first coated with 50 µl PBS containing 0.2 µg of the appropriate antigens. The plates were incubated overnight at 4°C and, then, washed five times with wash buffer containing PBS and 0.05% (v/v) Tween 20. Each well was then coated with 150 µl of a blocking buffer consisting of PBS, 20 mg/ml blotting-grade blocker (Bio-Rad), and 5% (v/v) FBS. The plates were incubated with the blocking buffer for 1 h at room temperature and, then, washed five times with wash buffer. Wild-type (WT) PGT121-class bNAbs and antibodies derived from the donor-17 scFv library were diluted in the blocking buffer to a maximum concentration of 1 or 10 µg/ml, followed by a 10-fold dilution series. For each antibody dilution, a total of 50 µl volume was added to the appropriate wells.

Each plate was incubated for 1 h at room temperature and, then, washed five times with wash buffer. A 1:5,000 dilution of goat anti-human IgG antibody (Jackson ImmunoResearch Laboratories, Inc.) was then made in the wash buffer, with 50 µl of this diluted secondary antibody added to each well. The plates were incubated with the secondary antibody for 1 h at room temperature and, then, washed five times with wash buffer. Finally, the wells were developed with 50 µl of TMB (Life Sciences) for 3–5 min before stopping the reaction with 50 µl of 2 N sulfuric acid. The resulting plate readouts were measured at a wavelength of 450 nm.

HIV-1 Neutralization

Neutralization assays were performed on TZM-bl reporter cells using a panel of six tier-2 isolates and two tier-1 isolates. Neutralization curves were fit by a non-linear regression analysis using a 5-parameter hill slope equation. The 50% inhibitory concentration (IC_{50}) is defined as the antibody concentration required for inhibiting HIV-1 infection by 50%.

Serum Binding and Competition ELISA

Mouse antisera from a previous trimer immunization were tested against a native-like BG505 gp140 trimer (gp140.664.R1), a V1V2 nanoparticle (V1V2-FR), and an N332 nanoparticle (1GUT_A_ES-FR) by ELISA following a previously described protocol (46). Competition ELISA was performed to measure the binding of mouse antisera elicited by a scaffolded gp140 trimer to an N332 nanoparticle in the presence of WT bNAbs or native intermediates (NINs) derived from a focused donor-17 library. A slightly modified protocol was used. Briefly, plates were coated with purified N332 nanoparticles at 0.2 µg per well and incubated overnight at 4°C. After blocking, 10–50 µg/ml of a bNAb variant was added to the plates at 50 µl per well for 1 h incubation. The mouse antisera were initially diluted by a factor of 5 in blocking buffer, followed by a 10-fold dilution series. A 50 µl volume of each dilution was then added to the wells without washing the plates. After incubation for 1 h and five washes, a 1:2,000 dilution of goat anti-mouse IgG antibody was added to each well at 50 µl per well, followed by development with 50 µl of TMB, stop with 50 µl of 2 N sulfuric acid, and measurement at 450 nm.

RESULTS

A Native-Like gp140 Trimer Probe for Identification of Env-Specific Antibodies

The HIV-1 Env spike, a trimer of gp120 and gp41 heterodimers, is the only target of neutralizing humoral immune response (61). Trimer-based HIV-1 vaccine design has long been hampered by the metastable nature of the Env. A cleaved, soluble BG505 SOSIP.664 gp140 trimer was recently developed that closely mimics the functional Env spike in the stable, prefusion conformation (62–68). The BG505 SOSIP.664 trimer has enabled high-resolution structural analysis of the Env spike in complex with many bNAbs by crystallography and EM (42, 69–75). Alternative trimer platforms have also been developed, including the native, flexibly linked trimer (76–78), the single-chain gp140 (sc-gp140) trimer (79), and the uncleaved, prefusion-optimized (UFO) trimer (47). In particular, the UFO design has demonstrated greater trimer yield and purity for diverse HIV-1 strains in comparison to the SOSIP design (47). Furthermore, a cleaved version of the UFO trimer has been displayed on various nanoparticles as multivalent immunogens (48).

Here, we designed a cleaved BG505 trimer probe for Env-specific antibody identification, which contains a redesigned HR1 bend—the basis of the UFO design (47)—and a C-terminal Avi-tag. This trimer probe was expressed transiently in HEK293 F cells with co-transfected furin as previously described (47). The secreted Env protein was purified using a GNL column followed by *in vitro* biotinylation and SEC on a Superdex 200 16/600 column. We first compared the SEC profiles of the tagged and untagged trimers based on the ultraviolet absorbance at 280 nm (**Figure 1A**). The biotinylated Avi-tagged trimer probe, termed gp140.664.R1-Avi-Biot, displayed high yield and purity on par with the untagged trimer. We then analyzed the trimer probe by negative-stain EM, which yielded a 3D reconstruction consistent with the previously reported structures of HR1-redesigned trimers and UFO trimers (**Figure 1B**; Figure S1 in Supplementary Material) (47). The unoccupied EM densities extending from the C-termini of the docked gp140 trimer model may correspond to the biotinylated Avi-tag. Using BLI, we assessed the antigenicity of the trimer probe against a panel of bNAbs and non-NAbs. Three bNAbs isolated from donor-17 (PGT121, 124, and 133) (34) were used to assess the recognition of the N332 supersite by the PGT121-class bNAbs and to establish a baseline for comparison with forthcoming antibodies identified from this donor (**Figure 1C**). Octet binding revealed similar kinetics for the trimer probe and the untagged trimer, showing fast on-rates and flat dissociation curves. Consistently, nearly identical kinetic profiles were observed for bNAbs targeting other conserved epitopes, including the apex recognized by PGDM1400 (80) and PG16 (35), the CD4bs by VRC01 (81), and the gp120-gp41 interface by PGT151 (82) and 35O22 (83) (**Figure 1D**). Five non-NAbs were utilized to assess the exposure of non-neutralizing epitopes on the Env surface (**Figure 1E**). While the trimer probe appeared to shield non-neutralizing epitopes to the same degree as the untagged parent trimer, both outperformed the SOSIP trimer (47).

Our results thus demonstrated that the addition of an Avi-tag and *in vitro* biotinylation had no adverse effect on trimer purity, structural integrity, or antigenicity. In recent studies, the BG505 SOSIP.664 trimer was used as a bait to isolate bNAbs from HIV-1 patient samples (84, 85). As one of the well-characterized trimer platforms (86, 87), the UFO trimer may provide an alternative probe to identify bNAbs of diverse specificities from antibody libraries or Env-specific B cells.

Long-Read (900 bp) NGS and H/L-Paired Antibodyomics Enable Digital Panning

Phage display (88, 89) has been widely used to produce monoclonal antibodies (mAbs) for research and therapeutic purposes (90–94). While the first HIV-1 bNAbs were isolated from phage libraries (95), single-cell approaches have contributed most of the new bNAbs (96). Nonetheless, both scFvs and fragment antigen binding regions can be displayed on the phage surface and subjected to a selection process known as biopanning. Although NGS has been used to estimate the diversity of antibody phage

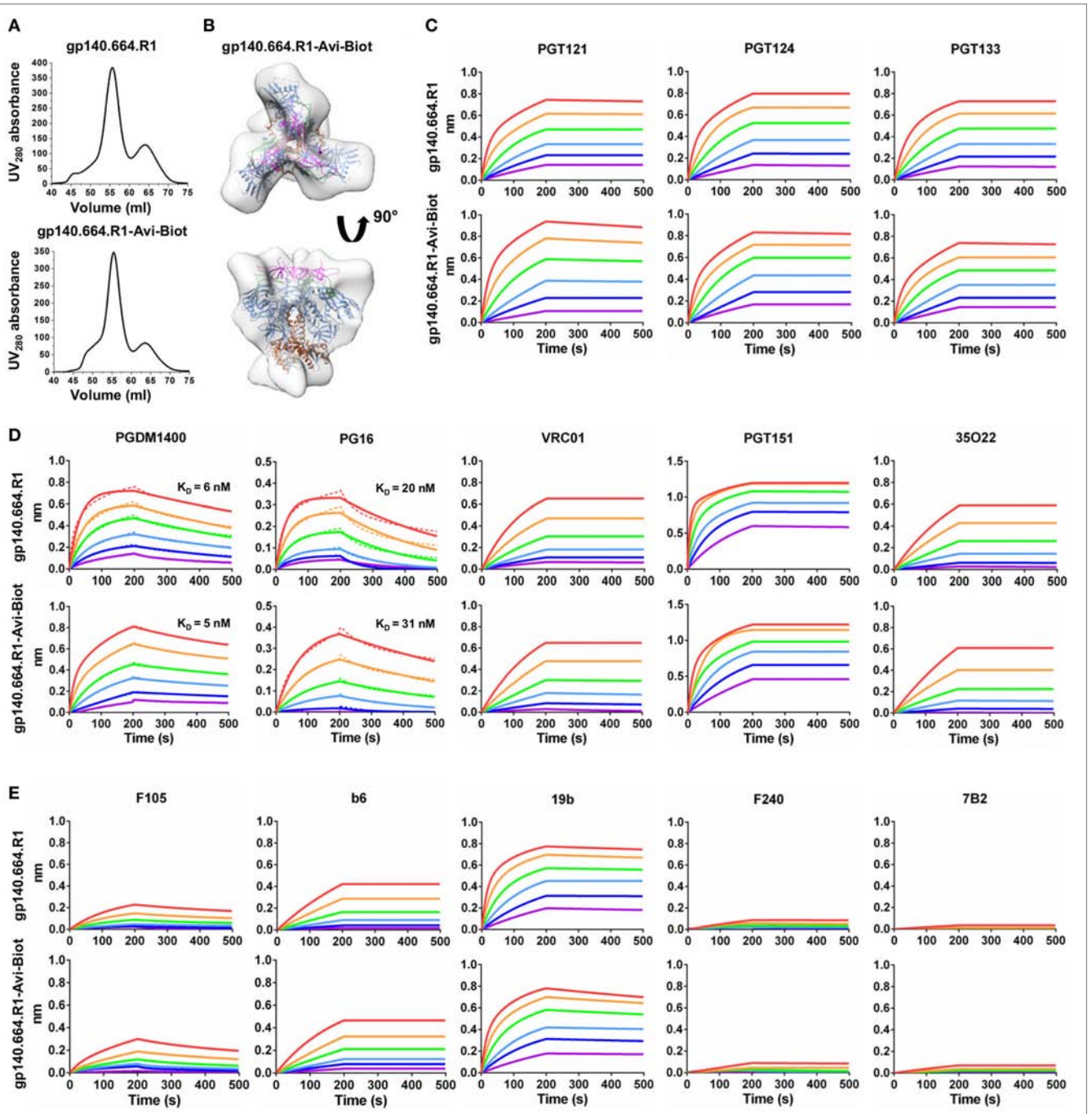


FIGURE 1 | Biophysical characterization of a native-like, prefusion-optimized trimer probe. This trimer probe was designed based on a BG505 gp140 construct containing a redesigned heptad repeat 1 bend (termed gp140.664.R1). **(A)** Size-exclusion chromatography (SEC) profiles of 293F-expressed, *Galanthus nivalis* lectin (GNL)-purified gp140.664.R1 trimer (top) and biotinylated, Avi-tagged gp140.664.R1 trimer probe, termed gp140.664.R1-Avi-Biot (bottom), from a Superdex 200 16/600 column. **(B)** 3D reconstruction of the gp140.664.R1-Avi-Biot trimer probe derived from negative-stain electron microscopy (EM). The trimer densities are shown in gray transparent surface with the fitted crystal structure of the SOSIP trimer (PDB ID: 4TVP, gp120 in blue with V1V2 in magenta, V3 in green, and gp41 in brown). **(C–E)** Antigenic profiles of the gp140.664.R1 and gp140.664.R1-Avi-Biot trimers measured for **(C)** three broadly neutralizing antibodies (bNAbs) of the PGT121 class (PGT121, 124, and 133) isolated from donor-17, **(D)** five bNAbs targeting other sites of vulnerability on the HIV-1 envelope glycoprotein, and **(E)** five representative non-NAbs. Sensorgrams were obtained from an Octet RED96 instrument using a trimer titration series of six concentrations (200–6.25 nM by twofold dilution). K_D values calculated from 1:1 global fitting are labeled for V1V2 apex-directed bNAbs (PGDM1400 and PG16) in panel **(D)**.

libraries (97, 98), its broader application has been restricted by the insufficient read length to cover both HC and LC in a scFv sequence (>800 bp). It was recently proposed that once this technical barrier is overcome, NGS can be used to directly select functional mAbs from a scFv library (99, 100), rendering a quantitative method for antibody discovery.

Here, we addressed this technical challenge in a case study of the PGT121 lineage. We first characterized three types of antibody libraries constructed from PBMCs of donor-17 using gel electrophoresis (**Figure 2A**). Two antibody libraries were generated using different PCR methods—multiplex PCR (with gene-specific primers) and 5′-RACE PCR (with single 3′-reverse primers)—and showed bands around 500 and 600 bp, respectively (15). By contrast, the scFv library generated using H/L-overlapping PCR and a large primer set (Table S1 in Supplementary Material), as well as four post-panning libraries, yielded distinctive bands around 900 bp. This comparison highlights the difficulty in sequencing full-length scFv libraries. Previously, we demonstrated an extended read length of 600–700 bp in the unbiased NGS analysis of antibody repertoires using the Ion Torrent PGM and a modified protocol with 1,200 nucleotide flows (15, 17, 46, 101). Can this NGS platform be adapted for sequencing scFvs? To test this possibility, we sequenced the donor-17 scFv libraries on the PGM using an Ion 314v2 chip and 1,500 nucleotide flows (**Figure 2B**). Markedly, over one million raw reads in the range of 750–950 bp were obtained for the scFv libraries, compared with 450 and 550 bp for the mixed HC/LC libraries generated by two different PCR methods.

We then combined 900 bp long-read NGS and H/L-paired antibodyomics with scFv library panning into a coherent strategy, termed digital panning, for identification of functional mAbs (**Figure 2C**). By design, this method is both analytical and deterministic, as it can capture the full-length scFvs during the panning process and select representative scFv clones based on their frequency and antibody characteristics. The previously developed *Antibodyomics* pipeline (55) was modified to facilitate *in silico* analysis of the sequenced scFv libraries (**Figure 2D**). Briefly, each scFv is assigned a unique index and divided into HC and LC by matching a 15-aa linker between the two chains. Following the chain-specific pipeline processing, HC and LC from the same scFv are identified based on their shared scFv indices, resulting in correlated HC and LC databases. Of note, the method used in the *Antibodyomics* pipeline for indel error correction (9) has been modified to achieve greater accuracy, as demonstrated for the 454 sequencing data from donor-17 (18) (Figure S2 in Supplementary Material). The scFv-derived HC and LC databases can then be analyzed in-depth using bioinformatics tools previously developed for repertoire profiling and lineage tracing (8–12, 15, 17). Further implementation of an H/L-paired, CDR3-based clustering method allows determination of representative scFv clones and their respective frequencies for facilitating mAb selection.

Digital Panning of a Diverse Donor-17 ScFv Library Identifies a New Lineage Variant

As described earlier, a diverse donor-17 scFv library has been constructed using primers covering all HC and LC germline gene families (Table S1 in Supplementary Material) and screened against a native-like trimer probe (**Figure 1**). The pre-panning scFv library and four post-panning scFv libraries were pooled at a ratio of 3:1:1:1:1 for 900 bp deep sequencing on the PGM (Table S2A in Supplementary Material). After data processing with the H/L-paired *Antibodyomics* pipeline (**Figure 2D**), the scFv-derived HC and LC databases were analyzed to generate library profiles and to select Env-specific mAb clones.

Quantitative library profiles displayed a distinct pattern of antibody enrichment and a rapid convergence after two panning steps. In terms of germline gene usage (**Figure 3A**, column 1), the IgHV4 and IgLV3 gene families accounted for >96% of the library upon convergence, with a 6–12-fold increase in frequency with respect to the pre-panning library. Further analysis revealed the prevalence of IgHV4-59 (~60%) and IgHV4-61 (~30%) within the IgHV4 family and IgLV3-21 (~96%) within the IgLV3 family. In contrast, the κ LCs exhibited a less discernible pattern of germline gene usage and a reduction to less than 100 sequences in the converged scFv library. A significant shift was observed in the SHM distribution (**Figure 3A**, column 2), with the average value increasing from 11% and 6% to 24% and 27% for HC and LC, respectively. Furthermore, nearly 90% of the HC sequences possessed long HCDR3 loops of 23–25 aa (**Figure 3A**, column 3). Overall, this scFv library appeared to have converged to the PGT121-class bNAbs, which are characterized by a specific germline gene usage (IgHV4-59 and IgLV3-21), high degree of SHM (HC: 19–22% and LC: 21–29%), and a long HCDR loop (24 aa). A two-dimensional (2D) identity-divergence analysis was then performed to visualize the scFv-derived HC and LC populations during the trimer panning process (**Figure 3B**). Indeed, the 2D plots revealed rapid enrichment of PGT124-like HCs and PGT124/PGT133-like LCs, which was further confirmed by the H/L-paired, CDR3-based clustering analysis (**Figure 3C**). While the most prevalent scFv family was characterized by PGT124-like HCDR3 loops, ~30% of HCs in this family were assigned to IgHV4-61 instead of IgHV4-59 due to a 2-aa insertion in HCDR1. Of note, the PGT133-like LCs appeared to be more prevalent than the PGT124-like LCs in the converged library, showing ~50-fold difference in their frequencies. In addition, a small group of 45 scFvs possessed a 16-aa HCDR3 loop partially matching the C-terminal portion of the PGT124 HCDR3 sequence. Taken together, digital panning of a diverse donor-17 scFv library resulted in a panel of mAbs including predominant bNAb-like clones and mAbs of unknown specificities.

The six most prevalent mAbs (Ab_{d17}-1–6) in the H/L-paired clustering analysis (**Figure 3C**) were synthesized for functional validation (Table S3A in Supplementary Material). Antigen binding was initially assessed by ELISA against a native-like trimer (gp140.664.R1) (47), a gp120-ferritin (gp120-FR) nanoparticle (48), and an N332 nanoparticle (1GUT_A_ES-FR) (46), with a V1V2-ferritin nanoparticle (V1V2-FR) (48) included as a negative control. Three representative bNAbs for the PGT121 class demonstrated differential binding profiles: PGT124 bound to the trimer and the N332 nanoparticle with comparable EC50s, whereas PGT121 showed preferential binding to the gp140 trimer (**Figure 4A**). This finding is consistent with the reports that PGT121 engages multiple glycans on the Env (41, 42), which may not all

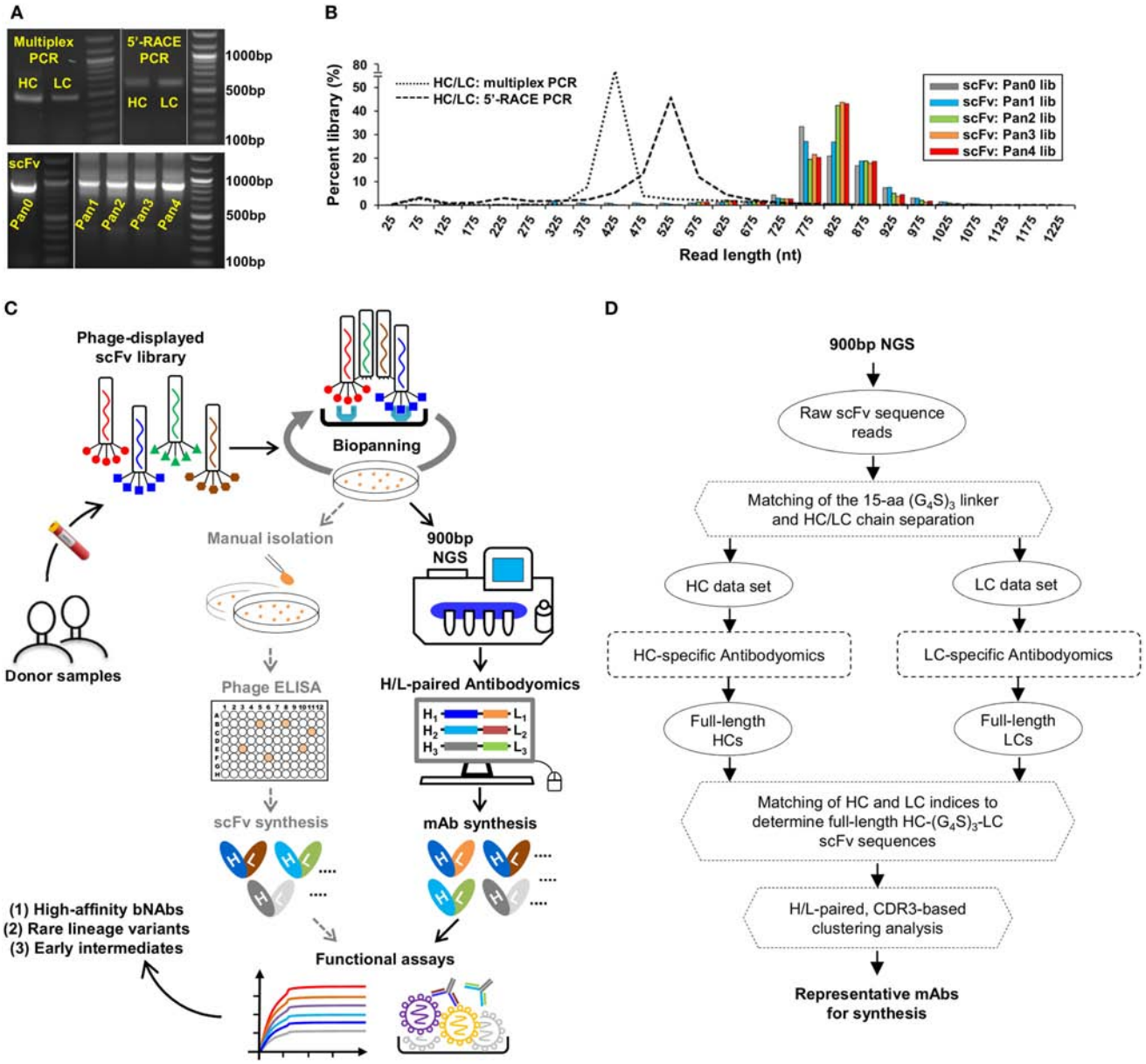


FIGURE 2 | Digital panning, a quantitative method for identification of functional antibodies. **(A)** Gel electrophoresis of three donor-17 antibody libraries generated from multiplex PCR using gene-specific primers, 5′-RACE PCR using single 3′-reverse primers, and H/L-overlapping PCR using gene-specific primers. For the purpose of formatting, the DNA gel has been rearranged with splicing (labeled with white lines). **(B)** Read length distributions of three donor-17 antibody libraries obtained from Personal Genome Machine (PGM) sequencing. Five single-chain variable fragment (scFv) libraries are plotted as histograms and color-coded according to their antigen panning steps: gray (Pan0), cyan (Pan1), green (Pan2), orange (Pan3), and red (Pan4). The two mixed HC/LC libraries generated from multiplex PCR and 5′-RACE PCR are shown as black dotted and dashed lines, respectively. **(C)** Schematic view of the digital panning method with the route of conventional panning included for comparison. Briefly, a scFv library is first constructed from donor peripheral blood mononuclear cells (PBMCs) and displayed on the phage surface for biopanning against an optimized antigen. The pre- and post-panning scFv libraries are sequenced on PGM to achieve 900 bp read length, with the sequenced full-length scFv libraries processed and analyzed by an H/L-paired antibodyomics pipeline. Representative scFv clones are selected for mAb synthesis and functional characterization. In HIV-1 bNAb studies, digital panning may identify high-affinity bNAb-like scFvs, rare bNAb lineage variants, and early intermediates. **(D)** Schematic view of an H/L-paired antibodyomics analysis.

be included in the scaffolded N332 supersite on the nanoparticle surface (46). Among the library-derived mAbs, the two PGT124-like clones (Ab_{d17}-1 and Ab_{d17}-2) appeared to be the best performers, binding to the N332-containing antigens with EC50 values ~10-fold lower than the WT bNAbs in addition to the reduced V1V2 recognition. Among the other mAbs, Ab_{d17}-5 and Ab_{d17}-6 bound the native-like trimer and the gp120 nanoparticle but not the N332 nanoparticle, whereas Ab_{d17}-3 and Ab_{d17}-4 displayed poor antigen binding. The "YYY–MDV" segment in the Ab_{d17}-6 HCDR3, which partially matches to the PGT124 HCDR3, might

be involved in Env binding, but detailed structural characterization would be required to confirm this hypothesis. When evaluated by BLI (**Figure 4B**), Ab_{d17}-1 and Ab_{d17}-2 exhibited PGT124-like trimer binding profiles in comparison with a weak binding signal observed for Ab_{d17}-6. HIV-1 neutralization was tested against a panel of six tier-2 viruses and two tier-1 viruses (**Figure 4C**). While Ab_{d17}-1 and Ab_{d17}-2 displayed the same neutralization breadth and potency as the WT bNAbs, Ab_{17}-6 only neutralized

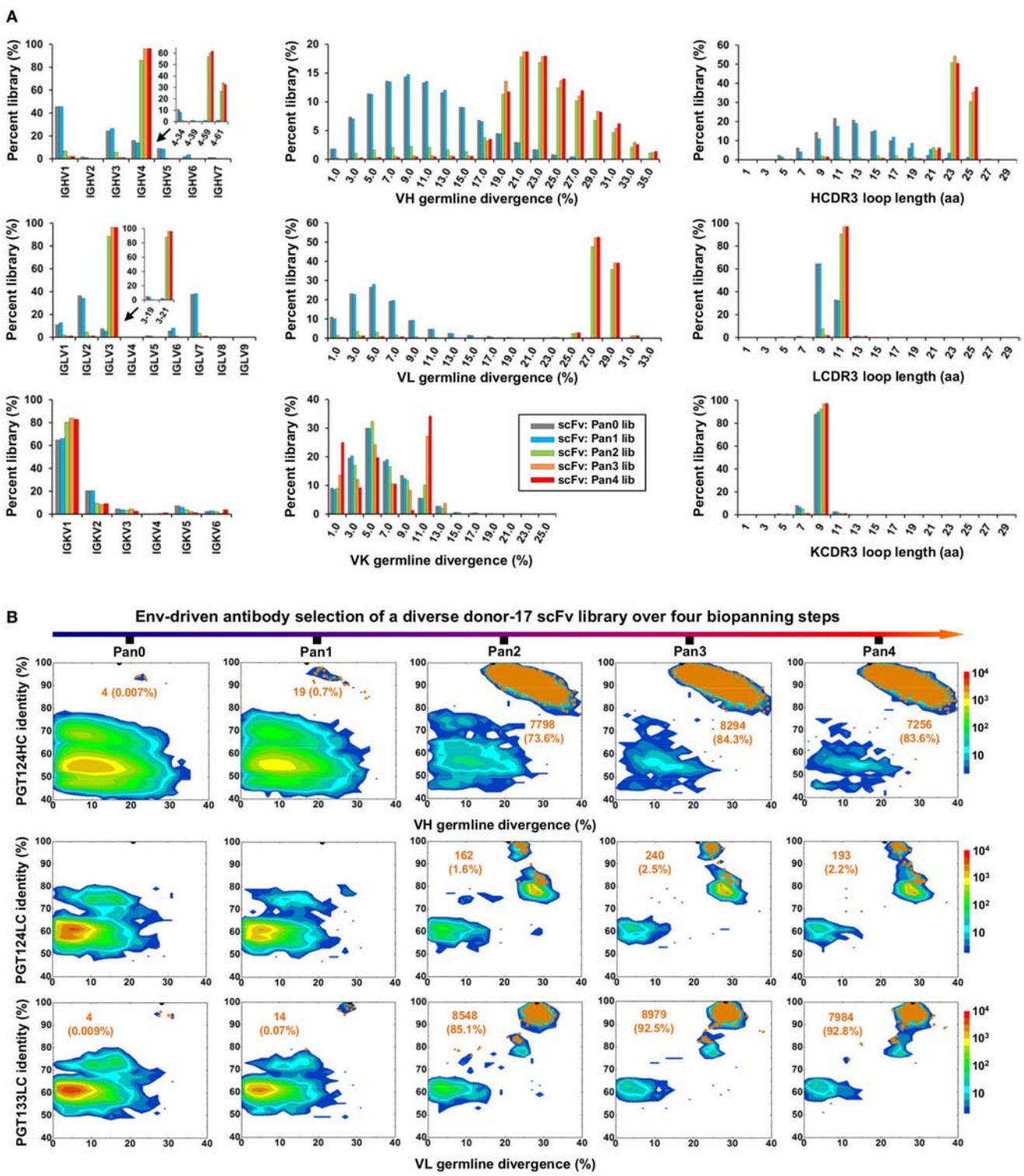


FIGURE 3 | Continued

C

Representative scFv/mAb clones identified by H/L-paired, CDR3-based clustering analysis [a]

Clone	HC name	VH identity	HCDR3 (aa)	Size	LC name	LCDR3 (aa)	Size
Ab_{d17}-1	T2P4H1	85.1%	ARRGQRIYGVVSFGEFFYYYYMDV	7,277	T1P4L1	HYWDSRSPISWI	6,910
Ab_{d17}-2	T2P4H2 [b]	84.3%			T1P4L2	HMWDSRSGFSWS	133
Ab_{d17}-3	T2P4H1				T2P4L3	QSYDSSLSGVV	10
Ab_{d17}-4	T2P4H3	88.2%	DGGDPDTPWFIGAFDV	31	T2P4L3	QSYDSSLSGVV	24
Ab_{d17}-5	T2P4H4	91.2%	ATAPWSH	19	T2P4L4	LLYHSGAQPYVV	15
Ab_{d17}-6	T2P4H5	93.2%	DGEWQIMNYYYKGMDV	45	T2P4K5	QQSFTTPQT	12

[a] Listed items include scFv/mAb clone name, heavy chain (HC) name, VH germline gene identity (%), HCDR3 sequence, HC cluster size, light chain (LC) name, LCDR3 sequence, and LC cluster size. The bracket indicates that Ab_{d17}-1-3 HCs share the same HCDR3 sequences, thus belonging to a single HCDR3 cluster.

[b] T2P4H2 shares the same HCDR3 sequence as WT PGT124 HC but contains a 2-aa mutation in the HCDR1 loop.

FIGURE 3 | Digital panning of a diverse donor-17 single-chain variable fragment (scFv) library against the native-like trimer. A diverse scFv library was constructed from donor-17 peripheral blood mononuclear cells (PBMCs) using a complete set of primers. **(A)** Distributions of germline gene usage (left), somatic hypermutation (middle), and CDR3 loop length (right) are plotted for the five scFv libraries obtained from the trimer panning process. Distributions of prevalent germline gene families within IgHV4 and IgLV3 (>1%) are shown as insert images. Histograms are color-coded following the same scheme as in **Figure 2B**. **(B)** Identity-divergence analysis of the five donor-17 scFv libraries obtained from the trimer panning process. For each library, sequences are plotted as a function of sequence identity to PGT124 HC (top), PGT124 LC (middle), and PGT133 LC (bottom), and sequence divergence from germline genes. Color-coding indicates sequence density at a particular point on the 2D plot. Wild-type (WT) PGT124 and PGT133 are labeled on the 2D plots as black dots. Sequences with CDR3 identity of 95% or greater are shown as orange dots, with the number of sequences and library percentage labeled on the 2D plot for comparison. **(C)** Representative scFv clones identified from a diverse scFv library by H/L-paired, CDR3-based clustering analysis for mAb synthesis.

two tier-1 viruses, indicating the presence of diverse NAb lineages, including the PGT121 lineage, in the donor repertoire.

A key finding thus far in the analysis of a diverse donor-17 scFv library was Ab_{d17}-2, which represented PGT124-like clones with a putative IgHV4-61 origin. Sequence alignment of Ab_{d17}-2 with IgHV1-59, IgHV4-61, and PGT124 revealed a unique 2-aa insertion preceding the "YY" motif in the HCDR1 loop (**Figure 4D**). Is this insertion biologically relevant or merely an error that occurred in the scFv library construction? To answer this question, we prepared an HC-only library from the donor-17 PBMCs (that were used to construct scFv libraries) with a previously reported 5′-RACE PCR protocol (15, 17, 46, 101) and sequenced this library on the Ion S5 platform using an Ion 530 chip. S5 sequencing yielded over 12 million raw reads, which were processed by the *Antibodyomics* 2.0 pipeline, resulting in 7.8 million full-length HCs for repertoire profiling and identity-divergence analysis (Figures S3A,B in Supplementary Material). HC populations with a PGT124 identity of 85–100% and a germline divergence of 15–25% were observed on the 2D plot (**Figure 4E**). Among the 761 sequences with an HCDR3 identity of 90% or greater to PGT124, 11 were assigned to IgHV4-61 with HCDR1 insertions. Interestingly, similar HCDR1 insertions were also found in PGT122 HC variants (Figure S3C in Supplementary Material). Ultra-deep repertoire sequencing thus provided evidence that this HCDR1 insertion was likely a result of lineage diversification. Structural modeling of Ab_{d17}-2 in complex with a native-like gp140 trimer (47) and an engineered gp120 outer domain revealed potential roles of this HCDR1 insertion (**Figure 4F**). As previously reported, the PGT121-class bNAbs utilize an open face formed by three HCDRs to interact with various components of the Env (37, 41, 42). Due to the close proximity of HCDR1 to HCDR2, some Env interactions might be shifted to HCDR1 as a means to accommodate the rapidly changing glycan shield. Indeed, the HCDR1 insertion site is ~30 Å from the glycan-rich V1, V3, and V4 loops, suggesting that an extended loop at this site may mediate interactions with multiple gp120 glycans.

In summary, digital panning of a diverse donor-17 scFv library identified bNAb-like scFv clones and a previously unknown HCDR1 variation within the PGT121 class. For comparative analysis, we also screened this scFv library against a clade-C V1V2 nanoparticle (48), which did not yield any neutralizing mAbs to the apex (Tables S2B and S3B, and Figures S4 and S5 in Supplementary Material), in line with the previous findings for this patient (34). Therefore, our results demonstrated a focused antibody repertoire tuned for specific recognition of a glycan supersite on the native Env.

Digital Panning of a Focused Donor-17 ScFv Library Identifies bNAb Intermediates

Previously, Sok et al. developed a novel phylogenetic method to derive putative intermediates of PGT121-134 based on the 454 sequencing data (18). These intermediates have been studied in detail to inform on the early events of bNAb development (41, 42) and to design trimer immunogens for sequential immunization of Ig knock-in mice (23, 31). Undoubtedly, patient samples will provide a more reliable source of bNAb precursors and intermediates, which, however, are only present at low frequencies within the memory B-cell repertoire. In this study, we hypothesized that digital panning of a bNAb-lineage-focused scFv library may capture these important but rare clones that would otherwise be inaccessible to the standard methods of antibody identification (96).

To test this hypothesis, we constructed a new scFv library from the donor-17 PBMCs using IgHV4- and IgLV3-specific primers to target the germline genes of the PGT121 class (Table S1 in

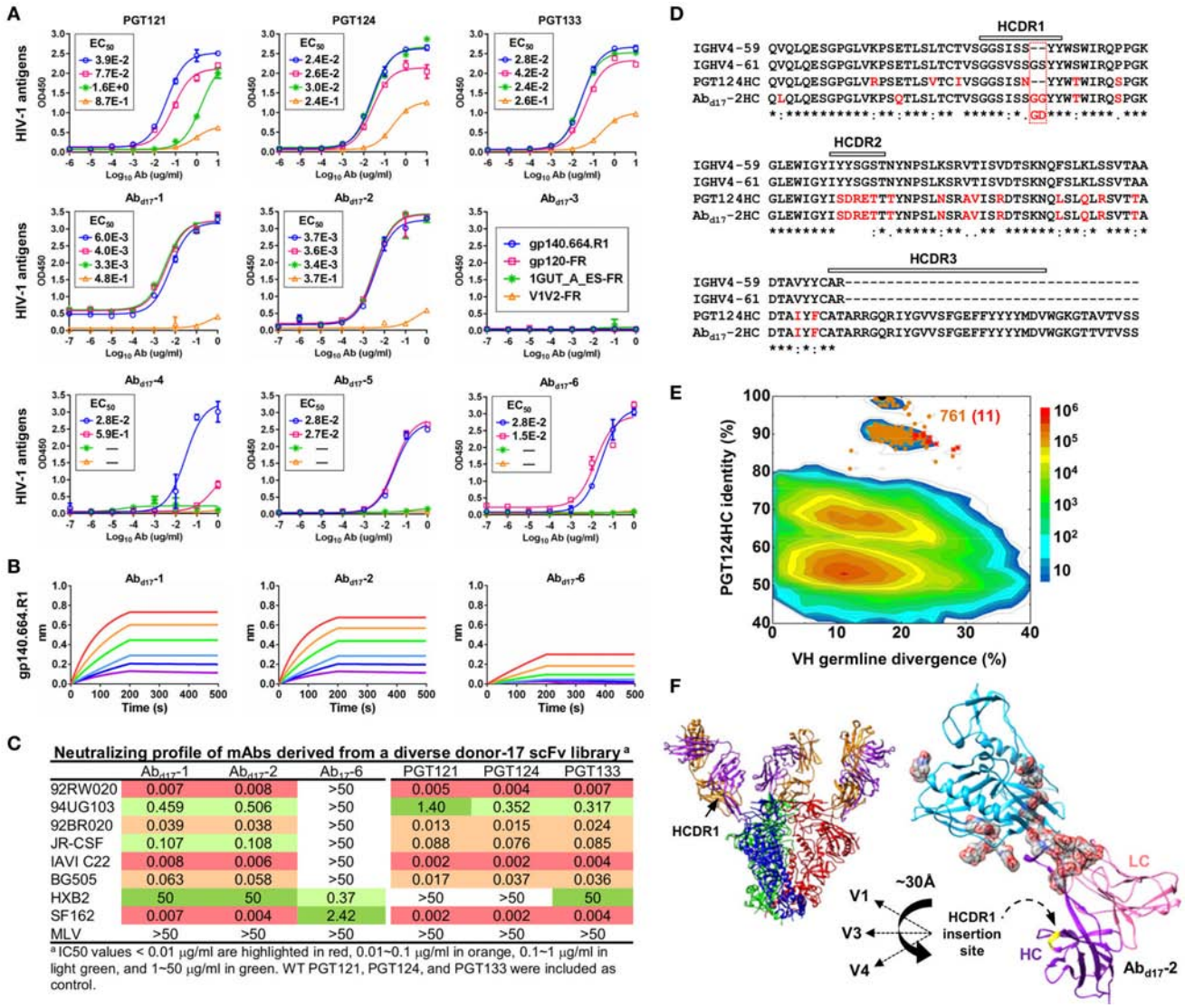


FIGURE 4 | Characterization of monoclonal antibodies (mAbs) identified from a diverse donor-17 single-chain variable fragment (scFv) library. **(A)** Enzyme-linked immunosorbent assay (ELISA) binding of three broadly neutralizing antibodies (bNAbs) of the PGT121 class (PGT121, 124, and 133) and six scFv-derived mAbs, Ab_{d17}-1–6, to four HIV-1 antigens including a native-like trimer (gp140.664.R1), a gp120 nanoparticle (gp120-FR), a nanoparticle presenting 24 copies of the scaffolded N332 supersite (1GUT_A_ES-FR), and a clade-C V1V2 nanoparticle (V1V2-FR). EC_{50} values are labeled for all ELISA plots except for instances in which the highest OD_{450} value is below 0.1 or in the cases of ambiguous data fitting. **(B)** Octet binding of Ab_{d17}-1, -2, and -6 to the native-like trimer gp140.664.R1. Sensorgrams were obtained from an Octet RED96 instrument using a titration series of six concentrations (200–6.25 nM by twofold dilution). **(C)** Neutralization profiles of Ab_{d17}-1, -2, and -6 against six tier-2 viruses and two tier-1 viruses, with WT bNAbs PGT121, 124, and 133 included as references. IC_{50} values are highlighted based on the following color-coding scheme: <0.01 μg/ml (red), 0.01–0.1 μg/ml (orange), 0.1–1 μg/ml (light green), and 1–50 μg/ml (green). **(D)** Sequence alignment of Ab_{d17}-2 with two HC germline genes (IgHV4-59 and IgHV4-61) and PGT124 HC. The three HCDR regions are marked above the sequences, with the mutations with respect to IgHV4-59 colored in red. **(E)** Identity-divergence analysis of the HC repertoire obtained from S5 sequencing. Sequences are plotted as a function of sequence identity to PGT124 HC and sequence divergence from germline genes. Color-coding indicates sequence density at a particular point on the 2D plot. PGT124 HC is labeled on the 2D plot as a black dot. Sequences with HCDR3 identity of 95% or greater to PGT124 are shown as orange dots on the 2D plots, with the ones containing a 2-aa heavy chain complementarity-determining region 1 (HCDR1) insertion shown as red asterisks. **(F)** Structural models of Ab_{d17}-2 in complex with a native-like trimer (gp140.664.R1) and engineered gp120 outer domain (eOD). The Ab_{d17}-2:gp140 complex was modeled upon two trimer structures (PDB IDs: 5JS9 and 5T3S), while the Ab_{d17}-2:eOD complex was modeled upon the PGT124:eOD complex (PDB ID: 4R2G). While the proteins are shown as ribbons in both cases, the molecular surface is also shown for glycans in the latter case. The distance (~30 Å) between the Ab_{d17}-2 HCDR1 and three gp120 glycan-rich loops is labeled.

Supplementary Material). This library was screened against the trimer probe, with the pre- and post-panning libraries sequenced on the PGM using an Ion 314 v2 chip. NGS yielded 809,354 raw reads, which were analyzed with the H/L-paired *Antibodyomics* pipeline (Table S2C in Supplementary Material). Quantitative library profiles revealed convergence patterns similar to those observed for the diverse donor-17 scFv library (**Figure 5A**). In brief, IgHV4-59 and IgLV3-21 accounted for over 90% of

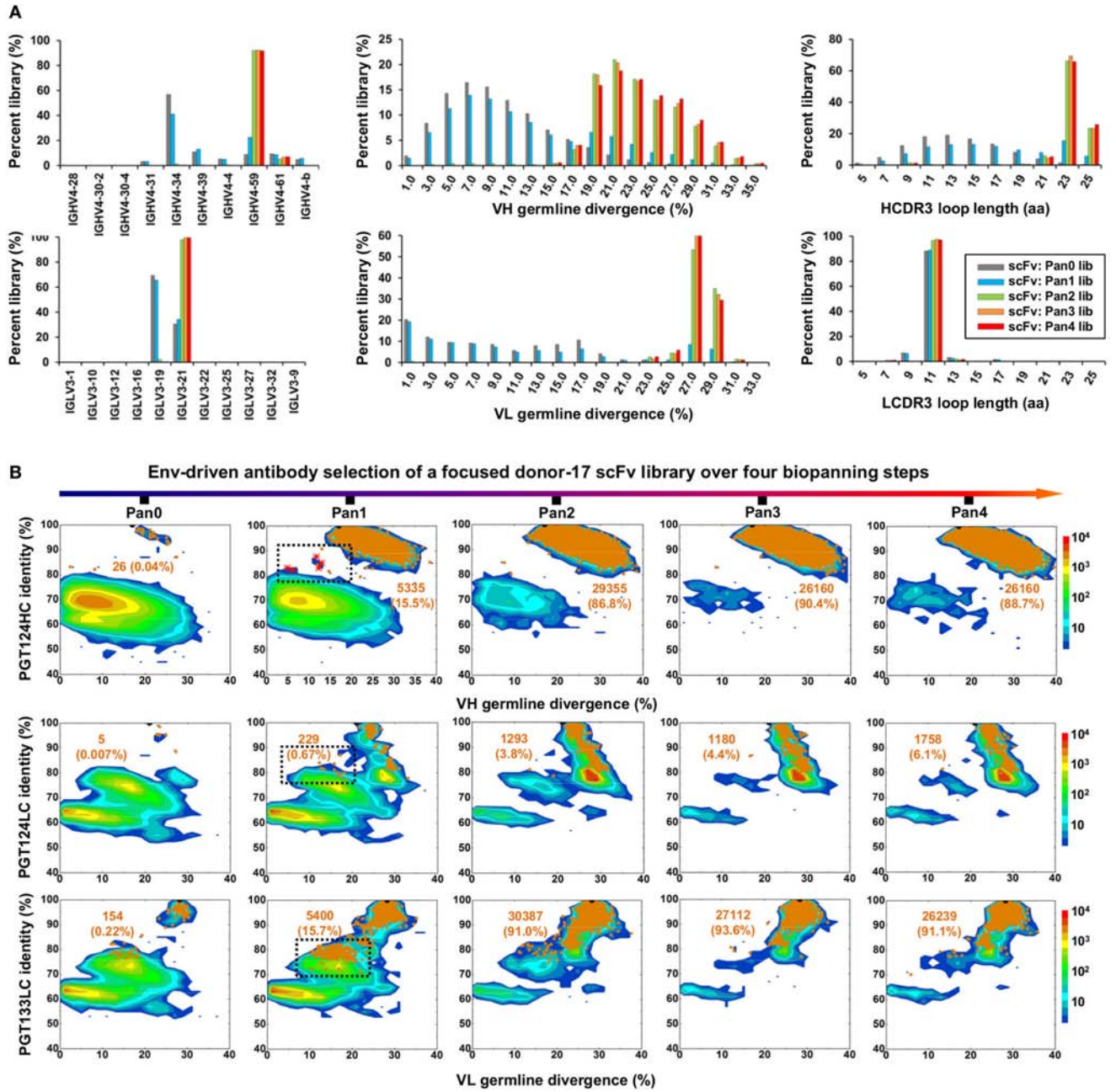


FIGURE 5 | Digital panning of a focused donor-17 single-chain variable fragment (scFv) library against the native-like trimer. A focused scFv library was constructed from donor-17 peripheral blood mononuclear cells (PBMCs) using IgHV4- and IgLV3-specific primers. **(A)** Distributions of germline gene usage (left), somatic hypermutation (middle), and CDR3 loop length (right) are plotted for the five scFv libraries obtained from the trimer panning process. Histograms are color-coded following the same scheme as in **Figure 2B**. **(B)** Identity-divergence analysis of the five donor-17 scFv libraries obtained from the trimer panning process. For each library, sequences are plotted as a function of sequence identity to PGT124 HC (top), PGT124 LC (middle), and PGT133 LC (bottom), and sequence divergence from germline genes. Color-coding indicates sequence density at a particular point on the 2D plot. WT PGT124 and PGT133 are labeled on the 2D plots as black dots. Sequences with CDR3 identity of 95% or greater are shown as orange dots, with the number of sequences and library percentage labeled on the 2D plot for comparison. Sequences with low levels of SHM and moderate identity to the respective broadly neutralizing antibodies, corresponding to native intermediates (NINs), are circled by black dotted line boxes on the 2D plots for the Pan1 library. Five NIN-like scFv clones selected from the Pan1 library based on the HC identity to PGT124 are shown as red asterisks on the 2D plot.

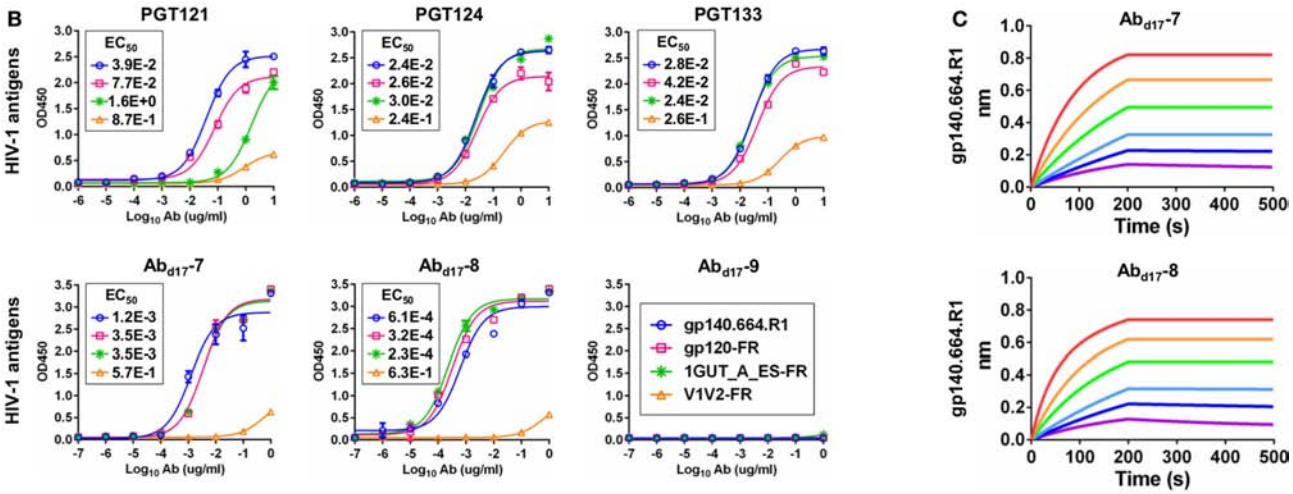
HCs and LCs upon convergence. The average degree of SHM increased from 12% and 11% to 23% and 27% for HCs and LCs, respectively. In addition, more than 90% of HCs contained long HCDR3 loops of 23–25 aa, characteristic of the PGT121-class bNAbs. Furthermore, the 2D plots demonstrated co-enrichment of PGT124-like HCs and PGT124/PGT133-like LCs, consistent

with our findings for the diverse scFv library but with a notable difference in the pattern of distribution (**Figure 5B**). Nonetheless, the H/L-paired clustering analysis identified three representative scFv clones from the converged library that resembled the WT bNAbs (**Figure 6A**; Table S3C in Supplementary Material). While the scFv-derived HC database appeared to be enriched for PGT124-like sequences of the IgHV4-59 origin, the PGT133-like LCs were more prevalent than the PGT124-like LCs, showing an ~16-fold difference in their frequencies. Three mAbs (Ab_{d17}-7–9) were synthesized for evaluation of antigen binding by ELISA (**Figure 6B**). Similar to Ab_{d17}-1 and Ab_{d17}-2 identified from the diverse scFv library, Ab_{d17}-7 and Ab_{d17}-8 bound to the three N332-containing antigens with EC50 values 10- to 100-fold lower than the WT bNAbs. Recognition of the native-like trimer by Ab_{d17}-7 and Ab_{d17}-8 was then confirmed by BLI (**Figure 6C**), which displayed binding profiles comparable to the WT bNAbs

A **Representative scFv/mAb clones identified by H/L-paired, CDR3-based clustering analysis** [a]

Clone	HC name	VH identity	HCDR3 (aa)	Size	LC name	LCDR3 (aa)	Size
Ab_{d17}-7	T1P4H1				T1P4L1	HYWDSRSPISWI	23,502
Ab_{d17}-8	T1P4H1	83.3%	ARRGQRIYGVVSFGEFFYYYYMDV	25,597	T1P4L2	HMWDSRSGFSWS	1,449
Ab_{d17}-9	T1P4H1				T1P4L3	QSYDSSLSGVV	50

[a] Listed items include scFv/mAb clone name, heavy chain (HC) name, VH germline gene identity (%), HCDR3 sequence, HC cluster size, light chain (LC) name, LCDR3 sequence, and LC cluster size. The bracket indicates that Ab_{d17}-7-8 HCs share the same HCDR3 sequences, thus belonging to a single HCDR3 cluster.



D **Neutralizing profile of mAbs derived from a focused donor-17 scFv library** [a]

	Ab_{d17}-7	Ab_{d17}-8	PGT121	PGT124	PGT133
92RW020	0.007	0.008	0.005	0.004	0.007
94UG103	0.358	0.318	1.40	0.352	0.317
92BR020	0.028	0.030	0.013	0.015	0.024
JR-CSF	0.080	0.081	0.088	0.076	0.085
IAVI C22	0.006	0.005	0.002	0.002	0.004
BG505	0.045	0.049	0.017	0.037	0.036
HXB2	50	50	>50	>50	50
SF162	0.003	0.005	0.002	0.002	0.004
MLV	>50	>50	>50	>50	>50

[a] IC50 values < 0.01 μg/ml are highlighted in red, 0.01~0.1 μg/ml in orange, 0.1~1 μg/ml in light green, and 1~50 μg/ml in green. WT PGT121, PGT124, and PGT133 were included for comparison.

FIGURE 6 | Characterization of monoclonal antibodies (mAbs) identified from a focused donor-17 single-chain variable fragment (scFv) library. **(A)** Representative scFv clones identified from a focused scFv library by H/L-paired, CDR3-based clustering analysis for mAb synthesis. **(B)** Enzyme-linked immunosorbent assay (ELISA) binding of three broadly neutralizing antibodies (bNAbs) of the PGT121 class (PGT121, 124, and 133) and three scFv-derived mAbs, Ab_{d17}-7–9, to the gp140.664.R1 trimer and three nanoparticles (gp120-FR, 1GUT_A_ES-FR, and V1V2-FR). EC_{50} values are labeled for all ELISA plots except for instances in which the highest OD_{450} value is below 0.1 or in the cases of ambiguous data fitting. **(C)** Octet binding of Ab_{d17}-7 and -8 to the gp140.664.R1 trimer. Sensorgrams were obtained from an Octet RED96 instrument using a titration series of six concentrations (200–6.25 nM by twofold dilution). **(D)** Neutralization profiles of Ab_{d17}-7 and -8 against six tier-2 viruses and two tier-1 viruses, with WT bNAbs PGT121, 124, and 133 included as references. IC_{50} values are highlighted based on the same color-coding scheme as in **Figure 4C**.

(**Figure 1C**). Consistently, Ab_{d17}-7 and Ab_{d17}-8 also demonstrated neutralizing breadth and potency almost identical to the PGT121-class bNAbs with some weak activity against a tier-1 HXB2 strain, which could not be neutralized by the WT bNAbs (**Figure 6D**). Overall, screening of a focused donor-17 scFv library by a native-like trimer probe resulted in mAbs with optimized binding properties and bNAb-like neutralization profiles.

Although trimer panning of two donor-17 scFv libraries converged to the bNAb-like clones with similar functions, visual comparison revealed "islands" of sequences present only on the 2D plots of the focused library, but not the diverse library (**Figure 5B**). To explore the cause of this discrepancy, we plotted the HC and LC sequences with a CDR3 identity of 95% or greater to the WT bNAbs. Surprisingly, these islands were mainly occupied by sequences with a low level of SHM (<15%) and moderate identity to the WT bNAbs (~80%), suggesting that they may be the native intermediates (termed NINs) of the PGT121-class bNAbs. Interestingly, these NINs were most visible after the first panning step (termed Pan1) but began to disappear as the library was further enriched for high-affinity bNAb-like clones. Five NINs (NIN_{d17}-1–5) were selected from the Pan1 library for experimental testing (**Figure 7A**; Table S4 in Supplementary Material). Of note, sequence analysis revealed that the HCs of all five NINs contained the mature PGT124 HCDR3 loop with a low level of SHM (3–10%) but were assigned to three different germline genes: IgHV4-34, IgHV4-59, and IgHV4-61. While Ab_{d17}-2 and NIN_{d17}-2 sharing the same IgHV4-61 germline gene with differing levels of SHM suggest an actively evolving PGT124 sub-lineage possessing the 2-aa insertion in HCDR1, the near-germline IgHV4-34 HCs with a mature PGT124 HCDR3 cannot be explained within the current framework of the PGT121-class bNAb lineage development. Due to the random HC/LC pairing in scFv library construction, diverse LCs were observed for the five NINs, with the PGT133-like LCs used by NIN_{d17}-4 and -5. Antigen binding of the five NIN-derived mAbs was then assessed by ELISA, with NIN_{d17}-3 and -4 displaying notable affinities for the gp140.664.R1 trimer (47), the gp120 nanoparticle (48), and the N332 nanoparticle (**Figure 7B**) (46). To eliminate the effect of non-functional LCs, we paired the five NIN HCs with the PGT124 and PGT133 LCs. As expected, pairing with the WT bNAb LCs could restore the antigen affinity of NIN_{d17}-2, -3, and -4, while moderately increasing the antigen affinity of NIN_{d17}-1 and -5. Lastly, a total of nine NIN-derived mAbs were tested for neutralization, with eight showing neutralizing activity (**Figure 7C**). Overall, NIN mAbs reconstituted from the IgHV4-59 and IgHV4-61 HCs outperformed those reconstituted from the IgHV4-34 HCs, exhibiting neutralization breadth on par with that of the WT bNAbs, although with reduced potency. Furthermore, the IgHV4-59 and IgHV4-61 HCs, when paired with the PGT124 LC, displayed more potent neutralization than their PGT133 LC counterparts. The results also suggested that an SHM level of 10% or lower may be sufficient for the PGT121-class bNAb intermediates to achieve effective Env recognition (by targeting the N332 supersite) and broad HIV-1 neutralization.

Sequence alignment and structural modeling revealed distinct features of the newly derived PGT121-class intermediates (**Figure 8**). For example, compared with the WT PGT124 and an inferred PGT124 intermediate (32H) (18), NIN_{d17}-2–4 showed fewer mutations in the HCDR2 loop, suggesting that the highly mutated HCDR2 motif in the mature bNAbs may not be as critical as previously thought. Another important finding was an *N*-linked glycosylation site in the HCDR1 loop of NIN_{d17}-4, suggesting that HCDR1 may be a focal point of maturation for facilitating Env interactions through various mechanisms. In contrast, the HCs of NIN_{d17}-1 and -5 were IgHV4-34 germline-like sequences with only four to six mutations within the V gene. Furthermore, while the HC mutations within the WT PGT124 and an inferred 32H (18) displayed similar distribution patterns across the protein surface, the HC mutations within NIN_{d17}-2 and NIN_{d17}-4 were focused mainly on HCDR1 of the functionally important open face (37). Additionally, we identified the NIN-like PGT124 and PGT133 LCs, which contained a similar deletion in FR1 but lacked the 3-aa FR3 insertion compared to the mature bNAbs (Figure S6 in Supplementary Material).

Our results confirmed that PGT124, and a related 10-1074 (36), represents a distinct branch of lineage development with respect to other members of the PGT121 class (18, 40–42), and define the minimal SHM needed to achieve broad HIV-1 neutralization. Since PGT124 and 10-1074 mainly require the N332 glycan for Env recognition (42, 45), and PGT124 intermediates with different levels of SHM have been found in the donor repertoire, the PGT124 sub-lineage may provide a more promising template than PGT121 for immunogen design targeting the N332 supersite.

PGT124 Intermediates Compete with Trimer-Elicited Mouse Sera for the N332 Supersite

Since the unmutated SOSIP.664 trimer was poorly recognized by the PGT121 precursor, random mutagenesis and structure-based design were undertaken to create sequential trimer immunogens to target the inferred PGT121 precursor and intermediates (18, 31). Although these modified trimers induced bNAb-like responses in Ig knock-in mice (23), they have not been tested in WT animals. In this study, PGT124 intermediates were selected by a native-like trimer probe presenting an intact glycan shield. Based on this finding, we hypothesized that an unmutated trimer may be able to elicit N332-speficic antibody responses in WT animals and that the bNAb intermediates (NINs) identified from the donor-17 scFv library can in turn be used in a competition assay to assess the N332 specificity of serum antibodies elicited by a trimer, or an N332-focused immunogen.

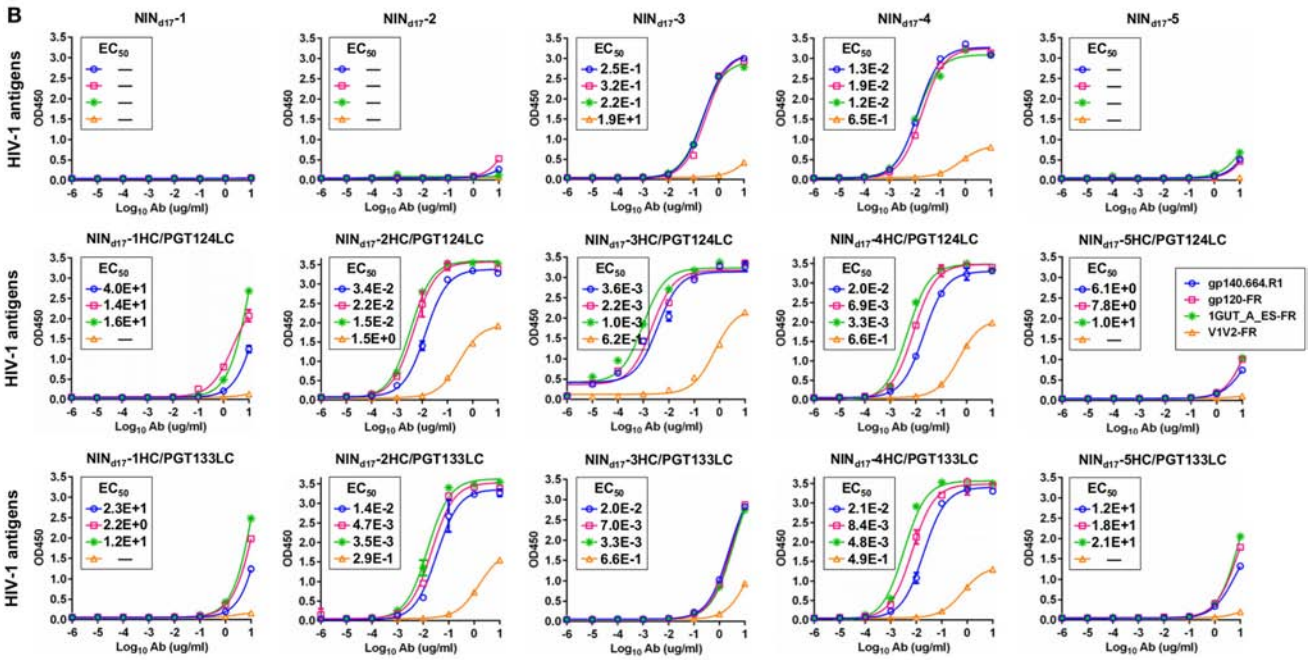
We briefly investigated this hypothesis by examining serum samples from a previous mouse immunization aimed to study the early B-cell responses to the N332 supersite and MPER in the context of multivalent scaffolds and native-like trimers (46). It was reported that a scaffolded full-length gp140 trimer elicited a robust antibody response to the apex, suggesting that the inclusion of MPER and a C-terminal scaffold domain can stabilize the glycan shield and facilitate bNAb recognition (46). In this study, we first performed ELISA to assess sera from mice immunized with the gp140.664.R1 trimer, which was also the basis of the trimer probe used for screening the donor-17 scFv libraries (**Figure 9A**). Consistent with the previous finding for the SOSIP trimer (102)

A **Five native intermediate (NIN) scFvs identified from a focused donor-17 scFv library after the first panning step [a]**

Library index	Clone	VH gene	VH ID	$HCDR3_{PGT124}$ ID	VL gene	VH ID	LCDR3 (aa)	$LCDR3_{PGT133}$ ID
4911	NIN_{d17}-1	IGHV4-34*01	97.3%	100%	IGLV3-19*01	98.6%	NSRDSSGNHRV	61.1%
27471	NIN_{d17}-2	IGHV4-61*02	91.3%	100%	IGLV3-21*02	93.5%	QVWDSSSDEVV	63.9%
106020	NIN_{d17}-3	IGHV4-59*01	91.8%	100%	IGLV3-19*01	100%	NSRDSSGNHRV	61.1%
118139	NIN_{d17}-4	IGHV4-59*01	90.8%	100%	IGLV3-21*01	77.1%	HYWDSRSPISWI	100%
123988	NIN_{d17}-5	IGHV4-34*01	94.2%	100%	IGLV3-21*01	77.0%	HYWDSRSPISWI	100%

[a] Listed items include scFv index in the Pan1 library (library index), scFv/mAb clone name, HC identity to VH germline gene (VH ID), HCDR3 identity to PGT124 HCDR3 ($HCDR3_{PGT124}$ ID), LC identity to VL germline gene (VL ID), amino acid sequence of LCDR3 (LCDR3 (aa)), and LCDR3 identity to PGT133 LCDR3.

B



C **Neutralizing profile of native intermediate (NIN)-derived mAbs identified from a focused donor-17 scFv library [a]**

	NIN_{d17}-1HC /PGT133LC[b]	NIN_{d17}-2HC /PGT124LC	NIN_{d17}-2HC /PGT133LC	NIN_{d17}-4[c]	NIN_{d17}-4HC /PGT124LC	NIN_{d17}-4HC /PGT133LC	NIN_{d17}-5HC /PGT124LC	NIN_{d17}-5HC /PGT133LC	PGT121	PGT124	PGT133[d]
92RW020	5.03	0.006	0.030	0.026	0.010	0.015	10.0	2.74	0.007	0.006	0.007
94UG103	10.0	0.591	1.384	1.654	1.055	1.760	10.0	10.0	0.500	0.372	0.317
92BR020	>10	0.020	0.123	0.075	0.032	0.060	10.0	10.0	0.012	0.014	0.024
JR-CSF	>10	0.046	0.332	0.166	0.097	0.098	10.0	10.0	0.040	0.039	0.085
IAVI C22	>10	0.007	0.037	0.019	0.008	0.023	10.0	3.06	0.005	0.005	0.004
BG505	10.0	0.030	0.226	0.182	0.119	0.116	10.0	10.0	0.043	0.035	0.036
HXB2	2.30	10.0	0.674	–	>10	>10	1.51	1.40	>10	>10	>10
SF162	>10	0.007	0.052	–	0.015	0.014	>10	4.21	0.012	0.007	0.004
MLV		>50	>50	>50	>50	>50	>50	>50	>50	>50	>50

[a] IC50 values < 0.01 μg/ml are highlighted in red, 0.01~0.1 μg/ml in orange, 0.1~1 μg/ml in light green, and 1~50 μg/ml in green. WT PGT121 and 124 were included for comparison. [b] NIN_{d17}-1HC/PGT124LC did not neutralize any of the 8 isolates tested and therefore is not shown. [c] For NIN_{d17}-4, HXB2 and SF162 were not tested but expected to produce IC50 values comparable to those obtained for the two LC variants (NIN_{d17}-4HC/PGT124LC and NIN_{d17}-4HC/PGT133LC). [d] WT PGT133 data from the previous neutralization (Figure 6D) was included here for comparison.

FIGURE 7 | Characterization of the PGT124 intermediates identified from a focused donor-17 single-chain variable fragment (scFv) library. **(A)** Five native intermediates, NIN_{d17}-1–5, identified from the Pan1 library based on the HCDR3 analysis. **(B)** Enzyme-linked immunosorbent assay (ELISA) binding of NIN_{d17}-1–5 and their LC variants to the gp140.664.R1 trimer and three nanoparticles (gp120-FR, 1GUT_A_ES-FR, and V1V2-FR). The two sets of LC variants were derived by pairing the NIN HCs with PGT124 and PGT133 LCs. EC_{50} values are labeled for all ELISA plots except for instances in which the highest OD_{450} value is below 0.1 or in the cases of ambiguous data fitting. **(C)** Neutralization profiles of selected NIN_{d17} variants against six tier-2 viruses and two tier-1 viruses, with WT broadly neutralizing antibodies PGT121, 124, and 133 included as references. IC_{50} values are highlighted based on the same color-coding scheme as in **Figure 4C**.

and with our previous experiment (46), the apex and the N332 supersite were not well recognized by mouse sera, as indicated by two nanoparticle probes. In comparison, enhanced binding to both glycan epitopes was observed for mouse sera elicited by the scaffolded gp140.681.R1-1NOG trimer (**Figure 9B**), suggesting that an unmutated trimer, when presented in a proper structural context, can induce antibody responses to the bNAb epitopes in the glycan shield. Two library-derived bNAb intermediates,

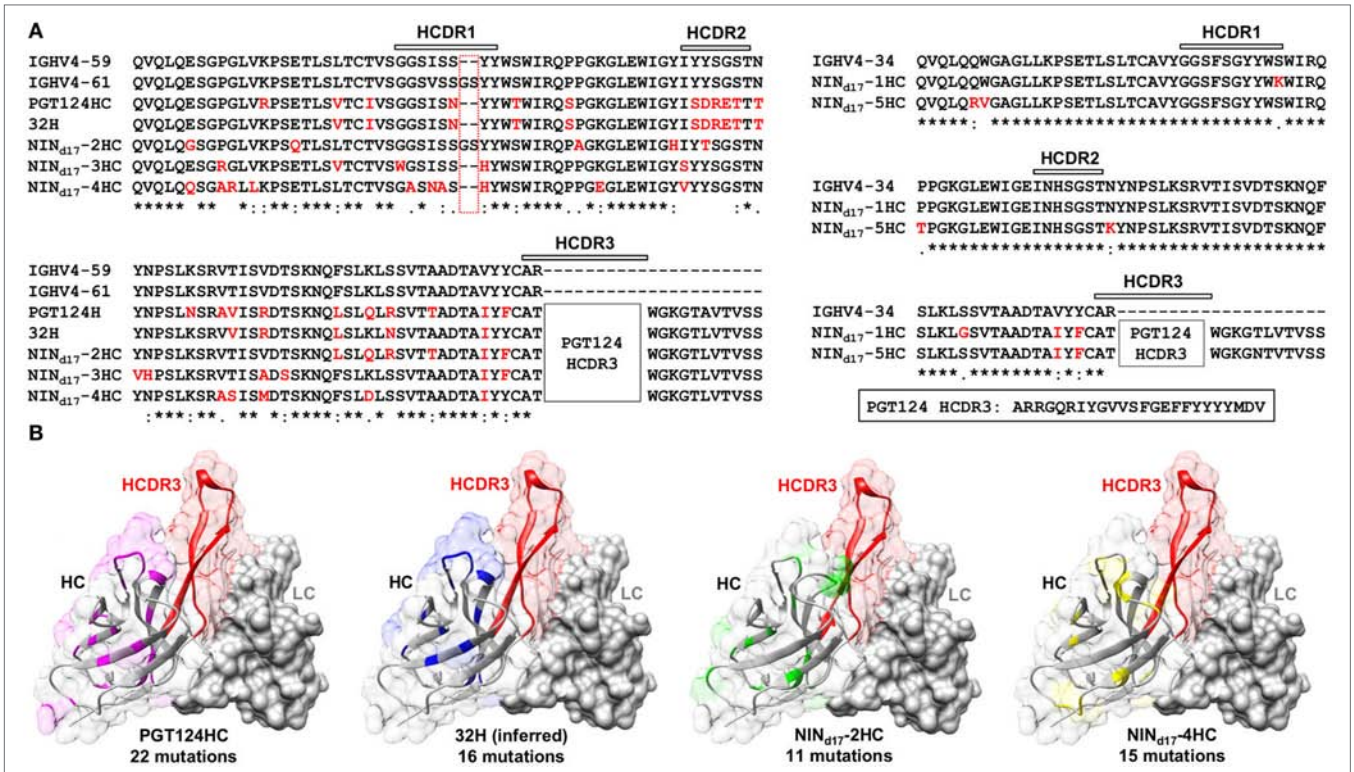


FIGURE 8 | Sequence and structure analyses of the PGT124 intermediates identified from a focused donor-17 single-chain variable fragment (scFv) library. **(A)** Sequence alignment of NIN$_{d17}$-2, -3, and -4 with two HC germline genes (IgHV4-59 and IgHV4-61), PGT124 HC, and inferred 32H (18) (left), and sequence alignment of NIN$_{d17}$-1 and -5 with the germline gene IgHV4-34 (right). The three HCDR regions are marked above the sequences, with the mutations with respect to IgHV4-59 (left) and IgHV3-34 (right) colored in red. As the HCDR3 region in all NIN$_{d17}$ sequences is identical to that of WT PGT124, it is highlighted separately and not shown in the sequence alignment. **(B)** Structural mapping of HC mutations onto the antibody surface for PGT124, inferred intermediate 32H, NIN$_{d17}$-2, and NIN$_{d17}$-4. The structures of antibody variable domains were modeled using the PGT124 structure as a template (PDB ID: 4R2G), with HCs shown as ribbons in transparent molecular surface and LCs as solid gray molecular surface. The HCDR3 loop is colored in red in all models, whereas HC mutations are colored in pink for PGT124, blue for 32H, green for NIN$_{d17}$-2, and yellow for NIN$_{d17}$-4. The number of amino acid mutations with respect to the germline gene (IgHV4-59) is indicated under the structural model.

NIN$_{d17}$-1HC/PGT133LC (with weak trimer binding and neutralization) and NIN$_{d17}$-4 (with moderate neutralizing potency), were utilized in a competition ELISA to assess the N332 specificity of mouse sera obtained from the scaffolded trimer group (**Figure 9C**). The WT PGT124 was included in the ELISA assay as a positive control. As expected, PGT124 and NIN$_{d17}$-4—but not the NIN$_{d17}$-1 variant with a near-germline HC—could effectively block serum binding to the N332 nanoparticle, illustrating the utility of the library-derived bNAb intermediates for immunogen evaluation. It also appeared that murine antibodies, with short (on average 9-aa) HCDR3 loops, failed to compete with human bNAb intermediates with a 24-aa HCDR3 loop poised to penetrate the glycan shield.

DISCUSSION

Diverse bNAbs identified from the elite neutralizers have served as useful templates for guiding rational HIV-1 vaccine design (103). However, there is a significant disparity between the degrees of SHM observed for bNAbs and for weak- or non-NAb antibody responses in chronic infection and vaccination. The minimal level of SHM required for neutralization and protection is poorly defined, although recent studies of bNAb-virus co-evolution have begun to shed some light on this critical issue (16, 17, 19–22). Limited by patient samples and antibody technologies, it remains a challenge to identify bNAb precursors and intermediates from infected donors, thus hindering B-cell lineage immunogen design (7). To overcome this challenge, germline reversion and phylogenic analysis have been used to derive putative bNAb precursors and intermediates (18). Analyses of the inferred antibodies have revealed distinct patterns for the PGT121 and PGT124 sub-lineages (40–42). These insights constituted the foundation for designing sequential trimer constructs capable of eliciting bNAb-like responses in Ig knock-in mice (23, 31). Despite these encouraging successes, it is necessary to acknowledge the risk in utilizing inferred antibodies resulting from assumptions which may or may not accurately reflect the true events of B-cell lineage development.

A variety of methods has been applied to the identification of HIV-1 bNAbs, including phage display, hybridoma, single B-cell culturing coupled with large-scale functional screening, and antigen-specific single B-cell sorting by flow cytometry

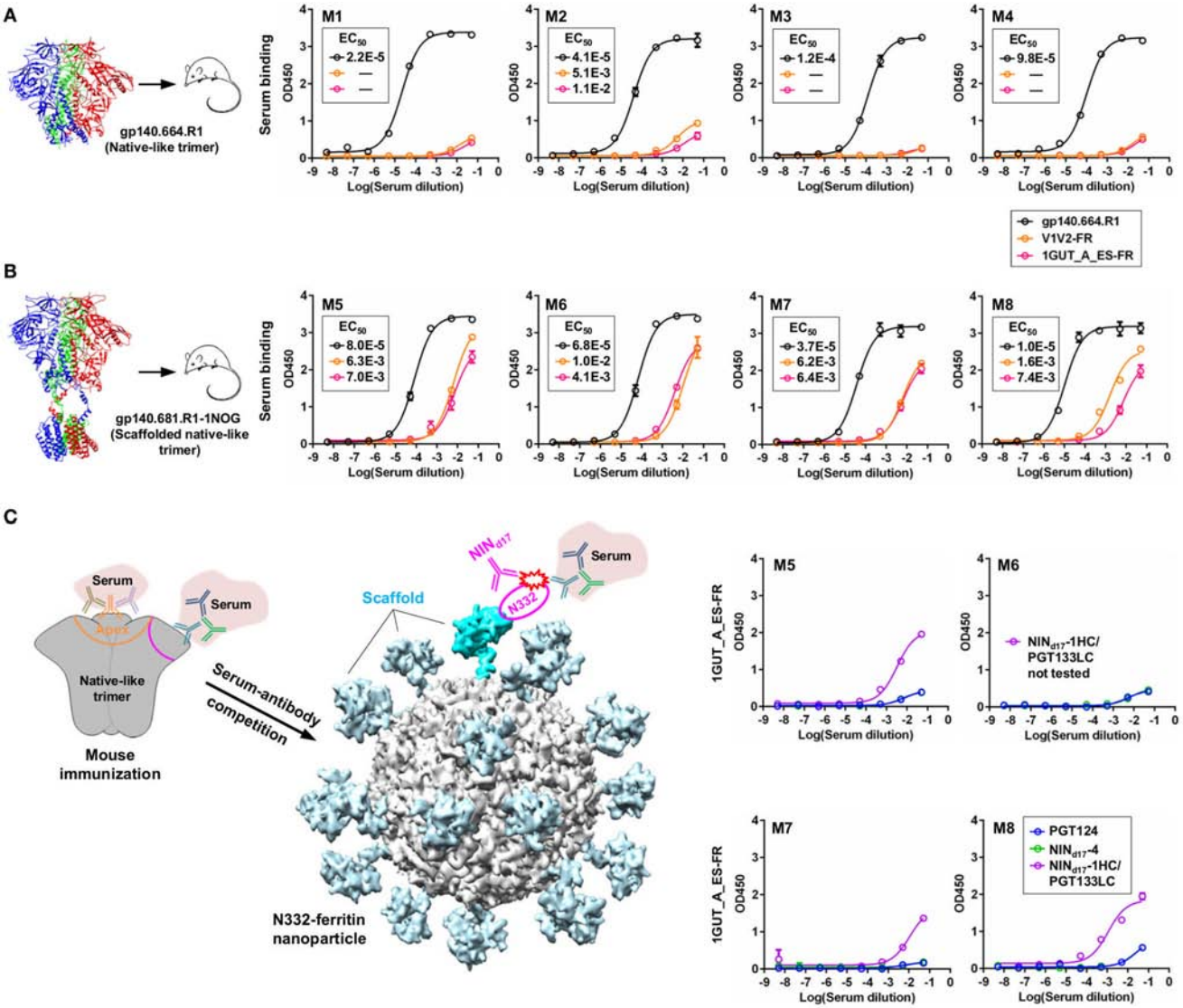


FIGURE 9 | Detection of N332-specific antibody responses in trimer immunization by library-derived PGT124 lineage intermediates. **(A)** Enzyme-linked immunosorbent assay (ELISA) binding of mouse sera elicited by a native-like gp140 trimer (gp140.664.R1) to the trimer immunogen and two nanoparticles (V1V2-FR and 1GUT_A_ES-FR). The structure of this trimer (PDB ID: 5JS9) is shown on the left, with ELISA plots for four subjects in this group (M1–M4) shown on the right. **(B)** ELISA binding of mouse sera elicited by a scaffolded full-length gp140 trimer (gp140.681.R1-1NOG) (39) to the parent trimer (gp140.664.R1) and two nanoparticles (V1V2-FR and 1GUT_A_ES-FR). The structural model of this scaffolded trimer is shown on the left, with ELISA plots for four subjects in this group (M5–M8) shown on the right. **(C)** ELISA binding of mouse sera elicited by the gp140.681.R1-1NOG trimer to an N332 nanoparticle (1GUT_A_ES-FR), in which PGT124 and two intermediate variants (NIN_{d17}-4 and NINd17-1HC/PGT133LC) were added to compete for the N332 supersite. A diagram of the competition ELISA is shown on the left, with ELISA plots for four subjects in this group (M5–M8) shown on the right. EC_{50} values are labeled for all ELISA plots except for instances in which the highest OD_{450} value is below 0.1 or in the cases of ambiguous data fitting.

(96). Among these, the single-cell methods are considered most advantageous, as they enable isolation of mAbs with natively paired HC and LC from live, functional B cells. It was also suggested that new methods would be necessary for identifying rare precursors and intermediates—a challenge facing both bNAb and immunization studies (96). Advances in NGS technology and templating methods have allowed unbiased analysis of B-cell repertoires during HIV-1 infection and vaccination (15, 17, 46, 101). Presently, NGS is applied to the characterization of scFv antibody libraries but not yet to the direct selection of functional scFv clones due to its insufficient lead length to cover both HC and LC within a scFv (100). In this study, a long-read NGS technology was established that permitted high-throughput sequencing of full-length scFv libraries and used in conjunction with an H/L-paired antibodyomics method for library profiling and clone selection. These advances have transformed the conventional scFv library panning method into a quantitative "digital panning" method and will likely improve the single-cell,

H/L-paired antibody isolation and repertoire analysis (104, 105). Using two scFv libraries constructed from the samples of an elite HIV-1 neutralizer (donor-17) (34), we demonstrated the utility of digital panning in the study of PGT121-class bNAbs, with due consideration of recent advances in gp140 trimer design (86, 87). A native-like trimer probe, with structural and antigenic profiles indistinguishable from its parent UFO trimer (47), was utilized to screen donor-17 scFv libraries so as to dissect the details of an N332-dependent bNAb lineage within the antibody repertoire. Digital panning of donor-17 scFv libraries with this trimer probe offered a wealth of novel information on the PGT121 class: PGT122/PGT124 HC variants with a 2-aa HCDR1 insertion, HC and LC intermediates with fewer mutations and different sequence motifs than the inferred intermediates (18), and a versatile open face utilizing HCDR1 and HCDR2 mutations to mediate Env interactions. However, due to the shortcomings of phage display such as primer bias and random H/L pairing, it is possible that only a subset of lineage variants and intermediates were captured. Nonetheless, these new findings will contribute to a more complete understanding of the lineage development of the PGT121-class bNAbs. The bNAb intermediates were also utilized to gauge N332-specific antibody responses elicited by two native-like gp140 trimers. A previously reported gp140 scaffolding strategy (46) appeared to notably enhance the trimer-induced antibody response to the N332 supersite in BALB/c mice, although such response could barely compete with the human bNAb intermediates. Future studies investigating the cause of this enhancement, or evaluating the scaffolded trimers using an extensive regimen, may prove useful for the development of an effective trimer vaccine.

In summary, our study has provided valuable tools, including a native-like trimer probe and the digital panning method, to facilitate bNAb studies, particularly those involving identification of rare bNAb intermediates in HIV-1 patient and vaccination samples. The PGT124 sub-lineage, possessing an invariable HCDR3 loop and multiple library-derived intermediates, may serve as a promising template for B-cell lineage vaccine design targeting the N332 supersite.

ETHICS STATEMENT

Blood samples were acquired from an HIV-1-infected donor (donor-17) of the Protocol G cohort under written consent. The samples were collected following clinical protocols approved by the Republic of Rwanda National Ethics Committee, the Emory University Institutional Review Board, the University of Zambia Research Ethics Committee, the Charing Cross Research Ethics Committee, the UVRI Science and Ethics Committee, the University of New South Wales Research Ethics Committee, the St. Vincent's Hospital and Eastern Sydney Area Health Service, the Kenyatta National Hospital Ethics and Research Committee, the University of Cape Town Research Ethics Committee, the International Institutional Review Board, the Mahidol University Ethics Committee, the Walter Reed Army Institute of Research (WRAIR) Institutional Review Board, and the Ivory Coast Comité National d'Ethique des Sciences de la Vie et de la Santé (CNESVS).

AUTHOR CONTRIBUTIONS

LH and JZ conceived the method. LH, XL, PA, CJM, RA, CDM, and JZ performed digital panning, antibody synthesis, antigen binding assays, and computational analysis. NV, GO, and AW performed negative-stain EM analysis of the gp140 trimer probe. BZ set up the MicroPulser system. KS-F, DS, and DB performed the HIV-1 neutralization assays. LH and JZ wrote the manuscript with input from all coauthors.

ACKNOWLEDGMENTS

Electron microscopy data were collected at the Scripps Research Institute EM Facility.

FUNDING

This work was partly supported by the International AIDS Vaccine Initiative Neutralizing Antibody Center and CAVD, by the Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery (CHAVI-ID UM1 AI00663) (AW, DB), by the HIV Vaccine Research and Design (HIVRAD) program (P01 AI110657) (AW), and by the HIV Vaccine Research and Design (HIVRAD) program [P01 AI124337 (JZ), AI084817 (AW), AI129698 (JZ)], and AI125078-01A1 (JZ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fimmu.2017.01025/full#supplementary-material.

FIGURE S1 | Negative-stain EM of the biotinylated, Avi-tagged BG505 gp140 trimer probe used for biopanning. This trimer probe contains a redesigned heptad repeat 1 bend (47) and a biotinylated Avi-tag located immediately downstream of residue 664 (termed gp140.664.R1-Avi-Biot). **(A)** Raw micrograph of the BG505 gp140.664.R1-Avi-Biot trimer. **(B)** Reference-free 2D class averages of the BG505 gp140.664.R1-Avi-Biot trimer. Percentages of native-like trimers (closed and partially open or breathing trimers) and non-native species (misfolded trimers as well as dimers and monomers) are indicated. **(C)** The estimated resolution (~21 Å) of the EM reconstruction was calculated from the Fourier Shell Correlation (FSC) using a cutoff of 0.5. **(D)** Top and side views of the 3D EM reconstruction of the BG505 gp140 SOSIP.664 trimer and the gp140.664.R1-Avi-Biot trimer. The EM densities of the SOSIP trimer are shown in gray transparent surface with the crystal structure (PDB 4TVP, gp120 in blue with V1V2 in magenta, V3 in green and gp41 in brown) fitted into the density. The EM densities of the gp140.664.R1-Avi-Biot trimer are shown in gray transparent surface with the SOSIP trimer densities overlaid as wire mesh (in orange). The contour level used for the gp140.664.R1-Avi-Biot trimer density was ~33.

FIGURE S2 | Assessment of the *Antibodyomics* 2.0 pipeline. **(A)** Schematic view of the chain-specific *Antibodyomics* pipeline, which consists of five steps including (1) data reformatting and cleaning, (2) germline gene assignment, (3) sequencing error correction, (4) calculation of sequence identity to a set of known antibodies, and (5) determination of CDR3 and variable domain boundaries. Step 3 (error correction) is highlighted with a red dashed-line box. **(B)** Distribution of improvement in sequence quality resulting from error correction. Quality improvement is measured by the change of amino acid sequence identity with respect to the germline V gene. For the donor-17 antibody chain data generated by 454 sequencing (18), *Antibodyomics* 1.0 yielded an average improvement of 19.4 and 21.7% for HC and LC, respectively, in comparison with 22.5 and 24.1% from *Antibodyomics* 2.0.

FIGURE S3 | Ultra-deep sequencing of the donor-17 HC repertoire. **(A)** Distributions of germline gene usage (left), somatic hypermutation (middle), and HCDR3 length (right). In the histogram of germline gene family distribution, IgHV4 is highlighted in black whereas other germline genes are shown in gray (left 1). A more detailed distribution within the IgHV4 family is also plotted (left 2). **(B)** Identity-divergence analysis of the donor-17 HC repertoire using the PGT121-class bNAb HCs as templates. Sequences are plotted as a function of sequence identity to WT bNAb HCs and germline divergence. Color-coding indicates sequence density at a particular point on the 2D plot. Wild-type bNAb HCs are labeled on the 2D plots as black dots. Sequences with HCDR3 identity of 90% or greater and those assigned to IgHV4-61 are shown as orange dots and red asterisks, respectively, with the number of sequences for each labeled on the 2D plot. **(C)** Sequence alignment of selected HCs of the putative IgHV4-61 origin with respect to two germline genes (IgHV4-59 and IgHV4-61) and WT bNAb HCs. The three HCDR regions are marked above the sequences, with the mutations with respect to IgHV4-59 colored in red. Below the sequence alignment, asterisk (*) indicates identical residues, colon (:) indicates residues with strongly similar properties, and period (.) indicates residues with weakly similar properties.

FIGURE S4 | Digital panning of a diverse donor-17 single-chain variable fragment (scFv) library against a clade-C V1V2-ferritin nanoparticle. This scFv library, constructed from the donor-17 peripheral blood mononuclear cells (PBMCs) using a large set of primers, has been screened against a native-like gp140 trimer probe, gp140.664.R1-Avi-Biot. Distributions of germline gene usage **(A)**, somatic hypermutation **(B)**, and CDR3 loop length **(C)** are plotted for the five scFv libraries obtained from the nanoparticle panning process. Histograms are color-coded according to their antigen panning steps: gray (Pan0), cyan (Pan1), green (Pan2), orange (Pan3), and red (Pan4).

FIGURE S5 | Identification and characterization of monoclonal antibodies (mAbs) from a diverse donor-17 single-chain variable fragment (scFv) library screened against a clade-C V1V2-ferritin nanoparticle. **(A)** Six prevalent scFv clones identified by H/L-paired, CDR3-based clustering analysis. **(B)** Enzyme-linked immunosorbent assay (ELISA) binding of three representative bNAbs of the PGT121 class (PGT121, 124, and 133) and six scFv-derived mAbs (VAb_{d17}-1–6) to four HIV-1 antigens including a native-like trimer (gp140.664.R1), a gp120-ferritin nanoparticle (gp120-FR), an N332 nanoparticle (1GUT_A_ES-FR), and a V1V2-ferritin nanoparticle (V1V2-FR). For VAb_{d17}s, ferritin was included in the ELISA as a negative control. EC_{50} values are labeled for all ELISA plots except for instances in which the highest OD_{450} value is below 0.1 or in the cases of ambiguous data fitting.

FIGURE S6 | Additional native intermediates (NINs) selected from a focused donor-17 single-chain variable fragment (scFv) library during the trimer panning process. **(A)** Sequence alignment of HCs with assigned germline genes and WT PGT124 HC. **(B)** Sequence alignment of LCs with germline gene IgLV3-21 and WT PGT124 LC or PGT133 LC. The three HCDR regions are marked above the sequences, with the mutations with respect to the assigned germline gene colored in red. Below the sequence alignment, asterisk (*) indicates identical residues, colon (:) indicates residues with strongly similar properties, and period (.) indicates residues with weakly similar properties.

REFERENCES

1. Kwong PD, Mascola JR. Human antibodies that neutralize HIV-1: identification, structures, and B cell ontogenies. *Immunity* (2012) 37:412–25. doi:10.1016/j.immuni.2012.08.012
2. Kwong PD, Mascola JR, Nabel GJ. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nat Rev Immunol* (2013) 13:693–701. doi:10.1038/nri3516
3. Burton DR, Mascola JR. Antibody responses to envelope glycoproteins in HIV-1 infection. *Nat Immunol* (2015) 16:571–6. doi:10.1038/ni.3158
4. Klein F, Mouquet H, Dosenovic P, Scheid JF, Scharf L, Nussenzweig MC. Antibodies in HIV-1 vaccine development and therapy. *Science* (2013) 341:1199–204. doi:10.1126/science.1241144
5. Walker LM, Burton DR. Rational antibody-based HIV-1 vaccine design: current approaches and future directions. *Curr Opin Immunol* (2010) 22:358–66. doi:10.1016/j.coi.2010.02.012
6. Burton DR, Ahmed R, Barouch DH, Butera ST, Crotty S, Godzik A, et al. A blueprint for HIV vaccine discovery. *Cell Host Microbe* (2012) 12:396–407. doi:10.1016/j.chom.2012.09.008
7. Haynes BF, Kelsoe G, Harrison SC, Kepler TB. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat Biotechnol* (2012) 30:423–33. doi:10.1038/nbt.2197
8. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* (2011) 333:1593–602. doi:10.1126/science.1207532
9. Zhu J, O'Dell S, Ofek G, Pancera M, Wu X, Zhang B, et al. Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinfornnatics. *Front Microbiol* (2012) 3:315. doi:10.3389/fmicb.2012.00315
10. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* (2013) 39:245–58. doi:10.1016/j.immuni.2013.04.012
11. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc Natl Acad Sci USA* (2013) 110:6470–5. doi:10.1073/pnas.1219320110
12. Zhu J, Wu X, Zhang B, McKee K, O'Dell S, Soto C, et al. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc Natl Acad Sci USA* (2013) 110:E4088–97. doi:10.1073/pnas.1306262110
13. Wu X, Zhang Z, Schramm CA, Joyce MG, Kwon YD, Zhou T, et al. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* (2015) 161:470–85. doi:10.1016/j.cell.2015.03.004
14. Sok D, Pauthner M, Briney B, Lee JH, Saye-Francisco KL, Hsueh J, et al. A prominent site of antibody vulnerability on HIV envelope incorporates a motif associated with CCR5 binding and its camouflaging glycans. *Immunity* (2016) 45:31–45. doi:10.1016/j.immuni.2016.06.026
15. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci Rep* (2014) 4:6778–6778. doi:10.1038/srep06778
16. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* (2014) 509:55–62. doi:10.1038/nature13036
17. Kong L, Ju B, Chen Y, He L, Ren L, Liu J, et al. Key gp120 glycans pose roadblocks to the rapid development of VRC01-class antibodies in an HIV-1-infected Chinese donor. *Immunity* (2016) 44:939–50. doi:10.1016/j.immuni.2016.03.006
18. Sok D, Laserson U, Laserson J, Liu Y, Vigneault F, Julien JP, et al. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog* (2013) 9:e1003754. doi:10.1371/journal.ppat.1003754
19. MacLeod DT, Choi NM, Briney B, Garces F, Ver LS, Landais E, et al. Early antibody lineage diversification and independent limb maturation lead to broad HIV-1 neutralization targeting the Env high-mannose patch. *Immunity* (2016) 44:1215–26. doi:10.1016/j.immuni.2016.04.016
20. Gao F, Bonsignori M, Liao HX, Kumar A, Xia SM, Lu X, et al. Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. *Cell* (2014) 158:481–91. doi:10.1016/j.cell.2014.06.022
21. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* (2013) 496:469–76. doi:10.1038/nature12053
22. Bonsignori M, Zhou T, Sheng Z, Chen L, Gao F, Joyce MG, et al. Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. *Cell* (2016) 165:449–63. doi:10.1016/j.cell.2016.02.022
23. Escolano A, Steichen JM, Dosenovic P, Kulp DW, Golijanin J, Sok D, et al. Sequential immunization elicits broadly neutralizing anti-HIV-1 antibodies in Ig knockin mice. *Cell* (2016) 166:1445–58. doi:10.1016/j.cell.2016.07.030

24. Briney B, Sok D, Jardine JG, Kulp DW, Skog P, Menis S, et al. Tailored immunogens direct affinity maturation toward HIV neutralizing antibodies. *Cell* (2016) 166:1459–70. doi:10.1016/j.cell.2016.08.005
25. Jardine J, Julien JP, Menis S, Ota T, Kalyuzhniy O, McGuire A, et al. Rational HIV immunogen design to target specific germline B cell receptors. *Science* (2013) 340:711–6. doi:10.1126/science.1234150
26. Jardine JG, Sok D, Julien JP, Briney B, Sarkar A, Liang CH, et al. Minimally mutated HIV-1 broadly neutralizing antibodies to guide reductionist vaccine design. *PLoS Pathog* (2016) 12(8):e1005815. doi:10.1371/journal.ppat.1005815
27. McGuire AT, Gray MD, Dosenovic P, Gitlin AD, Freund NT, Petersen J, et al. Specifically modified Env immunogens activate B-cell precursors of broadly neutralizing HIV-1 antibodies in transgenic mice. *Nat Commun* (2016) 7:10618. doi:10.1038/ncomms10618
28. McGuire AT, Hoot S, Dreyer AM, Lippy A, Stuart A, Cohen KW, et al. Engineering HIV envelope protein to activate germline B cell receptors of broadly neutralizing anti-CD4 binding site antibodies. *J Exp Med* (2013) 210:655–63. doi:10.1084/jem.20122824
29. Tian M, Cheng C, Chen X, Duan H, Cheng HL, Dao M, et al. Induction of HIV neutralizing antibody lineages in mice with diverse precursor repertoires. *Cell* (2016) 166:1471–84. doi:10.1016/j.cell.2016.07.029
30. Jardine JG, Ota T, Sok D, Pauthner M, Kulp DW, Kalyuzhniy O, et al. Priming a broadly neutralizing antibody response to HIV-1 using a germline-targeting immunogen. *Science* (2015) 349:156–61. doi:10.1126/science.aac5894
31. Steichen JM, Kulp DW, Tokatlian T, Escolano A, Dosenovic P, Stanfield RL, et al. HIV vaccine design to target germline precursors of glycan-dependent broadly neutralizing antibodies. *Immunity* (2016) 45:483–96. doi:10.1016/j.immuni.2016.08.016
32. Doores KJ. The HIV glycan shield as a target for broadly neutralizing antibodies. *FEBS J* (2015) 282:4679–91. doi:10.1111/febs.13530
33. Crispin M, Doores KJ. Targeting host-derived glycans on enveloped viruses for antibody-based vaccine design. *Curr Opin Virol* (2015) 11:63–9. doi:10.1016/j.coviro.2015.02.002
34. Walker LM, Huber M, Doores KJ, Falkowska E, Pejchal R, Julien JP, et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* (2011) 477:466–70. doi:10.1038/nature10373
35. Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, Goss JL, et al. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* (2009) 326:285–9. doi:10.1126/science.1178746
36. Mouquet H, Scharf L, Euler Z, Liu Y, Eden C, Scheid JF, et al. Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proc Natl Acad Sci USA* (2012) 109:E3268–77. doi:10.1073/pnas.1217207109
37. Julien JP, Sok D, Khayat R, Lee JH, Doores KJ, Walker LM, et al. Broadly neutralizing antibody PGT121 allosterically modulates CD4 binding via recognition of the HIV-1 gp120 V3 base and multiple surrounding glycans. *PLoS Pathog* (2013) 9:e1003342. doi:10.1371/journal.ppat.1003342
38. Kong L, Lee JH, Doores KJ, Murin CD, Julien JP, McBride R, et al. Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. *Nat Struct Mol Biol* (2013) 20:796–803. doi:10.1038/nsmb.2594
39. Doores KJ, Kong L, Krumm SA, Le KM, Sok D, Laserson U, et al. Two classes of broadly neutralizing antibodies within a single lineage directed to the high-mannose patch of HIV envelope. *J Virol* (2015) 89:1105–18. doi:10.1128/jvi02905-14
40. Sok D, Doores KJ, Briney B, Le KM, Saye-Francisco KL, Ramos A, et al. Promiscuous glycan site recognition by antibodies to the high-mannose patch of gp120 broadens neutralization of HIV. *Sci Transl Med* (2014) 6:236ra63. doi:10.1126/scitranslmed.3008104
41. Garces F, Sok D, Kong L, McBride R, Kim HJ, Saye-Francisco KF, et al. Structural evolution of glycan recognition by a family of potent HIV antibodies. *Cell* (2014) 159:69–79. doi:10.1016/j.cell.2014.09.009
42. Garces F, Lee JH, de Val N, de la Pena AT, Kong L, Puchades C, et al. Affinity maturation of a potent family of HIV antibodies is primarily focused on accommodating or avoiding glycans. *Immunity* (2015) 43:1053–63. doi:10.1016/j.immuni.2015.11.007
43. Barouch DH, Whitney JB, Moldt B, Klein F, Oliveira TY, Liu J, et al. Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* (2013) 503:224–8. doi:10.1038/nature12744
44. Caskey M, Schoofs T, Gruell H, Settler A, Karagounis T, Kreider EF, et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nat Med* (2017) 23:185–91. doi:10.1038/nm4268
45. Gristick HB, von Boehmer L, West AP Jr, Schamber M, Gazumyan A, Golijanin J, et al. Natively glycosylated HIV-1 Env structure reveals new mode for antibody recognition of the CD4-binding site. *Nat Struct Mol Biol* (2016) 23:906–15. doi:10.1038/nsmb.3291
46. Morris CD, Azadnia P, de Val N, Vora N, Honda A, Giang E, et al. Differential antibody responses to conserved HIV-1 neutralizing epitopes in the context of multivalent scaffolds and native-like gp140 trimers. *mBio* (2017) 8:e36–17. doi:10.1128/mBio.00036-17
47. Kong L, He L, de Val N, Vora N, Morris CD, Azadnia P, et al. Uncleaved prefusion-optimized gp140 trimers derived from analysis of HIV-1 envelope metastability. *Nat Commun* (2016) 7:12040. doi:10.1038/ncomms12040
48. He L, de Val N, Morris CD, Vora N, Thinnes TC, Kong L, et al. Presenting native-like trimeric HIV-1 antigens with self-assembling nanoparticles. *Nat Commun* (2016) 7:12041. doi:10.1038/ncomms12041
49. Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, et al. Automated molecular microscopy: the new Leginon system. *J Struct Biol* (2005) 151:41–60. doi:10.1016/j.jsb.2005.03.010
50. Lander GC, Stagg SM, Voss NR, Cheng A, Fellmann D, Pulokas J, et al. Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J Struct Biol* (2009) 166:95–102. doi:10.1016/j.jsb.2009.01.002
51. Sorzano CO, Bilbao-Castro JR, Shkolnisky Y, Alcorlo M, Melero R, Caffarena-Fernández G, et al. A clustering approach to multireference alignment of single-particle projections in electron microscopy. *J Struct Biol* (2010) 171:197–206. doi:10.1016/j.jsb.2010.03.011
52. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, et al. EMAN2: an extensible image processing suite for electron microscopy. *J Struct Biol* (2007) 157:38–46. doi:10.1016/j.jsb.2006.05.009
53. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF chimera – A visualization system for exploratory research and analysis. *J Comput Chem* (2004) 25:1605–12. doi:10.1002/jcc.20084
54. Zhu Z, Dimitrov DS. Construction of a large naïve human phage-displayed Fab library through one-step cloning. *Methods Mol Biol* (2009) 525:129–42. doi:10.1007/978-1-59745-554-1_6
55. Kwong PD, Chuang GY, DeKosky BJ, Gindin T, Georgiev IS, Lemmin T, et al. Antibodyomics: bioinformatics technologies for understanding B-cell immunity to HIV-1. *Immunol Rev* (2017) 275:108–28. doi:10.1111/imr.12480
56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* (1990) 215:403–10. doi:10.1006/jmbi.1990.9999
57. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* (1988) 85:2444–8. doi:10.1073/pnas.85.8.2444
58. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. *Bioinformatics* (2007) 23:2947–8. doi:10.1093/bioinformatics/btm404
59. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* (2004) 32:1792–7. doi:10.1093/nar/gkh340
60. Li WZ, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* (2006) 22:1658–9. doi:10.1093/bioinformatics/btl158
61. Wyatt R, Sodroski J. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* (1998) 280:1884–8. doi:10.1126/science.280.5371.1884
62. Khayat R, Lee JH, Julien JP, Cupo A, Klasse PJ, Sanders RW, et al. Structural characterization of cleaved, soluble HIV-1 envelope glycoprotein trimers. *J Virol* (2013) 87:9865–72. doi:10.1128/jvi.01222-13
63. Klasse PJ, Depetris RS, Pejchal R, Julien JP, Khayat R, Lee JH, et al. Influences on trimerization and aggregation of soluble, cleaved HIV-1 SOSIP envelope glycoprotein. *J Virol* (2013) 87:9873–85. doi:10.1128/jvi.01226-13
64. Ringe RP, Sanders RW, Yasmeen A, Kim HJ, Lee JH, Cupo A, et al. Cleavage strongly influences whether soluble HIV-1 envelope glycoprotein trimers adopt a native-like conformation. *Proc Natl Acad Sci USA* (2013) 110:18256–61. doi:10.1073/pnas.1314351110
65. Sanders RW, Derking R, Cupo A, Julien JP, Yasmeen A, de Val N, et al. A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. *PLoS Pathog* (2013) 9:e1003618. doi:10.1371/journal.ppat.1003618

66. Sanders RW, Moore JP. HIV: a stamp on the envelope. *Nature* (2014) 514:437–8. doi:10.1038/nature13926
67. Yasmeen A, Ringe R, Derking R, Cupo A, Julien JP, Burton DR, et al. Differential binding of neutralizing and non-neutralizing antibodies to native-like soluble HIV-1 Env trimers, uncleaved Env proteins, and monomeric subunits. *Retrovirology* (2014) 11:41. doi:10.1186/1742-4690-11-41
68. Derking R, Ozorowski G, Sliepen K, Yasmeen A, Cupo A, Torres JL, et al. Comprehensive antigenic map of a cleaved soluble HIV-1 envelope trimer. *PLoS Pathog* (2015) 11:e1004767. doi:10.1371/journal.ppat.1004767
69. Julien JP, Cupo A, Sok D, Stanfield RL, Lyumkis D, Deller MC, et al. Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* (2013) 342:1477–83. doi:10.1126/science.1245625
70. Lyumkis D, Julien JP, de Val N, Cupo A, Potter CS, Klasse PJ, et al. Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* (2013) 342:1484–90. doi:10.1126/science.1245627
71. Pancera M, Zhou T, Druz A, Georgiev IS, Soto C, Gorman J, et al. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* (2014) 514:455–61. doi:10.1038/nature13808
72. Kwon YD, Pancera M, Acharya P, Georgiev IS, Crooks ET, Gorman J, et al. Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 Env. *Nat Struct Mol Biol* (2015) 22:522–31. doi:10.1038/nsmb.3051
73. Stewart-Jones GB, Soto C, Lemmin T, Chuang GY, Druz A, Kong R, et al. Trimeric HIV-1-Env structures define glycan shields from clades A, B, and G. *Cell* (2016) 165:813–26. doi:10.1016/j.cell.2016.04.010
74. Lee JH, Ozorowski G, Ward AB. Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. *Science* (2016) 351:1043–8. doi:10.1126/science.aad2450
75. Lee JH, Andrabi R, Su CY, Yasmeen A, Julien JP, Kong L, et al. A broadly neutralizing antibody targets the dynamic HIV envelope trimer apex via a long, rigidified, and anionic beta-hairpin structure. *Immunity* (2017) 46:690–702. doi:10.1016/j.immuni.2017.03.017
76. Guenaga J, Dubrovskaya V, de Val N, Sharma SK, Carrette B, Ward AB, et al. Structure-guided redesign increases the propensity of HIV Env to generate highly stable soluble trimers. *J Virol* (2015) 90:2806–17. doi:10.1128/jvi.02652-15
77. Sharma SK, de Val N, Bale S, Guenaga J, Tran K, Feng Y, et al. Cleavage-independent HIV-1 Env trimers engineered as soluble native spike mimetics for vaccine design. *Cell Rep* (2015) 11:539–50. doi:10.1016/j.celrep.2015.03.047
78. Guenaga J, Garces F, de Val N, Stanfield RL, Dubrovskaya V, Higgins B, et al. Glycine substitution at helix-to-coil transitions facilitates the structural determination of a stabilized subtype C HIV envelope glycoprotein. *Immunity* (2017) 46:792–803. doi:10.1016/j.immuni.2017.04.014
79. Georgiev IS, Joyce MG, Yang Y, Sastry M, Zhang B, Baxa U, et al. Single-chain soluble BG505.SOSIP gp140 trimers as structural and antigenic mimics of mature closed HIV-1 Env. *J Virol* (2015) 89:5318–29. doi:10.1128/jvi.03451-14
80. Sok D, van Gils MJ, Pauthner M, Julien JP, Saye-Francisco KL, Hsueh J, et al. Recombinant HIV envelope trimer selects for quaternary-dependent antibodies targeting the trimer apex. *Proc Natl Acad Sci USA* (2014) 111:17624–9. doi:10.1073/pnas.1415789111
81. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* (2010) 329:856–61. doi:10.1126/science.1187659
82. Falkowska E, Le KM, Ramos A, Doores KJ, Lee JH, Blattner C, et al. Broadly neutralizing HIV antibodies define a glycan-dependent epitope on the prefusion conformation of gp41 on cleaved envelope trimers. *Immunity* (2014) 40:657–68. doi:10.1016/j.immuni.2014.04.009
83. Huang J, Kang BH, Pancera M, Lee JH, Tong T, Feng Y, et al. Broad and potent HIV-1 neutralization by a human antibody that binds the gp41–gp120 interface. *Nature* (2014) 515:138–42. doi:10.1038/nature136
84. Doria-Rose NA, Bhiman JN, Roark RS, Schramm CA, Gorman J, Chuang GY, et al. New member of the V1V2-directed CAP256-VRC26 lineage that shows increased breadth and exceptional potency. *J Virol* (2016) 90:76–91. doi:10.1128/jvi.01791-15
85. Kong R, Xu K, Zhou T, Acharya P, Lemmin T, Liu K, et al. Fusion peptide of HIV-1 as a site of vulnerability to neutralizing antibody. *Science* (2016) 352:828–33. doi:10.1126/science.aae0474
86. Sanders RW, Moore JP. Native-like Env trimers as a platform for HIV-1 vaccine design. *Immunol Rev* (2017) 275:161–82. doi:10.1111/imr.12481
87. Ward AB, Wilson IA. The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunol Rev* (2017) 275:21–32. doi:10.1111/imr.12507
88. Smith GP. Filamentous fusion phage – novel expression vectors that display cloned antigens on the virion surface. *Science* (1985) 228:1315–7. doi:10.1126/science.4001944
89. Huse WD, Sastry L, Iverson SA, Kang AS, Alting-Mees M, Burton DR, et al. Generation of a large combinatorial library of the immunoglobulin repertoire in phage lambda. *Science* (1989) 246:1275–81. doi:10.1126/science.2531466
90. Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR. Making antibodies by phage display technology. *Annu Rev Immunol* (1994) 12:433–55. doi:10.1146/annurev.iy.12.040194.002245
91. Rader C, Barbas CF. Phage display of combinatorial antibody libraries. *Curr Opin Biotechnol* (1997) 8:503–8. doi:10.1016/s0958-1669(97)80075-4
92. Hoogenboom HR, de Bruïne AP, Hufton SE, Hoet RM, Arends JW, Roovers RC. Antibody phage display technology and its applications. *Immunotechnology* (1998) 4:1–20. doi:10.1016/s1380-2933(98)00007-4
93. Kretzschmar T, von Ruden T. Antibody discovery: phage display. *Curr Opin Biotechnol* (2002) 13:598–602. doi:10.1016/s0958-1669(02)00380-4
94. Bradbury ARM, Marks JD. Antibodies from phage antibody libraries. *J Immunol Methods* (2004) 290:29–49. doi:10.1016/j.jim.2004.04.007
95. Burton DR, Pyati J, Koduri R, Sharp SJ, Thornton GB, Parren PW, et al. Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* (1994) 266:1024–7. doi:10.1126/science.7973652
96. McCoy LE, Burton DR. Identification and specificity of broadly neutralizing antibodies against HIV. *Immunol Rev* (2017) 275:11–20. doi:10.1111/imr.12484
97. Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods* (2013) 60:99–110. doi:10.1016/j.ymeth.2013.03.001
98. Xie J, Yea K, Zhang H, Moldt B, He L, Zhu J, et al. Prevention of cell death by antibodies selected from intracellular combinatorial libraries. *Chem Biol* (2014) 21:274–83. doi:10.1016/j.chembiol.2013.12.006
99. Hammers CM, Stanley JR. Antibody phage display: technique and applications. *J Invest Dermatol* (2014) 134(2):e17. doi:10.1038/jid.2013.521
100. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, et al. Deep sequencing in library selection projects: what insight does it bring? *Curr Opin Struct Biol* (2015) 33:146–60. doi:10.1016/j.sbi.2015.09.001
101. Dai K, He L, Khan SN, O'Dell S, McKee K, Tran K, et al. Rhesus macaque B-cell responses to an HIV-1 trimer vaccine revealed by unbiased longitudinal repertoire analysis. *mBio* (2015) 6:e1375–1315. doi:10.1128/mBio.01375-15
102. Hu JK, Crampton JC, Cupo A, Ketas T, van Gils MJ, Sliepen K, et al. Murine antibody responses to cleaved soluble HIV-1 envelope trimers are highly restricted in specificity. *J Virol* (2015) 89:10383–98. doi:10.1128/jvi.01653-15
103. Haynes BF, Mascola JR. The quest for an antibody-based HIV vaccine. *Immunol Rev* (2017) 275:5–10. doi:10.1111/imr.12517
104. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31:166–9. doi:10.1038/nbt.2492
105. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2015) 21:86–91. doi:10.1038/nm.3743

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.

Copyright © 2017 He, Lin, de Val, Saye-Francisco, Mann, Augst, Morris, Azadnia, Zhou, Sok, Ozorowski, Ward, Burton and Zhu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

METHODS
published: 04 September 2017
doi: 10.3389/fimmu.2017.01072

A Streamlined Approach to Antibody Novel Germline Allele Prediction and Validation

*Ben S. Wendel[1], Chenfeng He[2], Peter D. Crompton[3], Susan K. Pierce[3] and Ning Jiang[2,4]**

[1] McKetta Department of Chemical Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, United States, [2] Department of Biomedical Engineering, Cockrell School of Engineering, The University of Texas at Austin, Austin, TX, United States, [3] Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, United States, [4] College of Natural Sciences, Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX, United States

Advancements in high-throughput sequencing and molecular identifier-based error correction have opened the door to antibody repertoire sequencing with single mutation precision, increasing both the breadth and depth of immune response characterization. However, improvements in sequencing technology cannot resolve one key aspect of antibody repertoire sequencing accuracy: the possibility of undocumented novel germline alleles. Somatic hypermutation (SHM) calling requires a reference germline sequence, and the antibody variable region gene alleles collected by the IMGT database, although large in number, are not comprehensive. Mismatches, resulted from single nucleotide polymorphisms or other genetic variation, between the true germline sequence and the closest IMGT allele can inflate SHM counts, leading to inaccurate antibody repertoire analysis. Here, we developed a streamlined approach to novel allele prediction and validation using bulk PBMC antibody repertoire sequencing data and targeted genomic DNA amplification and sequencing using PBMCs from only 4 ml of blood to quickly and effectively improve the fidelity of antibody repertoire analysis. This approach establishes a framework for comprehensively annotating novel alleles using a small amount of blood sample, which is extremely useful in studying young children's immune systems.

Keywords: antibody, B cell, immune repertoire sequencing, *IGHV*, novel germline allele, polymorphism

OPEN ACCESS

Edited by:
Jacob Glanville, Distributed Bio, United States

Reviewed by:
To-Ha Thai, Harvard Medical School, United States
Paulo Vieira, Institut Pasteur de Paris, France

***Correspondence:**
Ning Jiang
jiang@austin.utexas.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *04 May 2017*
Accepted: *17 August 2017*
Published: *04 September 2017*

Citation:
Wendel BS, He C, Crompton PD, Pierce SK and Jiang N (2017) A Streamlined Approach to Antibody Novel Germline Allele Prediction and Validation. Front. Immunol. 8:1072. doi: 10.3389/fimmu.2017.01072

INTRODUCTION

V(D)J recombination and non-template nucleotide insertion in the junction regions generate the first level of antibody repertoire diversity. During an immune response, B cells that are activated by binding their matching antigens go through a clonal expansion process accompanied by somatic hypermutations (SHMs) that are quasi-randomly introduced to the antibody genes. These mutated antibodies are then selected based on binding strength to the antigen, leading to a second generation of higher affinity antibodies (1–3).

Antibody repertoire SHM patterns have been implicated in a wide range of applications, from the development of broadly neutralizing antibodies against HIV to the diminished effectiveness of vaccines in elderly subjects (4–6). The recent incorporation of molecular barcodes into high-throughput immune repertoire sequencing has improved the ability to discern individual SHMs from PCR and sequencing errors (7, 8); however, accurate SHM calling requires an accurate set

of reference germline sequences to align to. The polygenic and polyallelic nature of the variable domain locus confounds this issue. Currently, 259 functional human antibody heavy chain V gene alleles listed in the IMGT database can be broken into seven subfamilies that likely share common evolutionary ancestors based on sequence similarity (9), but recent studies have shown that individuals often carry novel alleles that have yet to be characterized in the IMGT database (10–12). These novel alleles can be problematic for antibody repertoire analysis because single nucleotide polymorphisms (SNPs) between the novel alleles and the nearest IMGT alleles will instead be counted as SHMs on every sequence utilizing that allele, inflating the SHM load and skewing the SHM patterns. Although there are several software tools (11, 12) to predict the existence of novel alleles, a simple method for novel allele prediction and validation is lacking, especially using a small amount of blood samples.

Here, we report a streamlined method for predicting novel alleles from bulk antibody repertoire data and validating them by sequencing unrecombined genomic DNA (gDNA) from non-B cells. This method can be applied to PBMCs and B cells purified from as little as 4 ml of blood. Six novel alleles across eight different subjects from a larger, ongoing malaria study cohort (13) were predicted and validated with perfect congruency between the expressed repertoire and gDNA. This method can quickly and easily be applied to any antibody repertoire data to mitigate the effects of germline mismatches on SHM patterns.

MATERIALS AND METHODS

Study Design and Cohort

PBMC samples from eight residents of Kalifabougou, Mali were collected from an ongoing malaria cohort study (13). Up to five million PBMCs were directly lysed for antibody repertoire sequencing, and T cells were FACS-sorted from the remaining PBMCs for unrecombined gDNA validation. 6 predicted novel alleles were chosen for validation. The Ethics Committee of the Faculty of Medicine, Pharmacy, and Dentistry at the University of Sciences, Technique, and Technology of Bamako and the Institutional Review Board of the National Institute of Allergy and Infectious Diseases, National Institutes of Health approved the malaria study, from which we obtained frozen PBMCs. Written informed consent was obtained from adult participants and from the parents or guardians of participating children. The study is registered in the www.ClinicalTrials.gov database (NCT01322581).

Antibody Repertoire Sequencing and Novel Allele Prediction

Antibody repertoire sequencing was performed as previously described (7, 8) with some modifications. Novel allele prediction schematic is summarized in **Figure 1A**. In short, unique IgM sequences from bulk PBMC samples were used to minimize the effects of SHM and clonal expansion, as they are more likely to be derived from naïve B cells and thus have fewer SHMs than other antibody isotypes. These sequences were first aligned to the reference germline allele database (e.g., IMGT) and assigned to the best-matched alleles. The ratios of perfectly matched sequences to those with 1, 2, 3, and 4 mismatches (putative SNPs) compared to the reference germline were determined. Ratios of less than 2 to 1 were then inspected for identical mutation patterns. If identical mutations were present in at least 20% of the unique sequences, with less than 2% of the sequences harboring different mutations at the same positions, the allele containing those SNPs was flagged as a possible novel allele.

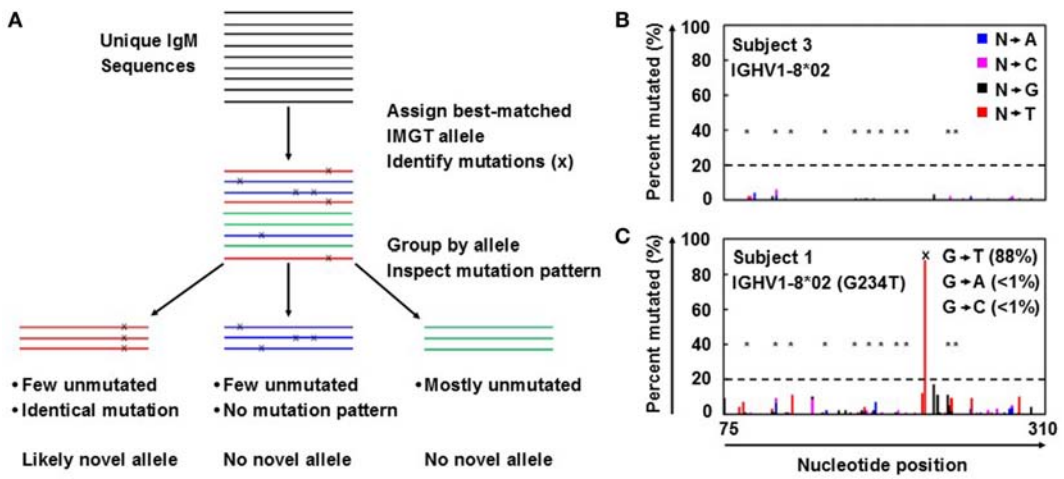


FIGURE 1 | Novel germline allele prediction from bulk repertoire sequencing data. **(A)** Overview of novel allele prediction schematic from bulk repertoire sequencing data. Color indicates best-matched IMGT reference germline allele assignment; x indicates single nucleotide polymorphisms (SNP) to reference germline allele. **(B,C)** Representative percent of unique IgM sequences mutated for each position along the sequence for the absence **(B)** and presence **(C)** of a novel allele resulted from SNP(s). Color indicates the nucleotide substitution: A (blue), C (pink), G (black), and T (red). *Indicates somatic hypermutation hotspots; dashed line indicates prediction threshold of SNP calling; x indicates SNP on predicted novel allele compared to the closest IMGT reference germline allele. SNP is broken down by nucleotide substitution as indicated in inset in **(C)**.

gDNA Sequencing and Reads Processing

Nested PCR was used to reduce non-specific amplification. Primers were designed such that the inner primers were no fewer than 14 bases away from the locations of the predicted IMGT/novel allele mismatches. Inner primers were fused to partial Illumina adaptors, and a third PCR was performed to add the full adaptor sequence (Table S1 in Supplementary Material). First PCR was performed on 10% of purified gDNA from 2,000 sorted T cells using Phusion Hot Start II DNA Polymerase (Thermo Scientific) with the following protocol: 98°C for 1 min; 10 cycles of 98°C for 30 s, 57°C for 1 min, and 72°C for 5 min; then 72°C for 10 min. Second PCR was performed on 10% of the first PCR product with the same protocol. Final adaptor ligation was performed on 10% of the second PCR product using TaKaRa Ex Taq DNA Polymerase Hot Start with the following protocol: 95°C for 3 min; 10 cycles of 95°C for 30 s, 57°C for 30 s, and 72°C for 2 min; then 72°C for 7 min. Libraries were pooled, gel-purified, and sequenced *via* Miseq 2 × 250 PE.

Sequencing reads were first merged using the SeqPrep tool.[1] IgBlast (14) was then used to align the reads to the established IMGT germline allele database. Reads mapping to the nearest germline allele to the novel allele of interested were filtered. Reads matching exactly to the IMGT germline allele or the novel allele sequence were tallied. If the exact novel allele sequence was found in 20% or more of the tallied reads, the sample was considered a positive hit.

Data Availability

Antibody repertoire sequencing data can be found in dbGaP under the accession number phs001209.v1.p1. gDNA sequencing data can be found in SRA under the accession number SRP112759.

[1]https://github.com/jstjohn/SeqPrep.

RESULTS

Bulk Antibody Repertoire Novel Allele Prediction

Antibody repertoire sequencing data from bulk PBMCs were collected and processed as described in Section "Antibody Repertoire Sequencing and Novel Allele Prediction" and summarized in **Figure 1A**. IgM sequences were used to calculate the mutation distribution by position because IgM is mostly expressed on naïve B cells that have not been activated and have fewer SHMs compared to other isotypes. As expected, the percentage of unique sequences mutated at each position in IgM is low, even for SHM hotspots (**Figure 1B**). However, a large spike at one (or more) specific position(s) could indicate the presence of a novel allele resulted from SNP(s) (**Figure 1C**).

A threshold of 20% of unique sequences harboring the identical predicted SNP(s) was applied to determine a positive hit on novel allele (**Figures 1B,C**, dashed horizontal line). Several *IGHV* genes, e.g., *IGHV1-69* and *IGHV3-30*, have copy-number variants (CNVs) that arose from chromosomal segmental duplication and insertion/deletion events, leading to a diploid copy number ranging from 0 to 4 alleles present for a given gene for an individual (15, 16). For genes with up to 4 copies, this 20% threshold can account for a heterozygous genotype with 1 of 4 copies being the novel allele, which should have a 25:75 split on the usage of these four alleles. Six genes with predicted novel alleles were chosen for validation (**Table 1**, column headers). Novel alleles were named according to the nearest IMGT allele followed by the substitution(s) in parenthesis, e.g., novel allele IGHV1-8*02 (G234T) has the same sequence as the IMGT allele IGHV1-8*02 with the G at position 234 substituted with a T. The full novel allele sequences can be found in **Table 2**. These novel alleles were also predicted independently using TIgGER (11), another novel germline allele detection tool. Overall, 17 positive novel allele hits were predicted from the 6 genes across the 8 subjects.

TABLE 1 | Summary of genomic DNA (gDNA) validation of novel alleles predicted by bulk repertoire sequencing data.

Subjects	Novel alleles					
	IGHV1-8*02 (G234T)	IGHV1-69*01 (G163A)	IGHV3-30*02 (T201C)	IGHV4-31*02 (C198T)	IGHV4-59*01 (T109C)	IGHV4-61*01 (C93T_C136G_A138C)
1	+/+	+/+	+/+[a]	–/–	+/+	–/–
2	–/N.D.	–/–	–/–	+/+	–/–	–/–[b]
3	–/–	+/+[a]	–/–	–/–	+/+	+/+
4	+/+	–/–	–/–[a]	–/–	–/–	–/–
5	–/–	–/–	–/–	+/+	–/–	–/–
6	+/+	–/–[a]	+/+	–/–[b]	+/+	–/–[b]
7	+/+	–/–	+/+[a]	–/–[b]	–/–	–/–
8	+/+	+/+	–/–	–/–[b]	–/–	–/–

+/+ (dark green) indicates positive in both the bulk repertoire and the gDNA data for predicted single nucleotide polymorphisms (SNPs); –/– (light green) indicates negative in both the bulk repertoire and the gDNA data for predicted SNPs; –/N.D. (yellow) indicates negative in bulk repertoire data but gDNA failed to amplify during gDNA validation for predicted SNPs.
[a]Indicates the existence of copy-number variants with more than two alleles detected in the gDNA data that belong to the same gene.
[b]Indicates the gene was not detected in the repertoire or gDNA, possibly due to gene deletion.

TABLE 2 | Novel allele sequences.

Novel allele	Sequence
IGHV1-8*02 (G234T)	CAGGTGCAGCTGGTGCAGTCTGGGGCTGAGGTGAAGAAGCCTGGGGCCTCAGTGAAGGTC TCCTGCAAGGCTTCTGGATACACCTTCACCAGCTATGATATCAACTGGGTGCGACAGGCC ACTGGACAAGGGCTTGAGTGGATGGGATGGATGAACCCTAACAGTGGTAACACAGGCTAT GCACAGAAGTTCCAGGGCAGAGTCACCATTACCAGGAACACCTCCATAAGCACAGCCTAC ATGGAGCTGAGCAGCCTGAGATCTGAGGACACGGCCGTGTATTACTGTGCGAGAGG
IGHV3-30*02 (T201C)	CAGGTGCAGCTGGTGGAGTCTGGGGGAGGCGTGGTCCAGCCTGGGGGGTCCCTGAGACTC TCCTGTGCAGCGTCTGGATTCACCTTCAGTAGCTATGGCATGCACTGGGTCCGCCAGGCT CCAGGCAAGGGGCTGGAGTGGGTGGCATTTATACGGTATGATGGAAGTAATAAATACTACC GCAGACTCCGTGAAGGGCCGATTCACCATCTCCAGAGACAATTCCAAGAACACGCTGTAT CTGCAAATGAACAGCCTGAGAGCTGAGGACACGGCTGTGTATTACTGTGCGAAAGA
IGHV4-61*01 (C93T_C136G_A138C)	CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCGGAGACCCTGTCCCTC ACCTGCACTGTCTCTGGTGGCTCCGTCAGTAGTGGTAGTTACTACTGGAGCTGGATCCGG CAGCCCGCCGGGAAGGGACTGGAGTGGATTGGGTATATCTATTACAGTGGGAGCACCAAC TACAACCCCTCCCTCAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTC TCCCTGAAGCTGAGCTCTGTGACCGCTGCGGACACGGCCGTGTATTACTGTGCGAGAGA
IGHV4-59*01 (T109C)	CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCGGAGACCCTGTCCCTC ACCTGCACTGTCTCTGGTGGCTCCATCAGTAGTCACTACTGGAGCTGGATCCGGCAGCCC CCAGGGAAGGGACTGGAGTGGATTGGGTATATCTATTACAGTGGGAGCACCAACTACAAC CCCTCCCTCAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTG AAGCTGAGCTCTGTGACCGCTGCGGACACGGCCGTGTATTACTGTGCGAGAGA
IGHV1-69*01 (G163A)	CAGGTGCAGCTGGTGCAGTCTGGGGCTGAGGTGAAGAAGCCTGGGTCCTCGGTGAAGGTC TCCTGCAAGGCTTCTGGAGGCACCTTCAGCAGCTATGCTATCAGCTGGGTGCGACAGGCC CCTGGACAAGGGCTTGAGTGGATGGGAAGGATCATCCCTATCTTTGGTACAGCAAACTAC GCACAGAAGTTCCAGGGCAGAGTCACGATTACCGCGGACGAATCCACGAGCACAGCCTAC ATGGAGCTGAGCAGCCTGAGATCTGAGGACACGGCCGTGTATTACTGTGCGAGAGA
IGHV4-31*02 (C198T)	CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCACAGACCCTGTCCCTC ACCTGTACTGTCTCTGGTGGCTCCATCAGCAGTGGTGGTTACTACTGGAGCTGGATCCGC CAGCACCCAGGGAAGGGCCTGGAGTGGATTGGGTACATCTATTACAGTGGGAGCACCTATT TACAACCCGTCCCTCAAGAGTCGAGTTACCATATCAGTAGACACGTCTAAGAACCAGTTC TCCCTGAAGCTGAGCTCTGTGACTGCCGCGGACACGGCCGTGTATTACTGTGCGAGAGA

Bold, yellow-highlighted bases indicate single nucleotide polymorphisms from the documented IMGT alleles referenced in the name of the novel allele.

gDNA Novel Allele Validation

Genomic DNA purified from FACS-sorted T cells from the same eight subjects was used to validate the presence of the predicted novel alleles as described in Section "gDNA Sequencing and Reads Processing" and summarized in **Figures 2A,B**. Due to the high degree of sequence homology among the V genes, a series of filtering steps was applied to eliminate reads that were distant from the putative novel allele or closest IMGT allele. Finally, the number of reads exactly matching the novel and original IMGT sequences were compared. If 20% of these reads matched the novel allele sequence, the subject was deemed positive for the novel allele. All 17 of the positive hits from the bulk repertoire data returned positive hits from the gDNA, and 30 out of 31 negative hits from the bulk repertoire data that were tested in parallel were also negative in the gDNA, with one library failing to amplify (**Table 1**).

The positive hits in the bulk repertoire data ranged from 22.7 to 99.9% of unique sequences containing the novel mutation, while the negative hits ranged from 0.0 to 1.6% (**Figure 2C**, *X*-axis). This tight range on the negative hits is consistent with the low rate of mutations expected for IgM antibodies. For the gDNA validation, the positive hits ranged from 29.0 to 100% of filtered reads exactly matching the novel sequence, while the negative hits all failed to detect a single novel allele read (**Figure 2C**, *Y*-axis). The densely packed clusters at the bottom left and top right of **Figure 2C** imply that this method is sensitive enough to distinguish between heterozygous and homozygous genotypes, and our threshold of calling a novel allele on both gDNA and repertoire data, which is 20% of reads mapped to either putative allele or its closest IMGT allele, is appropriate.

Another observation that increases confidence in novel germline allele prediction is the detection of the identical novel allele in multiple individuals (12). 5 of the 6 alleles tested were positively validated in two or more subjects (**Table 1**). The lone allele detected in a single individual, *IGHV4-61*01 (C93T_C136G_A138C)*, is three mismatches away from the nearest IMGT germline allele. None of the gDNA reads for this individual matched the reference allele, while all of the filtered reads exactly matched the predicted novel sequence. Additionally, none of the filtered reads from all seven negative subjects tested in parallel matched the novel sequence.

DISCUSSION

We developed a remarkably sensitive yet simple method for detecting and validating novel alleles from bulk antibody repertoire sequencing data. This approach requires little specialized bioinformatics analysis and no unique laboratory equipment or reagents. gDNA validation can be performed on DNA purified from as few as 2,000 FACS-sorted non-B cells, maximizing the proportion of the sample that can be utilized for antibody repertoire analysis.

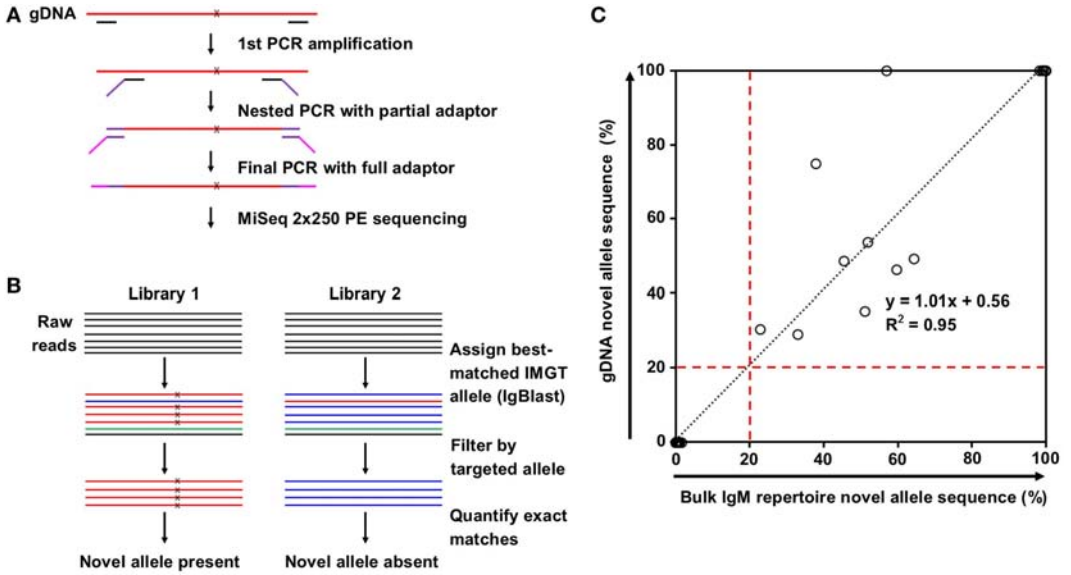


FIGURE 2 | Novel allele validation by targeted genomic DNA (gDNA) sequencing. **(A)** Overview of targeted gDNA amplification and library preparation. x indicates predicted single nucleotide polymorphism (SNP) on novel allele compared to the closest IMGT reference germline allele. **(B)** Overview of genomic DNA (gDNA) sequencing data analysis for the presence (left) and absence (right) of a novel allele resulted from SNP(s). Color indicates best-matched IMGT reference germline allele assignment; x indicates SNP to the closest IMGT reference germline allele. **(C)** Correlation between the percentage of novel allele sequences in bulk IgM repertoire data (%) and the percentage of novel allele sequences in gDNA data (%). Most points are clustered at the origin ($N = 30$) or the top right ($N = 9$). Black dotted line represents the linear regression; red dashed lines indicate the novel allele calling threshold.

In the era of Big Data and immune repertoire sequencing, researchers attempt to mine meaningful associations out of vast swaths of information. Reliable bioinformatics analysis is highly dependent on the quality data being analyzing. With respect to immune repertoire sequencing, great strides have been made toward mitigating sequencing and PCR errors, but even perfectly accurate sequencing data can result in erroneous SHM calling if there are mismatches between the reference germline alleles and the individual's true germline sequence. These systemic errors in SHM calling can propagate into faulty conclusions. For example, the SNP in IGHV4-59*01 (T109C) results in an amino acid substitution of tyrosine to histidine in the CDR1. The relative frequency of SHMs that lead to amino acid changes versus those that do not (replacement versus silent mutations) can be used to gauge antigen selection strength—more replacement mutations than expected indicates positive selection while fewer indicates negative selection (17). Mistakenly adding an extra replacement mutation in the CDR to a large portion of the sequences mapped to this allele could give the appearance of affinity maturation and antigen selection when perhaps no such selection took place. Additionally, tracking the evolution of antibody lineages has led to interesting results, particularly in broadly neutralizing HIV antibodies (18). Mistakenly attributing a germline SNP to a SHM could lead to incorrect root assignment or directionality during lineage tree formation, confounding the results. Therefore, it is important to verify germline allele sequences before performing detailed antibody repertoire analysis.

The combination of antibody repertoire and non-B cell gDNA sequencing allowed for advanced insight into the genotypes of the subjects. Novel allele *IGHV1-69*01 (G163A)* was found to be an exact match with IMGT allele *IGHV1-69*07*, except *IGHV1-69*07* is truncated at both ends. After performing nested PCR, these alleles were indistinguishable. However, in the three subjects predicted to have the novel allele, no unique sequences in the bulk repertoire data mapped to the truncated *IGHV1-69*07* allele; instead, they contained the full length *IGHV1-69*01* sequence with the G163A SNP.

IGHV1-69 and *IGHV1-69D* share common alleles that can range from 2 to 4 copies total on a diploid genome, and *IGHV3-30* and *IGHV3-30-5* share common alleles that can range from 0 to 4 copies total on a diploid genome (15). Interestingly, we detected more than 2 alleles in the gDNA of 2 of 8 subjects for *IGHV1-69/1-69D*, consistent with previous studies on *IGHV1-69* CNV in African populations (16), and more than 2 *IGHV3-30/3-30-5* alleles were detected in 3 of 8 subjects ([a] in **Table 1**). Conversely, *IGHV4-31* and *IGHV4-61* are associated with deletion events yielding 0 to 4 copies, each (15). *IGHV4-31* was not observed in the repertoire or gDNA of 3 of 8 subjects, and *IGHV4-61* was not observed in the repertoire or gDNA of 2 of 8 subjects ([b] in **Table 1**), likely indicating the absence of these genes in these subjects. These results demonstrated the sensitivity of our approach and emphasized the necessity of characterizing individual's own germline alleles in antibody repertoire sequencing studies in order to accurately count the number of SHMs.

The results were highly consistent with all 17 predicted positive hits and 30 of 31 predicted negative hits confirmed in gDNA. One limitation is that this method will only detect novel alleles that are similar to alleles within the IMGT reference database. Additionally, if a CNV results in more than four alleles present in the diploid genome for a given gene in an individual, then our threshold of a putative SNP call, which is least 20% of unique IgM sequences having the same mismatch at the same position,

would not be able to detect the novel allele initially in the antibody repertoire data. However, this is extremely rare based on current knowledge of antibody gene loci (15). In summary, at least 1 novel allele was found in each subject tested, highlighting the need for novel allele detection and correction in antibody repertoire analysis.

ETHICS STATEMENT

The Ethics Committee of the Faculty of Medicine, Pharmacy, and Dentistry at the University of Sciences, Technique, and Technology of Bamako and the Institutional Review Board of the National Institute of Allergy and Infectious Diseases, National Institutes of Health approved the malaria study, from which we obtained frozen PBMCs. Written informed consent was obtained from adult participants and from the parents or guardians of participating children.

AUTHOR CONTRIBUTIONS

BW designed and performed research, analyzed and interpreted data, and wrote the manuscript. CH helped perform data analysis. PC and SP provided samples and helped design research. NJ designed research, directed the study, provided funding, and wrote the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Michael Wilson and Jessica Podnar at the Genomic Sequencing and Analysis Facility at UT Austin for helping with the sequencing runs; Dr. Evan Cohen for helping with cell sorting; and Dr. Pengyu Ren for providing computational resources.

FUNDING

This work was supported by the National Institutes of Health [grants R00AG040149 (NJ) and S10OD020072 (NJ)] and the Welch Foundation [grant F1785 (NJ)]. NJ is a Cancer Prevention and Research Institute of Texas (CPRIT) Scholar and a Damon Runyon-Rachleff Innovator. BW is a recipient of the Thrust 2000—George Sawyer Endowed Graduate Fellowship in Engineering. The cohort study in Mali was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fimmu.2017.01072/full#supplementary-material.

REFERENCES

1. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* (2007) 76:1–22. doi:10.1146/annurev.biochem.76.061705.090740
2. Victora GD, Nussenzweig MC. Germinal centers. *Annu Rev Immunol* (2012) 30:429–57. doi:10.1146/annurev-immunol-020711-075032
3. De Silva NS, Klein U. Dynamics of B cells in germinal centres. *Nat Rev Immunol* (2015) 15(3):137–48. doi:10.1038/nri3804
4. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* (2013) 5(171):171ra119. doi:10.1126/scitranslmed.3004794
5. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol* (2013) 25(5):646–52. doi:10.1016/j.coi.2013.09.017
6. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc Natl Acad Sci U S A* (2013) 110(16):6470–5. doi:10.1073/pnas.1219320110
7. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110(33):13463–8. doi:10.1073/pnas.1312146110
8. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) 11:653–5. doi:10.1038/nmeth.2960
9. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res* (2015) 43(Database issue):D413–22. doi:10.1093/nar/gku1056
10. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) 184(12):6986–92. doi:10.4049/jimmunol.1000445
11. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* (2015) 112(8):E862–70. doi:10.1073/pnas.1417683112
12. Corcoran MM, Phad GE, Nestor VB, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
13. Tran TM, Li S, Doumbo S, Doumtabe D, Huang CY, Dia S, et al. An intensive longitudinal cohort study of Malian children and adults reveals no evidence of acquired immunity to *Plasmodium falciparum* infection. *Clin Infect Dis* (2013) 57(1):40–7. doi:10.1093/cid/cit174
14. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41(Web Server issue):W34–40. doi:10.1093/nar/gkt382
15. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. *Genes Immun* (2012) 13(5):363–73. doi:10.1038/gene.2012.12
16. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) 92(4):530–46. doi:10.1016/j.ajhg.2013.03.004
17. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) 40(17):e134. doi:10.1093/nar/gks457
18. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* (2013) 496(7446):469–76. doi:10.1038/nature12053

Conflict of Interest Statement: NJ is a scientific advisor of ImmuDX LLC. All other authors declare no conflict of interest.

Copyright © 2017 Wendel, He, Crompton, Pierce and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

METHODS
published: 14 September 2017
doi: 10.3389/fimmu.2017.01157

Novel Method for High-Throughput Full-Length IGHV-D-J Sequencing of the Immune Repertoire from Bulk B-Cells with Single-Cell Resolution

*Stefano Vergani[1,2], Ilya Korsunsky[3], Andrea Nicola Mazzarello[1], Gerardo Ferrer[1], Nicholas Chiorazzi[1] and Davide Bagnara[1,4]**

[1] Karches Centre for Chronic Lymphocytic Leukemia Research, The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, United States, [2] Hofstra-Northwell Health School of Medicine, Hempstead, NY, United States, [3] Robert S. Boas Center for Genomics & Human Genetics, The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY, United States, [4] Department of Experimental Medicine, University of Genoa, Genoa, Italy

OPEN ACCESS

Edited by:
Jacob Glanville,
Distributed Bio, United States

Reviewed by:
Michael P Murtaugh,
University of Minnesota,
United States
Johanne Tracey Jacobsen,
Whitehead Institute for
Biomedical Research,
United States

***Correspondence:**
Davide Bagnara
davide.bagnara@edu.unige.it

Specialty section:
This article was submitted
to B Cell Biology,
a section of the journal
Frontiers in Immunology

Received: 18 May 2017
Accepted: 01 September 2017
Published: 14 September 2017

Citation:
Vergani S, Korsunsky I,
Mazzarello AN, Ferrer G, Chiorazzi N
and Bagnara D (2017) Novel Method
for High-Throughput Full-Length
IGHV-D-J Sequencing of the Immune
Repertoire from Bulk B-Cells
with Single-Cell Resolution.
Front. Immunol. 8:1157.
doi: 10.3389/fimmu.2017.01157

Efficient and accurate high-throughput DNA sequencing of the adaptive immune receptor repertoire (AIRR) is necessary to study immune diversity in healthy subjects and disease-related conditions. The high complexity and diversity of the AIRR coupled with the limited amount of starting material, which can compromise identification of the full biological diversity makes such sequencing particularly challenging. AIRR sequencing protocols often fail to fully capture the sampled AIRR diversity, especially for samples containing restricted numbers of B lymphocytes. Here, we describe a library preparation method for immunoglobulin sequencing that results in an exhaustive full-length repertoire where virtually every sampled B-cell is sequenced. This maximizes the likelihood of identifying and quantifying the entire IGHV-D-J repertoire of a sample, including the detection of rearrangements present in only one cell in the starting population. The methodology establishes the importance of circumventing genetic material dilution in the preamplification phases and incorporates the use of certain described concepts: (1) balancing the starting material amount and depth of sequencing, (2) avoiding IGHV gene-specific amplification, and (3) using Unique Molecular Identifier. Together, this methodology is highly efficient, in particular for detecting rare rearrangements in the sampled population and when only a limited amount of starting material is available.

Keywords: next generation sequencing, immunoglobulin repertoire, Illumina Miseq sequencing, VDJ rearrangement, cDNA library, unique molecular identifier, B lymphocytes

INTRODUCTION

The diversity of the adaptive immune system is the key to its ability to respond to a wide variety of antigens. Extensive knowledge of the adaptive immune receptor repertoire (AIRR) could have a major impact on basic and translational research since it can help to better understand the dynamics and diversity of the AIRR, study immune responses induced by vaccines and infectious agents, and determine minimal residual disease, intra-clonal diversity, and evolution in lymphoma/leukemia.

Recent advances in next generation sequencing allow in-depth studies of AIRR of B (Ig-seq) and T lymphocytes.

B lymphocytes originate in the bone marrow where precursors pass through a series of highly regulated processes to generate a functional B-cell receptor that is necessary for the survival of mature B cells (1). In humans, the variable region of the IGH chain is created by the recombination of one of ~50 variable (IGHV) genes, one or more of ~30 diversity (IGHD), and one of 6 joining (IGHJ) genes. Within the recombined IGHV-D-J, the CDR3 is the most variable segment and is the major contributor for antigen contact and the antigen-binding site. Its variability is increased as a consequence of imperfect joining with random nucleotide insertion and deletions occurring during the recombination process. This yields an antigen-inexperienced B cell with a virtually unique IGHV-D-J rearrangement, without a germline reference and, therefore, a characteristic antigen-binding site. Finally, the diversity and complexity of the AIRR obtained by recombination is further increased in secondary lymphoid tissues by another biological process termed somatic hypermutation, whereby the enzyme activation-induced deaminase introduces mutations in the rearranged IGHV-D-J.

Determining the DNA sequence of the AIRR presents major challenges compared to targeted sequencing of other genes because of its linkage, at the mRNA level, to the constant region of IGH. The latter can change with B-lymphocyte maturation, moving from IgM to IgM + IgD to non-IgM (IgG, IgA, and IgE) isotypes. In addition, AIRR DNA sequencing is made even more challenging because of the absence of germline reference for the VH CDR3. This makes reliably reconstructing the sequences from short reads challenging, although certain library preparations have successfully addressed this problem (2). Finally, the extent of *in vivo* repertoire diversity and limited biological sampling (3) pose major problems, especially in human studies. In addition, the full diversity of the already limited sampled material is often not reflected leading to poor repertoire overlapping of technical replicates (4). It is, therefore, important and challenging to obtain a comprehensive repertoire representation of as many sampled cells as possible.

In studies where attention is focused on expanded B-lymphocyte clones, such as an immune response to a specific environmental insult, overrepresented rearrangements can be easily detected. On the other hand, in cases where attention is focused on non-expanded/rare cells (i.e., naïve, immature, and long-term memory B cells, or in detection of minimal residual disease in the context of a B-cell malignancy), the resulting repertoire is far more susceptible to biases intrinsic to methodology. Biases often result from amplicon length in the case of template-switch PCR, and in IGHV gene-specific primer for multiplex PCR. Despite bias correction to obtain quantitative data can be performed (5), it is not possible to recovery rearrangements that have not been detected in the sequencing process. Since high yields are crucial to obtain a reasonable representation, this poses challenges, especially when dealing with B-cell fractions from a limited sample. For example with the template-switch method, efficiency is less than 1 molecule per naïve B cell (6).

In light of these issues, we present a method that allows Ig-seq of the full IGHV-D-J-CH transcript, with single-cell resolution, from a pool of B cells.

RESULTS

Library Preparation

A defined number of B lymphocytes (<100–25,000) were sorted directly into 200 µl PCR tubes containing cell lysis buffer, and mRNA was isolated using poly-T coupled to magnetic beads (**Figure 1**). The entire amount of isolated mRNA was reverse transcribed in this solid phase, with the poly-T stretch working as primer for the reaction. Beads containing the resultant single-stranded cDNA were then purified with a magnet, and cDNA was used for the synthesis of double-strand (ds) cDNA of the IGHV-D-J rearrangements employing multiplex primers annealing to the 5' of the leader sequence (Table S1 in Supplementary Material). During ds-cDNA synthesis, a unique molecular identifier (UMI) consisting of 13–16 random nucleotides and containing in addition a partial Illumina adaptor, were introduced into each second strand of the cDNA. Then, the IGHV-D-J-CH ds-cDNA—purified by a magnet as above—was used to perform PCR amplification with a universal forward primer and a mix of CH isotype-specific reverse primers. The PCR product was used as a template for a semi-nested PCR with inner CH primers that allowed introduction of partial Illumina adaptors, which were used for library indexing.

In each of the above steps until the first PCR amplification, the entirety of mRNA, cDNA, and ds-cDNA was used, never being diluted. After each step, the material obtained was washed while attached to the original set of beads and resuspended directly in the reaction buffer of the following step. This poly-T magnetic bead purification of mRNA and of cDNA (purification for ds-cDNA not tested) was at least 7-times more efficient than a column-based method when tested on a starting material of 100,000 cells (data not shown). Moreover, the column-based method gave inconsistent results when starting from less than 10,000 cells, indicating even lower efficiency (not shown). Finally, our method was easily carried out in 96-well plates. The indexed library was sequenced with Illumina MiSeq v3 (600 cycles).

Raw Sequence Analysis and Error Correction

The defined raw Illumina sequences were analyzed using pRESTO tools (7). Sequences were clustered based on UMI identity (allowing one error in the UMI region). Only those sequences having at least 90% identity in the first 150 nt (region with higher quality) and including the HCDR3 on both Illumina reads from each UMI group (UMIG) were used to build a single consensus sequence; this step compensated for possible errors in the UMI region and for independent molecules that could be tagged erroneously with the same UMI. For analysis, we used sequences for which the consensus was obtained by matching at least three reads or by three identical sequences from different UMIGs.

Effects of Read Clustering

At a sequencing depth of 40× per starting cell for PBMC-derived naïve B cells, analyses were performed as described above with UMI and read clustering with regular UMI grouping solely by identity (**Figure 2**). Upon clustering, the total number of UMIGs

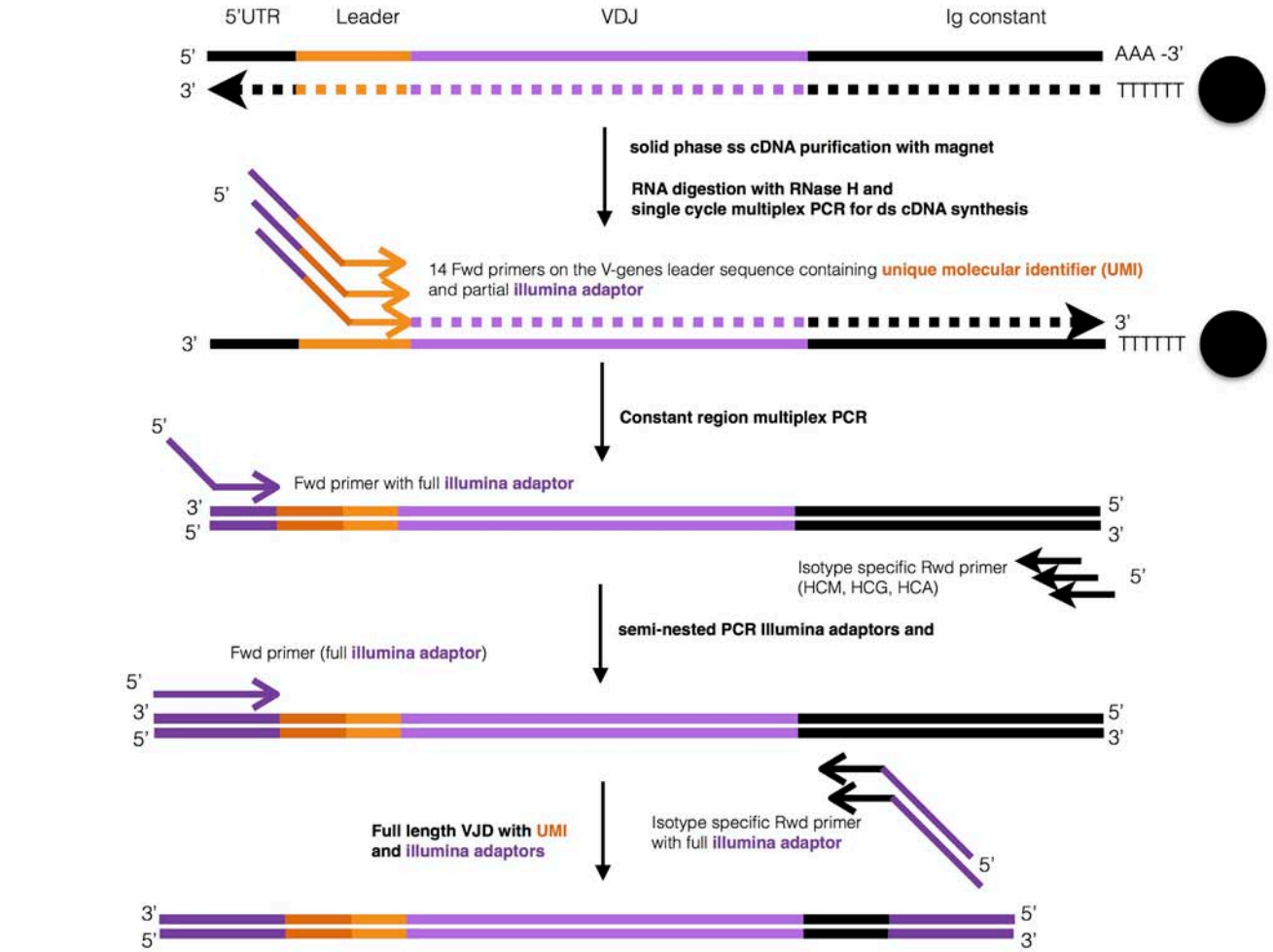


FIGURE 1 | Graphic representation of library preparation. Ig-seq library preparation consists of 4 steps: (1) mRNA purification and reverse transcription, (2) IGH-specific double strand cDNA synthesis with multiplex primers, both while bound to poly-T magnetic beads, () two step semi-nested PCR, and (1) Illumina indexing.

increased more than twofold, the number of UMIGs passing the filter increased by 30%, and the sequences belonging to a group composed of at least three independent reads increased by 300% (**Figure 2A**). After consensus filtering (**Figure 2B**), the mean read count per unique sequence increased from 29 to 31, and the mean UMIG count per unique sequence increased from 4.7 to 7.6; concomitantly, sequences with UMIG counts equal to one decreased from ~6 to ~1.8%. Clustering increased significantly singletons, which result from removing unrelated sequences with the same UMIG, and lead to less ambiguous nucleotide calls in forming the consensus sequence. After paired-end assembly, the presence of sequences containing an N nucleotide derived from the consensus of three or more reads was ~7 times lower when UMI and read clustering were performed (**Figure 2A**).

Measurement of Specific IGHV Gene Detectability

To estimate the ability to detect individual rearrangements containing specific IGHV genes, naïve B-cell repertoires, defined as $CD19^{+}CD27^{-}IgD^{+}CD38^{dim}CD24^{+}$ cells (Figure S1 in Supplementary Material), were analyzed at 40× depth per starting cell. We assumed that the naïve B-cell subpopulation had not yet encountered foreign antigen and hence would have not undergone clonal expansion. Therefore, a unique IGHV-D-J rearrangement would indicate the presence of a single cell in the sample of 40,000 cells analyzed per donor. Hence, we used the UMIG count per unique IGHV-D-J sequence—proxy of the number of mRNA molecules sequenced—as an indicator of the IGHV-specific detectability of the methodology.

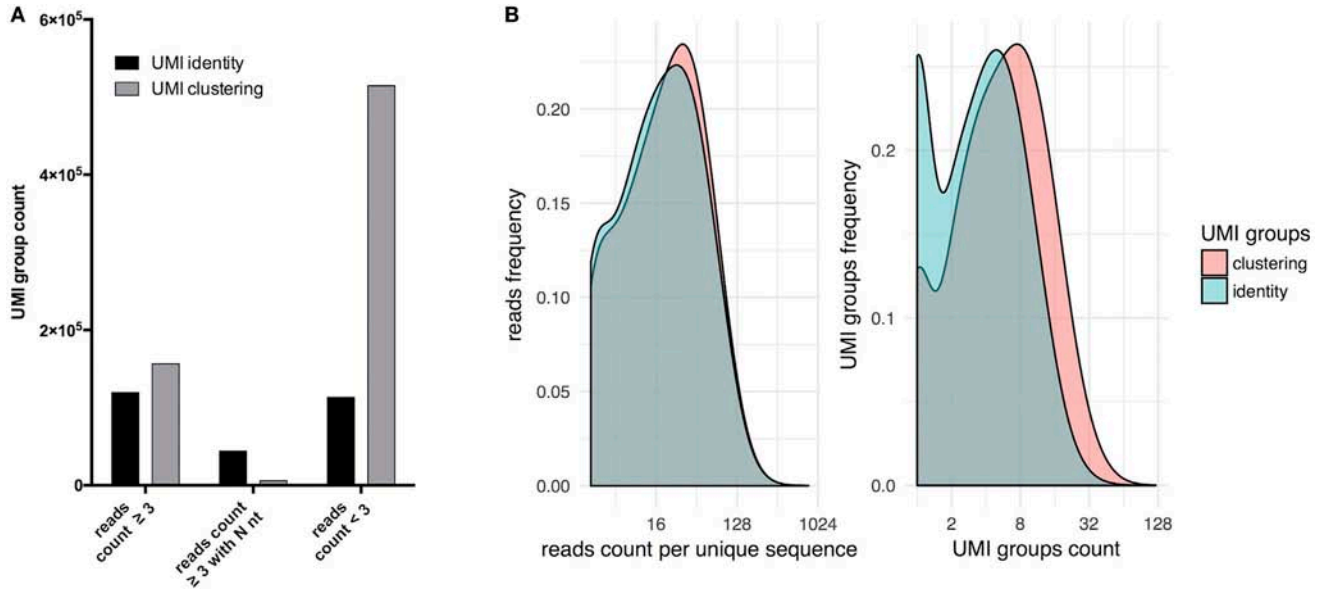


FIGURE 2 | Forming unique molecular identifier (UMI) groups by sequence clustering increases the number of sequences passing filter. **(A)** Allowing one error in the UMI sequence increases the sequences deriving from three or more raw reads compared to the grouping of UMI by read identity. In addition, this leads to a decrease in the number of ambiguous calls (N nt) during consensus assignment. **(B)** Sequences deriving from UMI groups by sequence clustering have higher UMI counts.

To quantify relative IGHV gene detectability, we compared the UMIG distribution of a specific gene to the distribution of all genes in the corresponding sample; this was done using a robust, non-parametric quantile-based method (see Methods) to estimate gene-specific detectability (**Figure 3A**). This indicated that, between the least and the most detectable IGHV gene, there was almost a fourfold difference, and detectability positively correlated with IGHV gene-use frequency observed (**Figure 3B**).

Since the least detectable genes were also the less frequent, we investigated if a lower detection efficiency biased the abundance by correlating our observed IGHV gene frequencies with those from PBMC-derived naïve B cells from six previously published datasets involving a total of 34 donors (**Figure 3C**; Figure S2 in Supplementary Material). Notably, the libraries used for sequencing in these studies were prepared by methods distinct from ours: multiplex PCR from Adaptive Biotechnologies (8), multiplex with primers on the FW1 (9–11), and RACE PCR (6, 12). In our dataset of 11 donors, we observed a relatively high intra-sample biological variability in the frequency of IGHV use (**Figure 3D**). Therefore, the data from different samples were averaged, thereby minimizing the effect of biologic- and method-specific variability. There was a very strong correlation of IGHV gene frequencies in our dataset with those in the six other sets ($R^2 = 0.93$, $P < 0.0001$), suggesting that specific IGHV gene use defined in our library was not significantly biased. Thus, the IGHV frequencies for individual genes were not solely the consequence of low efficiency for specific alleles. In addition, since the calculated IGHV gene-specific detectability was proportional to the number of UMIs sequenced per unique rearrangement, the data suggest that in some instances—at least for naïve B cells—IGHV gene use and IGH mRNA expression might be connected. Therefore, the estimated IGHV gene use frequency observed might reflect, at least in part, a true, not yet described biological phenomenon.

Use of Chronic Lymphocytic Leukemia (CLL) Cell Spike-In to Assess Sequencing Sensitivity

We investigated the extent that we could detect every individual IGHV-D-J rearrangement using the basic error correction and filtering approaches mentioned above. To do so, we spiked into a PBMC-derived polyclonal B cells population leukemic B cells from patients with CLL, a disease of clonal B lymphocytes presenting the morphology of resting B cells and with a known, discriminatory IGHV-D-J sequence. Specifically, 100 leukemic cells from 58 different CLL samples were sorted into a single tube containing cell lysis buffer. This collection represented 41 different IGHV genes, of which 37 were identical to the germline sequence and 21 exhibited somatic mutations with 1–10% differences from the corresponding germline sequence (**Figure 4A**; Table S2 in Supplementary Material). A fraction of the CLL lysate (1/200, 1/100, or 1/50 dilutions containing equivalent genetic material to 0.5, 1, or 2 CLL cells) was then mixed with a cell lysate created from 5,000 polyclonal B cells from a healthy donor. Using our Ig-seq method, we identified the CLL-specific rearrangements and assessed the presence of each of the 58 different CLL signatures in each condition/replicate. Since each B cell should contain multiple copies of its signature IGH mRNA, even at the higher dilution (0.5 equivalent cells per CLL) material from each CLL would be present and detectable.

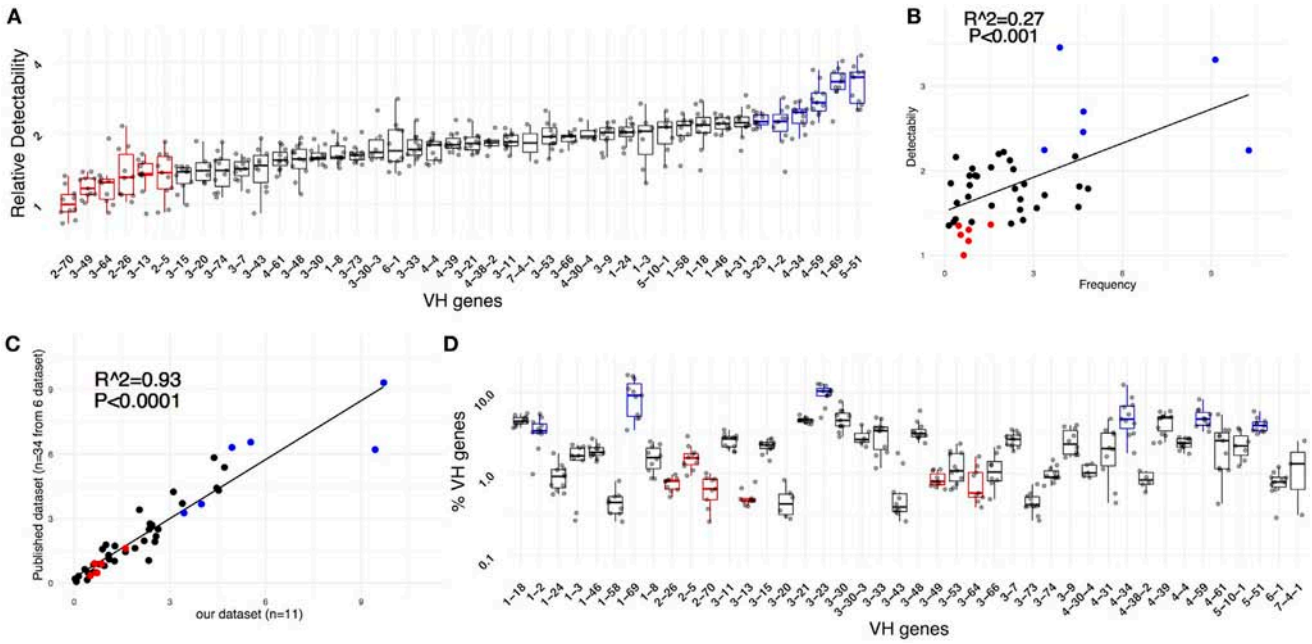


FIGURE 3 | IGHV gene-specific detectability. **(A)** Relative detectability for all genes. For ease of comparison, the efficiencies were scaled to set the minimal median efficiency (corresponding to IGHV2-70) to 1. **(B)** The IGHV gene-specific detectability positively correlates with the IGHV gene frequency observed in unique sequences derived from naïve peripheral blood B cells ($R^2 = 0.27$, $P < 0.001$). **(C)** The average IGHV gene frequency observed in our dataset ($n = 11$) correlates with the average frequency observed in 34 samples from 6 different published datasets ($R^2 = 0.93$, $P < 0.0001$). **(D)** IGHV gene frequencies and intra-sample variability.

Relative Detection Frequency per CLL

Each CLL spike-in (0.5, 1, or 2 cells) was performed in triplicate, and the resulting library sequenced independently at 10×, 20×, and 40× relative to the number of starting cells (i.e., for 5,000 starting cells, 40× equals ~200,000 raw sequences).

The UMIG count per CLL IGHV-D-J was highly variable across CLLs but consistent across replicates (**Figure 4B**); this possibly reflected different IGHV-D-J mRNA expression levels in individual samples and/or sequence-specific efficiency differences. On average, the UMIG count increased proportionally to the amount of starting genetic material (**Figure 4C**), and increasing the depth of sequencing only marginally affected the UMIG count (**Figure 4D**).

Impact of Sequencing Depth and B-Cell Type

An appropriate depth of sequencing is crucial to obtain complete coverage of the starting material and to allow appropriate error correction using the UMIs. Sophisticated error correction techniques (13) require a high depth of sequencing, and this results in greater cost to perform extensive studies with the current technology. Hence, the ideal parameter to use when choosing the depth of sequencing would be the number of starting molecules determined using digital PCR (5) or qPCR. However, a more practical approach is to consider the number of starting cells (6). Here, we focused on obtaining comprehensive coverage—with the basic error correction described above—by correlating the starting number of cells from discrete B-cell populations (i.e., naïve, memory, or plasma cells) quantified by FACS during cell sorting.

As expected, both the depth of sequencing and the amount of starting material per CLL (equivalent number of starting cells) influenced the ability to reproducibly detect each leukemic rearrangement within the healthy PBMC material (**Figure 4E**). Specifically, starting with two cells per CLL sample, a depth of 10× was sufficient for complete coverage; however for 1 and 0.5 cells, a depth of at least 40× was needed. Also, modulating the stringency of filtering led to a change in sensitivity (Figure S3 in Supplementary Material). For example, by choosing a read count of two sequences or higher, full CLL coverage from 1 cell at 20× depth was obtained; however, at a read count of 5 or higher, information for 0.5 cells at 40× depth was lost. Note that at 40× depth, in this experiment, we maintained full CLL coverage with read counts 7 or higher.

Impact of Genetic Material Dilution

We also used the spike-in data to assess the effect of genetic material dilution during library preparation on sequencing resolution. Lysate from PBMC spike-in with 1 cell per CLL was used to prepare ds-cDNA without genetic material dilution between steps, as described above. Then the ds-cDNA from each tube was divided into multiple aliquots at defined dilution factors (undiluted, 1:2, 1:4, 1:8, or 1:16). Each dilution was analyzed in triplicate, and each aliquot sequenced independently at depth 40× (**Figure 5A**). The results indicated that diluting the genetic material by only 50% compromised the ability to consistently detect each CLL IGHV-D-J (detected 56 out of 58 CLL); this deficiency became even more significant with further dilutions (**Figure 5B**). Thus,

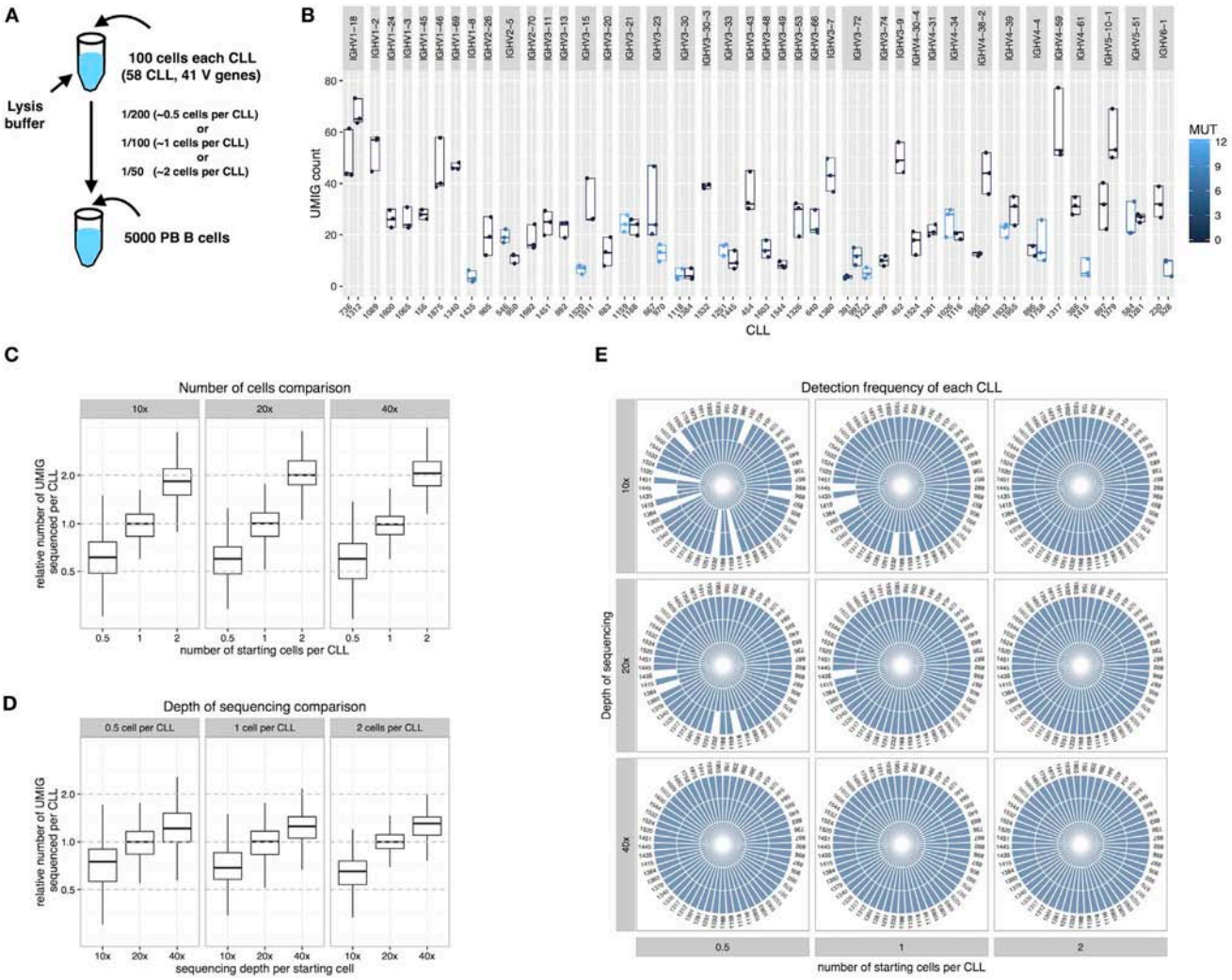


FIGURE 4 | Chronic lymphocytic leukemia (CLL) spike-in experiments. **(A)** A cell lysate containing 100 cells sorted from 58 different CLL samples was added to a lysate of 5,000 PBMCs derived from a normal person. The diluted material represents the equivalent of 0.5, 1, or 2 cells per CLL. **(B)** UMI group (UMIG) count across three replicates for every CLL and IGHV gene. **(C)** Relative UMIG count relative to input material. **(D)** Relative UMIG count relative to depth of sequencing. **(E)** Each slice in the pie charts represents a different CLL, and splice size indicates the frequency of detection among the three replicates. The depth of sequencing is calculated as a multiple of the number of starting cells for the library preparation.

our use of mRNA, cDNA, and ds-cDNA purification using poly-T coupled magnetic beads was crucial. As expected, the UMIG count for each CLL correlated with the detection reproducibility upon genetic material dilution (**Figure 5C**).

DISCUSSION

We have devised a protocol for Ig-seq that reaches single cell resolution. Being able to sequence every cell in a sample is particularly important in studies involving non-expanded B-cell clones such as for analyses of B-cell development, naïve B lymphocytes, long-term immunological memory where the larger clonotypes occur in <0.5% of memory B cells in the peripheral blood (6), or minimal residual disease, where the IGV-D-J rearrangement of interest is by definition not frequent. The methodology is also applicable for less polyclonal repertoires. For example, in studies of intraclonal diversification in leukemia/lymphoma, this method allows the following of clonal evolution by detecting with high accuracy and sensitivity subclonal variants even when present at low frequency (manuscript in preparation).

Our methodology results in a highly efficient process that yields a comprehensive repertoire representation of the starting sample, even when this is as low as ~100 cells. Indeed, we show for the first time the importance of circumventing genetic material dilution in the pre-amplification phases. Capturing mRNA from the entire cell lysate on poly(dT) magnetic beads and then carrying out cDNA synthesis on the same solid phase leads to yields that are several fold higher than conventional approaches (data not shown). This is because in most sequencing protocols, only a fraction of the original genetic material contributes to the

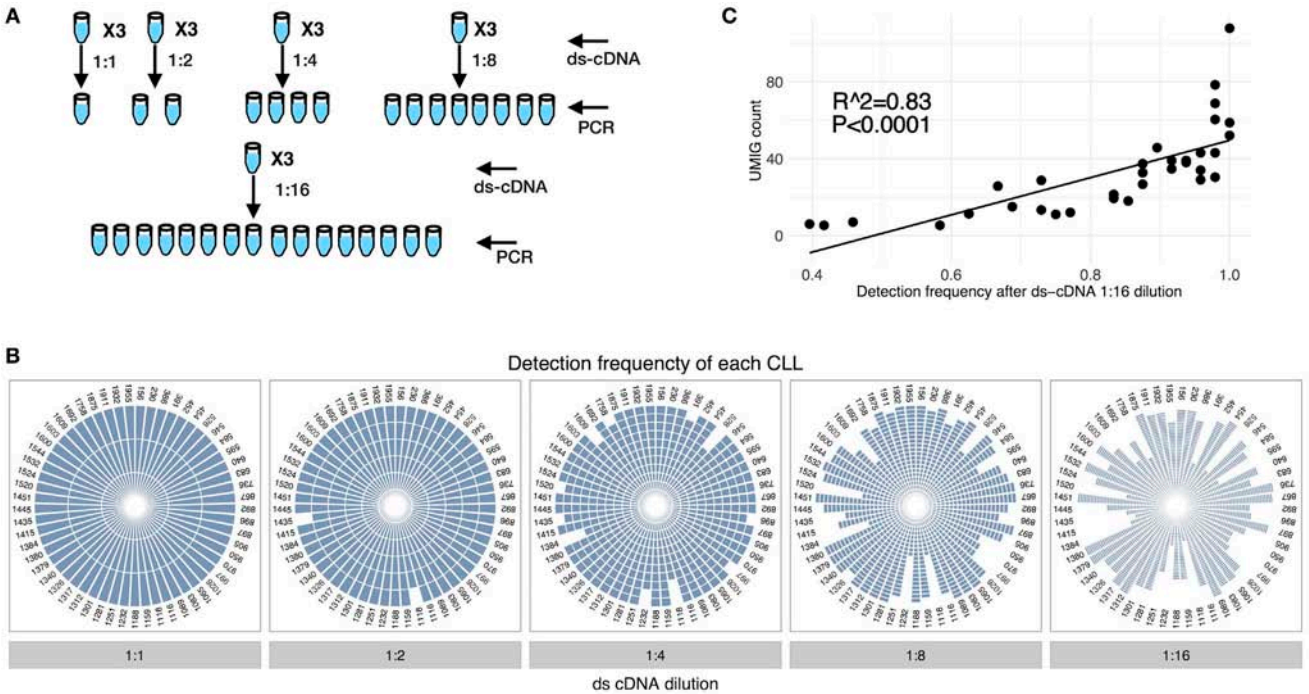


FIGURE 5 | Identification of chronic lymphocytic leukemia (CLL) spike-in upon dilution. **(A)** CLL spike-in was performed as in **Figure 4A**. The ds-cDNA obtained from a CLL spike-in with 1 cell per CLL was diluted at different levels (1:1, 1:2, 1:4, 1:8, and 1:16) to simulate genetic material dilution before PCR amplification step. **(B)** Impact of the dilution of genetic material on the identification of specific CLL IGHV genes. Every fraction of the three different replicates was sequenced. **(C)** The UMIG count of CLL spike-in correlates with the detection frequency after 1:16 dilution.

resulting library due to dilution of the extracted mRNA and/or the cDNA prior to PCR amplification.

Overall, our data agree with the mRNA quantification estimates for plasma cells, memory, and naive B cells reported by Turchaninova et al. (500:5:2). Thus, theoretically, when dealing with B-cell populations with higher IGH mRNA content than naïve B cells (e.g., memory B cells and plasma cells), the starting genetic material might require dilution based on IGH mRNA content relative to that in naïve cells. However, for memory B cells, diluting the mRNA did not have major impact on the resulting sequences, leading only to a shift in UMIG and raw reads count per unique sequence (data not shown). This was not the case for plasma cells that contain more than 100 time mRNA, indicating the requirement for tight control of the amount of B cells from which mRNA is collected. With 40× depth per starting cell, ~500,000 B cells can be sequenced in a single MiSeq run, although the system is calibrated to work with a range from a few hundred up to tens of thousands of cells per reaction tube (data not shown).

Moreover, by introducing a universal forward priming site during the ds-cDNA synthesis [as done by Vollmer et al. (4)], we reduced a potential bias that can occur when employing multiplex primers that undergo exponential PCR amplification. In this regard, we measured the differential detectability of IGHV-D-J rearrangements containing different IGHV genes. This indicated a relative IGHV gene-specific detectability of approximately fourfold. Notably, at least part of these differences appeared to reflect true *in vivo* biology and not solely the consequence of a technical artifact. Although methodological biases can come into play—such as those occurring as a consequence of the multiplex approach for the ds-cDNA synthesis—differential IGH mRNA content should be considered. This latter possibility requires further investigation to assess and understand the extent of these phenomena.

We chose to filter out sequences with read counts less than three. This threshold provided sequences where basic error correction was performed, without a major loss in sensitivity since monoclonal B cell spike-in experiments indicated that we could increase the threshold up to sevenfold without losing information (Figure S1 in Supplementary Material). The latter might not be true for B cells with very low IGH mRNA content, in which case a lower threshold might be preferable, although this might artificially increase diversity. Overall, using sequences with read counts three or higher, at 40× sequencing depth per starting cell provided error corrected sequences with good coverage at a reasonable cost. Increasing coverage and using more sophisticated error correction methods will give a more reliable dataset, compatible with our protocol.

Every sequencing platform and library preparation protocol results in a certain level of error. For this reason, allowing errors in the UMI region is becoming common practice (2, 5, 14). However, we observed that more important than possible errors in the UMI region is the level of UMI diversity. Even with a

limited starting population of only 5,000 naïve B cells per reaction and a theoretical diversity between 4^13 and 4^16 (depending on the specific primer in the multiplex), the apparent UMI diversity was insufficient. Within each UMI cluster, we created sub-clusters based on sequence identity. This led to a striking 30% increase in the number of sequences passing the filter of ≥3 raw reads per unique sequence. The approach taken by Khan et al. (5), where the UMI design does not follow a simple NNN… pattern, might mitigate the problem of the decreased real UMI diversity reducing the complexity of the UMI nucleotide sequences.

In conclusion, we have developed a protocol for Ig-seq where virtually every IGHV-D-J rearrangement in the starting B-cell population(s) can be detected. To achieve this result, we used a methodology with an overall efficiency sufficient to retain the "full" repertoire diversity of the sample analyzed. The key aspects of the method consist in starting from a defined number of cells for which one wants to know the repertoire, avoiding primer-specific PCR amplification and dilution of the starting genetic material for low IGH mRNA content cells, and achieving a minimum of 40× depth per starting number of cell.

METHODS

Samples

The study was approved by the Institutional Review Board of Northwell Health. Written, informed consent was obtained before blood collection from CLL patients in accordance with the Declaration of Helsinki. PBMCs from the CLL patients and from anonymous healthy blood donors were separated by density gradient centrifugation (Ficoll, GE Healthcare), frozen (10% DMSO, 45% FBS, and 45% RPMI), and stored in liquid nitrogen until used.

Cell Sorting

PBMCs from normal blood donors were incubated with the following anti-human Abs: V500 anti-CD19 (BD Biosciences), PerCPcy5.5 anti-CD38 (BioLegend), PE-cy7 anti- CD24 (BioLegend), FITC anti-IgD (ThermoFisher), and allophycocyanin anti-CD27 (BD Bioscien). CLL patient PBMCs were exposed to the following anti-human Abs: V500 anti-CD19 (BD Biosciences) and PE-cy7 anti-CD5 (Invitrogen). Non-B cells were excluded with efluor-450 anti-CD3 and anti-CD16, and dead cells were excluded by Sytox Blue staining (ThermoFisher). B cells were sorted directly into 200 µl PCR tubes containing 100 µl Dynabeads Oligo(dT) (ThermoFisher) lysis buffer and stored at −80°C.

Library Preparation and Sequencing

mRNA isolation from B-cell lysates was performed using Dynabeads mRNA DIRECT Micro Kit (ThermoFisher). The protocol used was that suggested by the manufacturer, except that mRNA isolation was performed in 200 µl 96-well PCR plates to enable parallel processing with the support of a 96-well magnetic stand. mRNA was used in its entirety for reverse transcription in 10 µl (50°C 1 h, 72°C 10 min) using SuperScript III Enzyme (ThermoFisher) in solid phase with Dynabeads Oligo(dT) as primer. After RNase H treatment, second-strand synthesis was performed in solid phase in 10 µl using Q5 Polymerase (NEB) and a mix of 13 primers covering all IGHV leader sequence segments reported in the IMGT database with a maximum of one mismatch, containing 13 to 16 random nt and partial Illumina adaptor sequences (37°C 20 min, 98°C 30 s, 62°C 2 min, and 72°C 10 min). Double-stranded cDNA was washed three times in 10 mM tris–HCl to remove the remaining primers, and the entire sample was used as template for PCR amplification in 10 µl using Q5 Polymerase with universal FW primer and mix of reverse isotype specific primer (98°C 30 s; 10 cycles of 98°C 10 s, 58°C 15 s, and 72°C 1 min; 72°C 10 min). Two microliters of the PCR product were used for a semi-nested PCR with inner RV primers for the constant region, which also introduce partial Illumina adaptors. This reaction was carried in 20 µl (98°C 30 s; 15 cycles of 98°C 10 s, 58°C 15 s, and 72°C 1 min; 72°C 10 min). The PCR product was purified with Ampure XP beads at a ratio of 1:1, and 1–10 ng used to add Illumina Index with Nextera XT kit (Illumina). The MiSeq Illumina (v3 2 × 300 kit, Illumina MS-102-3003) was used to sequence the library. The library was loaded at 12 pm with 10% PhiX. The list of the primers is in Table S1 in Supplementary Material. Raw data are deposited at SRA (BioProject ID PRJNA381394—http://www.ncbi.nlm.nih.gov/bioproject/381394).

Bioinformatic Analysis

Processing of raw reads was performed using a custom workflow built with pRESTO (REpertoire Sequencing TOolkit) (7). IGHV sequences obtained were then submitted to IMGT/HighV-QUEST (15) and analyzed using ChangeO (16), and custom R scripts.

Relative Detectability Estimation

A metric to quantify the relative abundance of gene-specific UMIG counts and to compare this to the total abundance of all IGHV genes within a sample was developed. In order to focus on gene-specific patterns, relative measure was used. Starting with two UMIG count distributions, the relative abundance metric summarizes the position of one distribution relative to the other. Each distribution was encoded as a vector of fine grained quantiles and performed a linear regression between the paired sets of quantiles of the two distributions. The slope of this line represents the relative shift of one distribution against the other and was thus termed the relative efficiency. In this paper, were used 100 quantiles, from 0 to 99, evenly spaced at 1% intervals for the detectability estimation. For this reason, were discarded any distributions with fewer than 100 points from the analysis. Figure S4 in Supplementary Material shows intermediate steps of this analysis for a highly detectable gene (IGHV5-51) and a low detectable gene (IGHV2-70), across all samples.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Belmont Report, and the Office of Human Research

Protection Program Institutional Review Board at Northwell Health System. All CLL samples were from individuals who provided written informed consent for the collection and use of samples for research purposes according to the Declaration of Helsinki. The Protocol was approved by the Northwell Health Institutional Review Board.

AUTHOR CONTRIBUTIONS

DB and SV performed the experiments; DB, SV, and IK analyzed the data; NC provided project funding; DB, SV, AM, GF, and NC interpreted the results and wrote the manuscript; DB designed the experiments and directed the project.

ACKNOWLEDGMENTS

This work was supported in part by philanthropic contributions from The Karches Foundation, Marks Foundation, Nash Family Foundation, the Mona and Edward Albert Foundation, and the Jean Walton Fund for Leukemia, Lymphoma Myeloma Research, and Fondazione Umberto Veronesi.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://journal.frontiersin.org/article/10.3389/fimmu.2017.01157/full#supplementary-material.

REFERENCES

1. Lam KP, Rajewsky K. Rapid elimination of mature autoreactive B cells demonstrated by Cre-induced change in B cell antigen receptor specificity in vivo. *Proc Natl Acad Sci U S A* (1998) 95:13171–5. doi:10.4172/1745-7580.1000056
2. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier-guided amplicon assembly. *J Immunol* (2016) 196:2902–7. doi:10.4049/jimmunol.1502563
3. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* (2012) 135:183–91. doi:10.1111/j.1365-2567.2011.03527.x
4. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110:13463–8. doi:10.1073/pnas.1312146110
5. Khan TA, Friedensohn S, de Vries ARG, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) 2:e1501371–1501371. doi:10.1126/sciadv.1501371
6. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) 11:1599–616. doi:10.1038/nprot.2016.093
7. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30:1930–2. doi:10.1093/bioinformatics/btu138
8. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A public database of memory and naive B-cell receptor sequences. *PLoS One* (2016) 11:e0160853. doi:10.1371/journal.pone.0160853
9. Bagnara D, Squillario M, Kipling D, Mora T, Walczak AM, Da Silva L, et al. A reassessment of IgM memory subsets in humans. *J Immunol* (2015) 195:3716–24. doi:10.4049/jimmunol.1500753
10. Tipton CM, Fucile CF, Darce J, Chida A, Ichikawa T, Gregoretti I, et al. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat Immunol* (2015) 16:755–65. doi:10.1038/ni.3175
11. Martin V, Wu Y-CB, Kipling D, Dunn-Walters D. Ageing of the B-cell repertoire. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140237. doi:10.1098/rstb.2014.0237
12. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* (2016) 7:11112. doi:10.1038/ncomms11112
13. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) 11:653–5. doi:10.1038/nmeth.2960
14. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova OV, Sharonov GV, Staroverov DB, et al. Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J Immunol* (2015) 194:6155–63. doi:10.4049/jimmunol.1500215
15. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
16. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) 31:3356–8. doi:10.1093/bioinformatics/btv359

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.

Copyright © 2017 Vergani, Korsunsky, Mazzarello, Ferrer, Chiorazzi and Bagnara. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 16 October 2017
doi: 10.3389/fimmu.2017.01286

Immune Repertoire after Immunization As Seen by Next-Generation Sequencing and Proteomics

Martijn M. VanDuijn[1], Lennard J. Dekker[1], Wilfred F. J. van IJcken[2], Peter A. E. Sillevis Smitt[1] and Theo M. Luider[1]*

[1] Department of Neurology, Erasmus MC, Rotterdam, Netherlands, [2] Erasmus Center for Biomics, Erasmus MC, Rotterdam, Netherlands

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Gunnar Houen, Statens Serum Institut, Denmark
*Evan W. Newell, Singapore Immunology Network (A*STAR), Singapore*
Andrew Horton, University of Texas at Austin, United States

***Correspondence:**
Martijn M. VanDuijn
m.m.vanduijn@erasmusmc.nl

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *30 June 2017*
Accepted: *25 September 2017*
Published: *16 October 2017*

Citation:
VanDuijn MM, Dekker LJ, van IJcken WFJ, Sillevis Smitt PAE and Luider TM (2017) Immune Repertoire after Immunization As Seen by Next-Generation Sequencing and Proteomics. Front. Immunol. 8:1286. doi: 10.3389/fimmu.2017.01286

The immune system produces a diverse repertoire of immunoglobulins in response to foreign antigens. During B-cell development, VDJ recombination and somatic mutations generate diversity, whereas selection processes remove it. Using both proteomic and NGS approaches, we characterized the immune repertoires in groups of rats after immunization with purified antigens. Proteomics and NGS data on the repertoire are in qualitative agreement, but did show quantitative differences that may relate to differences between the biological niches that were sampled for these approaches. Both methods contributed complementary information in the characterization of the immune repertoire. It was found that the immune repertoires resulting from each antigen had many similarities that allowed samples to cluster together, and that mutated immunoglobulin peptides were shared among animals with a response to the same antigen significantly more than for different antigens. However, the number of shared sequences decreased in a log-linear fashion relative to the number of animals that share them, which may affect future applications. A phylogenetic analysis on the NGS reads showed that reads from different individuals immunized with the same antigen populated distinct branches of the phylogram, an indication that the repertoire had converged. Also, similar mutation patterns were found in branches of the phylogenetic tree that were associated with antigen-specific immunoglobulins through proteomics data. Thus, data from different analysis methods and different experimental platforms show that the immunoglobulin repertoires of immunized animals have overlapping and converging features. With additional research, this may enable interesting applications in biotechnology and clinical diagnostics.

Keywords: immune repertoire, immunization, NGS, mass spectrometry, immunoglobulins

INTRODUCTION

The basic understanding of the molecular biology that leads to diversity in the adaptive immune response emerged around 1980 (1), an effort that was awarded with a Nobel prize for Physiology and Medicine for Tonegawa. Yet, it is only in recent years that technology has advanced sufficiently to study the population of sequences that results from this recombination process and the subsequent

mutation and selection pressures for the formation of mature immunoglobulins (2). The high-throughput sequencing methods that are available allow researchers to obtain a listing of the repertoire of sequences that make up the antibodies or T-cell receptors that mediate the adaptive immune response. Research groups have started using and refining such tools to understand the development of immune responses, and envision potential applications of information on the immune repertoire.

Yet, it is challenging to obtain a sample for sequencing that properly reflects the repertoire of antibody proteins that is present in the serum, and even more the repertoire of an antigen-specific subset of sequences. One challenge is that not all cells with a rearranged immunoglobulin locus express immunoglobulin protein. Distinctions have been found between the B-cell receptor repertoire and the plasma cell repertoire that drives immunoglobulin expression (3). Another challenge is the tissue niche that is sampled for obtaining sequence information. The immune response is a compartmentalized process that takes place in circulating blood, in the interstitial space of (inflamed) tissues, and in lymphoid organs, such as lymph nodes, the spleen, or bone marrow. The timing and location of the sampling sites are likely to affect the immune repertoire that is observed, and not all sites are easily accessible, especially in human subjects. However, antibodies that are produced as a result of an immune response will generally end up in the circulation regardless of the site of production. Antibody proteins can be collected from serum and affinity-enriched in order to study an antigen-specific subset of molecules. For these reasons, we here study the immune repertoire with a combination of proteomics and NGS. In this way, we can obtain a more comprehensive picture of the differences but also similarities that exist between individuals after an immune response to a particular antigen. The techniques were already combined successfully in the past, and can help provide unique but not always consistent views on the repertoire (4–8).

We previously found evidence for common features between antigen-specific immune sera. The findings are consistent with an increasing body of literature that reports commonalities in the sequence of immunoglobulins targeting a particular antigen (9–14). An immune repertoire consisting of sequences that are not unique to an individual is referred to as a public or stereotyped response. It is thought that such responses result from the selection of specific rearrangements during the initial immune response, or the selection of similar somatic mutations through a process of convergent evolution of the repertoire.

This experiment was designed with a number of distinguishing characteristics that define the data that were collected. First, the immune repertoire was studied in a group of laboratory outbred animals rather than in a heterogeneous population of human subjects. Second, the animals were all immunized with a purified antigen rather than with a pathogen that exposes a multitude of antigens and epitopes. Finally, the samples were analyzed with a combination of proteomics and long-read NGS, two techniques that provide complementary on the immune repertories and both allow us to examine the entire variable domain of the immunoglobulins. With proteomics, affinity-enriched antibodies can be studied, but with limited sequence length or sequence accuracy. Our NGS method offers superior depth, read length, and sequence accuracy, and in combination the strengths of both can be combined. With this dataset, we aim to validate our earlier proteomics observations on convergence in antigen-specific immune repertoires, perform an extended analysis with the NGS data, and establish the value of both techniques in the study of immune repertoires.

MATERIALS AND METHODS

Wistar rats were immunized and analyzed by proteomics as described earlier, and spleen material collected from these animals is now used for NGS analysis (10). Rat immunization and sample collection was performed by Genovac GmbH (Freiburg, Germany) under their local permits and regulatory framework. The immunization and three boosts were performed with either recombinant HuD or Keyhole Limpet Hemocyanine modified with dinitrophenyl (DNP) residues, each time with 2-week intervals. HuD is an onconeural antigen related to a paraneoplastic neurological syndrome (15), and DNP was chosen as a well-defined small epitope. Pre-immune and immune sera were collected, as well as a spleen cell suspension. IgG was isolated from the sera with Melon Gel (Invitrogen, Carlsbad, CA, USA), optionally affinity enriched against HuD or DNP-ovalbumin immobilized on Aminolink Plus particles (Thermo Fisher Scientific, Rockford, IL, USA), digested with trypsin and analyzed by a 90 min gradient on a Pepmap Acclaim column, coupled to on an LTQ Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany) set at 30,000 resolution MS1 in the Orbitrap and using dynamic data acquisition to produce CID MS–MS spectra in the ion trap.

For NGS, RNA was collected from 50×10^6 total splenocytes using Trizol (Life Technologies) and additional cleanup with an RNeasy spin column (Qiagen, Germantown, MD, USA). 5 μg RNA collected from splenocytes was reverse transcribed to cDNA with Superscript III (Invitrogen, Waltham, MA, USA) and primers complementary to the constant domains, which included a Unique Molecular Identifier segment (UMI; Supplementary Material). After the addition of Superscript III, reverse transcription proceeded for 40 min at 50°C, after which the enzyme was inactivated at 70°C for 15 min. After cDNA product cleanup with AmpureXP beads (Beckman Coulter, Indianapolis IN, USA), PCR was performed with Phusion proofreading polymerase in HighFidelity buffer (New England Biolabs, Ipswich, MA, USA). The cDNA was amplified with a multiplex degenerate forward primer set and a common reverse primer. Forward PCR primers were designed with the HYDEN degenerate primer design tool (16) and are listed in the Supplementary Material. The PCR was run in a touchdown fashion for 26 cycles, each cycle reducing the annealing temperature by 0.5°C starting from 68°C. A final eight cycles were performed at 55°C. All extensions were performed at 72°C. The PCR product was purified using AmpureXP and concentrated by speedvac. Dual-indexed sequencing libraries were constructed from 120 ng of PCR amplicon according to the manufacturer's instructions of the Ovation ultralow library kit (Nugen, San Carlos, CA, USA) using custom diversity adaptors with 1–8 random nucleotides before the PCR amplicon.

The library was quantified by qPCR and sequenced on a MiSeq with 2× 300 bp paired end chemistry (Illumina, San Diego, CA, USA). Material from all 10 samples was multiplexed in a single MiSeq run. Sequencing data were demultiplexed on index as well as PCR primer. Paired end reads were combined with PEAR and, subsequently, assembled using the MIGEC (17) pipeline, which processes the molecular barcode information for sequence error correction and to report expression levels without PCR bias. Default parameters were used except a minimum UMI count of 1 in MIGEC. The resulting sequences were annotated for germline alleles and regions with the High-VQuest service (18), and additional analysis was performed with the VDJTools package (19) for clustering of samples and tcR (20) to enumerate sequences overlapping between samples. VDJTools clustering was performed with the ClusterSamples function using the default distance parameter (clonotype overlap frequencies). Phylogenetic trees were built with the FastTree program and visualized with the Archaeopteryx viewer (21, 22).

Analyses on proteomics data were performed using ProgenesisQI for Proteomics 2.3 (Waters, Milford, MA, USA) for label-free quantitation and PEAKS Studio 6 (BSI Inc., Waterloo, ON, Canada) for sequence identification. Quantification is reported by ProgenesisQI as an integrated intensity under the isotopic peaks in the mass spectra, normalized for sample loading. Search parameters in PEAKS allowed for a fixed cysteine carbamidomethylation and variable methionine oxidation, a precursor mass tolerance of 10 ppm and 0.5 Da tolerance for ion trap MS–MS spectra and 1 missed cleavage. A search database was constructed from productive reads reported by High-VQuest from all samples combined. At 239 MB, the FASTA file for this database was slightly smaller in size than the common Uniprot database. Peptide spectrum matches with a −10logP confidence better than 15 were included in further analysis.

Proteomics data are made available for public use at the ProteomeXchange Consortium (23) (DOI 10.6019/PXD006484). The NGS data can be obtained from the NCBI Gene Expression Omnibus as study GSE98855 (24).

RESULTS

The immunizations and proteomics dataset were described in earlier work (10). The sequence data were demultiplexed, yielding between 1.0×10^6 and 2.2×10^6 paired end reads per biological sample. After processing the raw reads and collapsing those sharing the same barcodes, we assessed the class distribution of the reads. 79% of the reads related to IgG, followed by IgA, IgM, IgD, and IgE (**Figure 1**). As observed in other studies, the repertoire distribution was very skewed, showing a limited number of clones making up the majority of the expressed repertoire (Figure S1 in Supplementary Material).

Sample Clustering by Antigen

It was previously observed that animals immunized with different antigens could be distinguished from each other based on a cluster analysis on a proteomics dataset of affinity-enriched antibodies from the immune sera. A similar approach was performed on the immune repertoire data that were obtained from the splenocytes of these animals. The dataset consists of entire variable domain sequences, rather than the short peptide fragments that were identified in the proteomics data. Within the variable domain of the immunoglobulin, the complementarity-determining regions 1, 2, and 3 can be found as well as the surrounding framework regions. Unique rearrangements can be found in the CDR3 regions, and somatic mutations focus on the CDR regions but are not uncommon in the framework regions either. It is, therefore, of interest to assess several sections of the variable domain separately to assess similarities that may cluster the samples into groups. The dataset was processed with the High-VQuest service to annotate the various immunoglobulin regions in the sequence and to enable filtering for functional transcripts. The segments of interest were collected, collapsed to a unique set annotated with the read number, and the ClusterSamples function in the VDJTools package was then used to cluster the CDRs 1, 2, and 3 individually, as well as these CDRs together with their flanking framework regions. It was found that samples could be clustered according to the antigen that was used for the immunization based on all segments that were assessed, but clustering was strongest based on the CDR3 region, and became stronger by the inclusion of the flanking framework regions (**Figure 2**; Figure S2 in Supplementary Material).

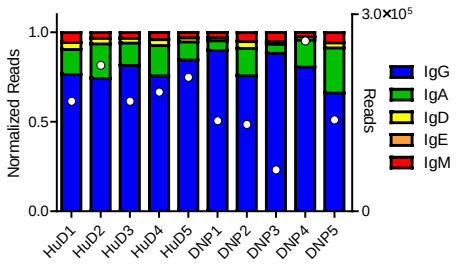


FIGURE 1 | The NGS dataset included a small section of the constant domain that allows identification of the class of immunoglobulin. The reads were normalized, but the total number of processed reads included for each sample has also been plotted in white markers on the right axis. In all samples, the majority of reads belonged to the IgG class, followed by IgA and the other classes. Differences were observed between subjects but no relation to the treatment could be shown.

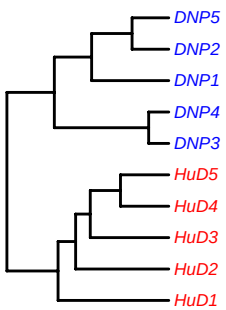


FIGURE 2 | Unsupervised hierarchical clustering of samples based on NGS data on the repertoire of unique CDR3 sequences. Cluster analysis based on other regions is provided in the Supplementary Figures.

Correlation between Proteomics and NGS Data

Proteomics data were acquired for both total IgG and affinity-enriched IgG from all immune sera. The NGS data were derived from splenocytes taken at the same point in time from these animals. However, the splenocytes do not represent the only site of IgG production, and it was, therefore, of interest to investigate the proportion of mass spectra that could be matched to reads in the NGS dataset, as well as the correlation between the number of reads and the intensity of mass spectrometer signals for a matching spectrum.

A database was constructed from all unique immunoglobulin sequences observed in all samples combined. This database was used to match MS/MS spectra with a PEAKS DB search. Affinity-enrichment yields previously suggested that the specific IgG makes up about 0.1% of the total amount of IgG. On the other hand, the spleen may be enriched for immune cells related to an active immune response, which would be the case after immunization and boosts. As shown in Table S1 in Supplementary Material, the fraction of MS/MS spectra that could be matched to the NGS results was larger in the case of the samples of total IgG than in the case of the affinity-enriched samples. This fits with the notion that the affinity-enriched IgG is a subset of the total repertoire, and that the splenocytes can be involved in immune responses against both the immunogen and numerous other antigens. Splenocyte IgG sequences that target such other antigens will remain unmatched to proteomics data on IgG affinity-enriched for the immunogen. The pre-immune sera show an intermediate number peptide spectrum matches, indicating an overlap between serum IgG and splenocytes in spite of being sampled 3 weeks apart.

While a single unified database was used for identifications, we separately compared searches performed with a database matching or mismatching the animal used for a proteomics sample. As a mismatching database, NGS data from an animal of the alternate immunization was used. It was found that a matched proteomics-NGS dataset typically yielded more peptide spectral matches than an unmatched set, and that a matched set yielded more unique hits that were not found in the unmatched set than vice versa (Table S2 in Supplementary Material). This supports the expectation that some of the unique rearrangements in the animals can be detected by both the proteomics data and the NGS data, but still about 75% of the identifications were seen in both the matching and non-matching search. Additional searches against a UniProt database revealed that the proteomics samples consisted primarily of immunoglobulin-related peptides, but abundant serum proteins such as albumin or complement factors could be observed as well. The total number of PSMs against the rat Uniprot database was about 25% of the number found against the NGS database.

For all NGS sequences that had a match in the proteomics dataset, the number of (UMI corrected) reads containing that peptide was annotated for all samples, as well as the signal intensity in the proteomics data as determined by label-free analysis (ProgenesisQI). A good correlation between RNA and proteome data would suggest that the splenocytes are a good representation of the expressed repertoire in the serum. It was found that the correlation between affinity-enriched IgG and the splenocyte RNA was almost absent (median of pairwise correlations 0.06, Figure S3 in Supplementary Material). As above, this fits with affinity-enriched IgG as a subset of the total repertoire. The correlation of the total serum IgG with the splenocyte RNA, while still modest (median of pairwise correlations 0.24), was significantly stronger than that of the affinity-enriched data. Correlations were not increased in a subset of the data with only higher confidence PSMs ($-10\log P > 40$). The incomplete correlation suggests that many of the serum antibodies were not expressed by the B-cells from the spleen or cells clonally related to them, but rather other cell populations, possibly bone marrow-resident plasma cells. However, it cannot be excluded that the incomplete repertoire coverage depth of, in particular, the proteomics data affects the correlations that were found.

Shared Sequence Motifs

From proteomics data alone, it was previously concluded that certain peptides from antigen-specific immunoglobulins are shared among different animals, yet unique for the immunogen. In the current work, the NGS dataset presents an opportunity to validate the proteomics findings with an independent technique and to perform a more extensive exploration of the immune repertoire than was possible with the proteomics approach alone.

A small subset of peptides that was previously identified by proteomics as selective for one of the immunogens was evaluated in the NGS dataset. The peptide sequence was initially uncertain as only *de novo* interpretation of MS/MS spectra was available. As described in the previous section, these MS/MS spectra can be searched against a database from the matching NGS dataset, and the number of reads for matching sequences was enumerated and compared to the proteomics label-free quantitation. It was found that, in several cases, multiple unique sequences could be a potential match to the MS/MS spectra in the proteomics dataset. However, such sequences were always quite similar (leucine/isoleucine variants, or residue position swaps), and one variant always dominated the number of reads (Table S3 in Supplementary Material). The latter is most likely to correspond to the peptide that was observed in the proteomics data. Again, we found that the quantitative correlation was low, but that both datasets agree qualitatively and corroborate the observation that these peptides associate with one of the immunogens. This has been illustrated for two of these peptides in **Figure 3**. Some other peptides that were quite abundant and shared in proteomics data could only be matched to low numbers of reads, or did not find a match at all, showing that the overlap between the datasets is not complete.

The size and read length of the NGS repertoire dataset permits a more detailed analysis of shared sequence segments. Similar to the cluster analysis, it was chosen to compare sharing for CDRs 1, 2, and 3, for CDRs expanded with the flanking framework regions, and for the complete variable domain. The datasets were processed in R with the shared.repertoire function of the tcR library (20). The output lists the number of reads that were observed in each sample for a given shared sequence segment. In these data files, it was assessed how many sequences were

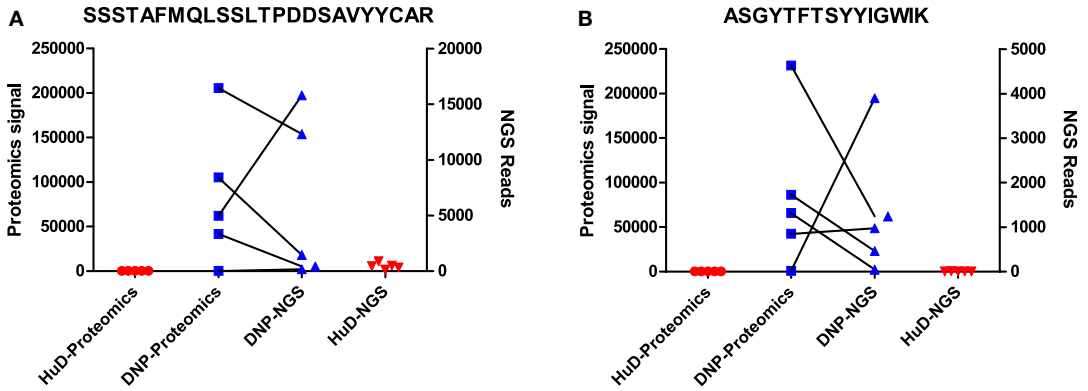


FIGURE 3 | Data for two peptides [SSSTAFMQLSSLTPDDSAVYYCAR **(A)**; ASGYTFTSYYIGWIK **(B)**] that were found to be associated with DNP immune sera in proteomics data. Shown are signal intensities in proteomics, and the number of reads in NGS data containing the same peptide. Matching DNP samples have been plotted with a connecting line.

shared within but unique to one of the immunogen groups. This was done for both immunogen groups and also for randomized controls where sequences were to be shared between a mixed set of five animals from the dataset. For the latter, the average and SD of 10 randomized sets is shown. For all immunoglobulin segments that were tested, shared sequences were found in the datasets. The number of shared sequences was plotted against the number of animals among whom they were shared (**Figure 4**; Figure S4 in Supplementary Material). It was observed in all comparisons that the number of shared sequences was largest within animals immunized with the HuD antigen and less with the DNP antigen. In the scrambled control sets, sharing was less than within either HuD- or DNP-treated animals, which were typically outside the 95% confidence limit of these controls. The presence of shared sequence segments within a mixed control group could indicate some sharing events occur by chance alone, although all immunized animals were also littermates that did share exposure to other environmental factors. Therefore, all subjects can share immune responses against antigens other than the intended immunogen as well. **Figure 4** and Figure S4 in Supplementary Material showed the extent of sharing for peptides that are found in animals treated with one antigen but that are not found for the alternate antigen. The analysis was repeated for shared peptides but without further constraint on presence or absence in animals with the alternate antigen, and the results for the CDR3 were also included in a panel of Figure S4 in Supplementary Material. Without the constraint more shared peptides were found, but differences between antigens and random controls were reduced. Similar results were found for the other regions of the immunoglobulin molecule. Although the absolute number of shared sequences varied depending on the region of the immunoglobulin molecule, in all cases the number of shared sequences decreased in a log-linear fashion as a function of the number of samples in which they were shared. If this trend also holds for a larger number of subjects than studied here, this implies that the sharing of any single sequence segment among all members of a large population that responded to an immunogen will be quite rare. This may affect the design

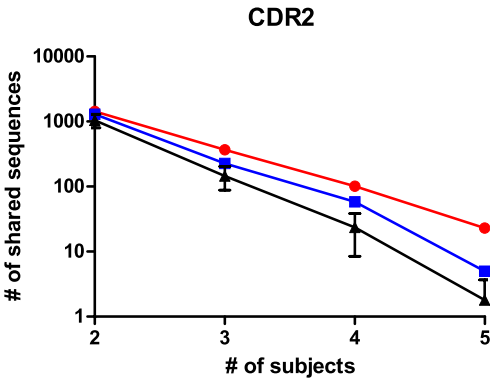


FIGURE 4 | The number of shared peptides that were observed for CDR2 sequences in the NGS dataset. Peptides that were shared among animals, but still unique to one of the antigens were enumerated and shown as a function of the number of subjects among whom they were shared, for HuD immunized animals (Red), DNP-immunized animals (Blue) and randomized groups of mixed composition (black, average ± SD, $n = 10$). A similar analysis for other regions in the immunoglobulin molecule is presented in Figure S4 in Supplementary Material.

of diagnostic applications that rely on shared motifs in the immune repertoire.

For the CDRs 1 and 2, the number of shared sequences is increased when the flanking framework regions are included in the analysis. It may seem counterintuitive that longer sequences are shared more, but this could be explained by CDR sequences that are split up into unique entities because of distinct motifs in their framework regions and that are, thus, counted multiple times versus only once when considering the CDR alone. This observation does not hold for the CDR3, which shows similar levels of sharing with and without the surrounding frameworks. Possibly this relates to the higher diversity in de CDR3 and the lower number of reads per unique CDR3.

We analyzed the size distribution of the CDR3 sequences that were shared among animals and compared them to the size

distribution for the CDR3 in the entire dataset. We found that the shared CDR3 sequences were significantly shorter, which is consistent with other reports in the literature (13) (Figure S5 in Supplementary Material). While very long CDR3s have been associated with long-lasting and well-developed immune responses, it has also been shown that shared and shorter CDR3 sequences still encode for antigen-specific sequences (13, 25). While the sharing of shorter CDR3 sequences seems to be favored; it is, therefore, still expected that the shared sequences represent normal antigen-specific antibodies.

Phylogenetic Analysis

In order to interpret the relations between the immunoglobulin sequences in the sample, a phylogenetic analysis was performed. First, the 200 most abundant reads were taken from each biological sample, combined and aligned as IMGT-gapped amino acid sequences 1–108 and processed with FastTree. The resulting data were visualized as a circular phylogram with the Archaeopteryx viewer and color coded for the treatment group (**Figure 5**). Several branches can be identified in the phylogram that contained reads from only one treatment group, but that still represented all animals from that group. This suggests that the immunization led to homologous sets of sequences, also among the more highly expressed sequences.

While such sequences probably relate to the antigens of interest, we further explored phylogenetic relations based on a peptide (ASGYTFTSYYIGWIK) that emerged from the proteomics data of affinity-enriched anti-DNP antibodies. While this peptide was also found among the high abundant subset analyzed above, we computed phylogenetic trees based on all productive reads from the MIGEC/High-V-Quest processing but for each animal separately. Subsequently, all reads containing the peptide were highlighted in an unrooted tree diagram (**Figure 6**; Figure S6 in Supplementary Material). These reads, probably related to anti-DNP immunoglobulins, clustered in distinct branches in the total repertoire. Subsequently, sequences from all nodes of such branches were listed, and for each sample a weblogo plot was constructed of the sequence repertoire, as well as one for the most homologous germline sequence (26). In these diagrams, it can be observed that, within the consensus sequence, up to four residue mutations are favored at selected positions, while otherwise the repertoire does not deviate much from the germline sequence (**Figure 7**).

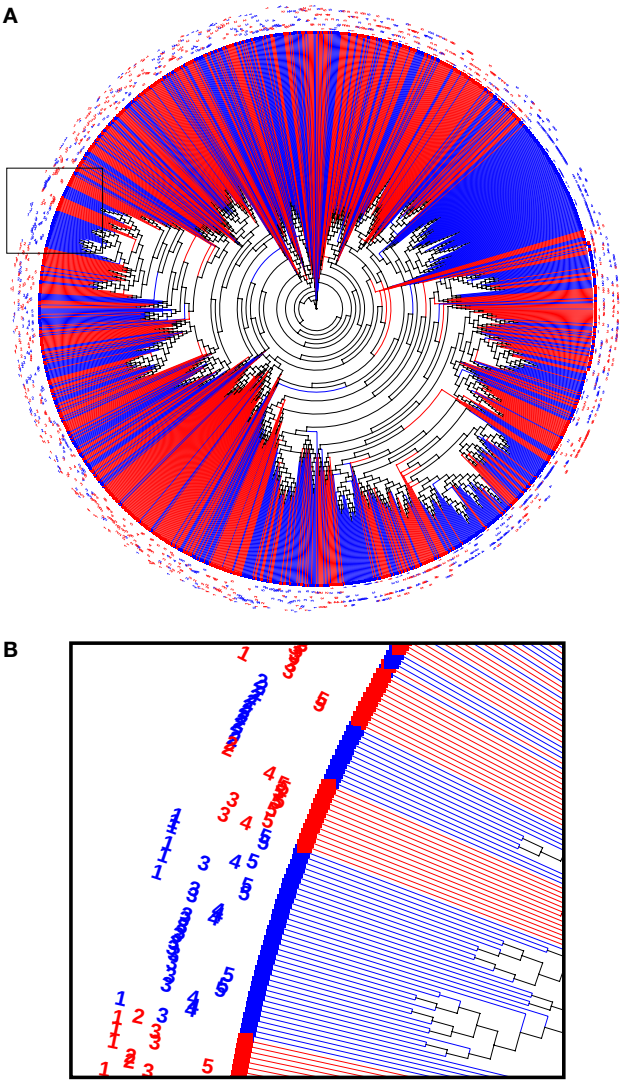


FIGURE 5 | A phylogenetic tree computed from a combined set of the 200 most abundant reads from each sample **(A)**. Nodes from HuD were plotted in red and nodes from DNP-treated animals were plotted in blue, and sample identifiers (1–5) were included in the label. Branches that were homogeneous for the antigen that was used (single color) often contained reads derived from all of the individual animals of that group. The indicated region that was enlarged **(B)** highlights examples of both mixed sample branches and a single sample (DNP-2) branch.

Proteogenomic Analysis of Repertoires

The availability of NGS data enables an extended analysis of the proteomics data, and also an analysis of some discrepancies that were observed in the past. One limitation of bottom-up proteomics is the limited sequence length covered by a tryptic peptide. The large set of homologous peptides in an immunoglobulin digest makes is impossible to reconstruct the full protein sequence associated with a given peptide of interest. By matching the peptide MS/MS spectrum to the NGS dataset, full variable region sequences can be identified that contain the peptide of interest. Moreover, these full length sequences may contain other tryptic peptides that are also represented in the proteomics dataset. Thus, observations from one peptide may now be supported by observations on additional peptides from the same chain, which could otherwise not be recognized as related. We found several examples of such peptides, and indeed such peptides showed similarities in their abundance in the samples (Figure S7 in Supplementary Material). Still, differences were observed as well, which may relate to the fact that, while peptides occur within the same chain for a subset of the reads, also other reads are present within the repertoire that contain either one peptide or the other but not both.

During the analysis of the proteomics dataset, it was observed that CDR3 sequences were underrepresented in the results.

It was unclear whether this related to limitations in sample preparation, detection in the instrument, or proper identification and recognition of these polymorphic regions as CDR3. One cause for poor detection can be long peptide lengths. For proteomics, the optimal peptide lengths range between 7 and 15 amino acids. We performed an *in silico* digestion of the NGS dataset using the Bio:Protease Perl package, and enumerated the length of tryptic peptides encompassing the CDR3, defined as a peptide with tryptic sites that surround IMGT-numbered residue 107 (**Figure 8**). The analysis revealed that the distribution peaked around 50-aa length, which is a length that adversely affects peptide detection. A secondary peak was observed for peptides of 2–5 aa, which is rather too short for detection and specificity. For comparison, the same algorithm was used to process a database with the human subset of the Uniprot protein database. That distribution, although long-tailed, peaks within the optimal range of 7–15 amino acids and is clearly different from CDR3 peptides. The data show that the amino acid composition of the CDR3 region is poorly suited for conventional trypsin-based proteomics and would benefit from either alternative proteases or from instrumental capabilities for larger molecules, such as high mass resolution and ETD fragmentation rather than only CID.

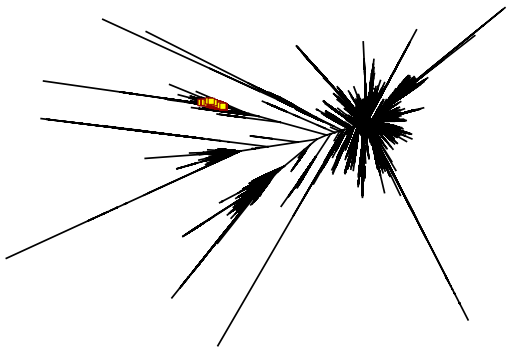


FIGURE 6 | Phylogram generated from productive reads of one of the DNP-immunized animals. The peptide ASGYTFTSYYIGWIK was abundantly found in affinity-enriched immunoglobulins in the proteomics data. NGS reads that contain peptide motif were highlighted in red in the phylogram, and show a set of DNP-specific reads in relation to the overall repertoire. Figures for the other animals in this group are in Figure S7 in Supplementary Material.

DISCUSSION

The data presented combine NGS and proteomics analysis to show that immune responses result in antibody sequence fragments that are shared among subjects exposed to the same immunogen. The new NGS data provide a much deeper view on the immune repertoire, as well as an improved sequence accuracy. The proteomics data, however, still allows us to focus on an antigen-specific subset of immunoglobulin sequences.

Similarity

A cluster analysis of the NGS data confirmed that, indeed, the immune repertoire of animals exposed to the same antigen contains similarities that allow them to be grouped accordingly. While this agrees with a similar analysis on proteomics data, this finding is nevertheless remarkable. First, the NGS data are based on the entire splenocyte repertoire and not on an antigen-specific subset of it. The majority of the repertoire is expected to relate to antigens other than the immunogen, but as the animals were treated the same except for the immunogen the antigen-specific response is the most likely component that drives the clustering. Second, it was found that sample clustering could be based on all complementarity-determining regions in the immunoglobulin molecule, including the CDR3. The latter was poorly represented in the proteomics dataset, and is considered the most diverse region of the molecule, driven by the random recombination of V, D, and J segments. Still, clustering of samples based on CDR3 sequences followed similar patterns as for the other CDRs, showing that homology and convergence occur in all of them.

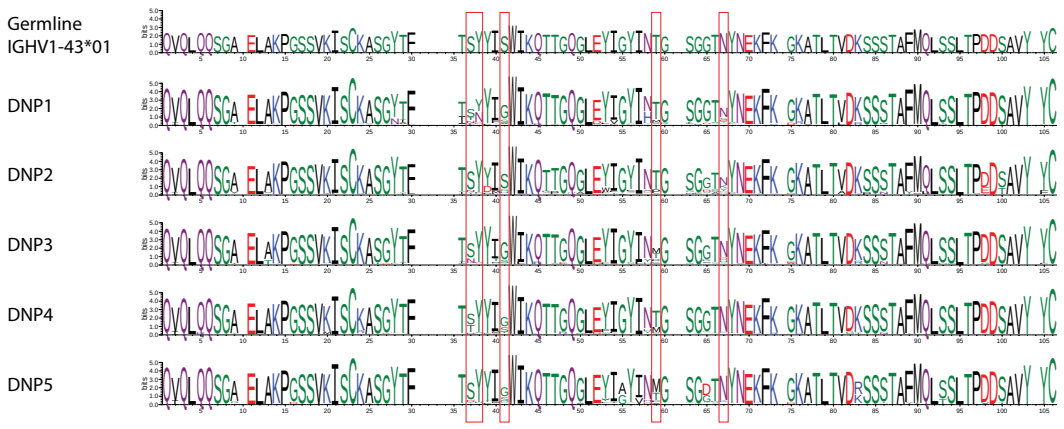


FIGURE 7 | Weblogo plots of the sequences contained in the phylogram branch encompassing all nodes highlighted in red in Figure S7 in Supplementary Material. A consensus sequence was extracted from the plots, and the most homologous germline sequence was determined with the DomainGapAlign function provided by IMGT. The sequence of this was included as a similar weblogo in the Figure as a reference.

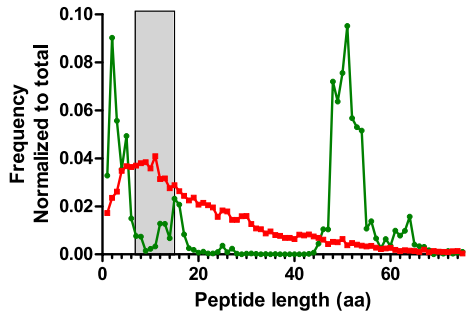


FIGURE 8 | Length distribution of tryptic peptides encompassing the CDR3 region obtained by *in silico* digestion of reads combined from all samples (green). The NGS data were appended with rat IgG2B sequence in order to include the first tryptic site located in the CH1 domain. For comparison, a similar digestion was performed on the human subset of the Uniprot protein database (red). The peptide size range that is optimal for shotgun proteomics analysis (7–15aa) has been highlighted in gray.

Shared Motifs

The repertoires of the samples were not only similar, many sequences could be identified that were exact matches between several animals in the dataset. The sharing of sequences in adaptive immune responses has been controversial in literature. While sharing has frequently been observed, interpretation whether this is a stochastic event driven by random chance, an experimental artifact or a selective process that demonstrates convergence in the immune repertoire remains under debate (11–13, 27–29). We looked into the NGS data for evidence of index hopping, an artifact that might explain shared sequences. However, we did not find reads with the same sequence and UMI barcode in different samples, indicating that the dual-index library preparation successfully prevented this problem. From the current dataset, we conclude that both chance and selection contribute to the sharing of sequences. Observations that support stochastic contributions are the decrease in the numbers of shared sequences as a function of the number of animals in which it is observed, the presence of shared sequences among animals treated with different immunogens, and shorter size distribution of CDR3 sequences in shared subsets compared to the total distribution of CDR3 sizes. On the other hand, the data show that sequence sharing is consistently higher among samples that share the same immunogen, and that samples can be properly grouped based on sequence similarities in their immune repertoires. Such observations support the view that repertoire convergence does indeed occur for the antigens that were investigated in this work.

Studies in the literature typically report convergence and sharing of the CDR3 region of the heavy chain, as this is known as the most diverse and most decisive for antigen specificity (13, 30). Here, we extend the analysis to the entire length of the variable domain and find that shared motifs can also be identified in and around the other two CDRs in the molecule. While the CDR3 may be the primary determinant of immunoglobulin specificity, finding that CDRs 1 and 2 similarly converge in response to an antigen indicates that they are still subject to somatic mutations and selection pressures and, thus, have a considerable contribution to binding (31).

Antigen Dependence

Repertoire convergence has been reported for a limited but increasing number of antigens and conditions, including HIV, tetanus toxoid, *Streptococcus pneumoniae*, *Haemophilus influenzae*, Dengue fever and Sjögren's syndrome (9, 11, 12, 32, 33). We now add evidence for the HuD antigen and DNP epitopes and have done so with two independent techniques. The question arises whether such sharing can be expected from a majority of antigens, or that it should be considered an exception. The only definitive answer to this question can be provided by collecting a large collection of repertoire data for different individuals and antigens. While such comprehensive data are still lacking, the number of datasets showing sharing or convergence is increasing while datasets that demonstrate absence thereof remain lacking. We, therefore, expect the sharing of motifs in the immune repertoire to be a more general phenomenon. We did find differences between both antigens in terms of the degree of sequence sharing between the repertoires, with the HuD antigen consistently sharing more sequence motifs than the DNP antigen. The HuD antigen is bigger and can expose more epitopes than the small DNP group. However, this should not have a big effect, as the NGS data consider not a specific subset but all antibodies in the repertoire, which may also include those targeting the KLH carrier protein used for DNP. Potential artifacts that could explain our observations were investigated, but no evidence was found for problems in index cross-contamination, or for significant differences in the number of reads for the samples. It is, therefore, more plausible that antigens differ intrinsically in the amount of sequence sharing that they elicit. It is conceivable that the fraction of distinct VDJ rearrangements that leads to specific binding early in the B-cell response inversely correlates with the degree of sharing between individuals, but a more extensive dataset covering more than two antigens is required to support that concept with evidence.

Immuno Proteogenomics

The combined use of proteomics and NGS was intended to obtain complementary information on the state of the immune repertoire. Indeed, it was possible to find evidence for repertoire convergence in both datasets, and specific observations could be cross-validated. On the other hand, the degree of quantitative correlation between the datasets was limited, and the fraction of MS–MS spectra that could be matched to NGS reads was relatively low. This suggests that the repertoires of expressed serum IgG and splenocyte mRNA overlap, but only partially. It has been reported that proteomics identification of immunoglobulins is more challenging than that of other proteomes (34). Also, as shown in **Figure 8**, workflows using trypsin have challenges for proteomics data that cover the CDR3. While specific peptide size distributions may shift for other immunoglobulin classes due to their CH1 sequence and also for other species, we expect that the length of CDR3 tryptic peptides can be a common problem. We did not observe an

improvement of correlation between proteomics and NGS in a subset of high scoring PSMs from the dataset, and we, therefore, do not believe that misidentification plays an important role in the discrepancies that are observed. Rather, we suspect the expression of serum immunoglobulins in niches that were not covered by RNA sequencing, such as bone marrow plays an important role, as well as the sequencing of RNA of other cell lineages that does not encode immunoglobulins, but rather the B-cell receptor. Refinement of experimental choices and protocols should bring NGS and proteomics data closer into alignment, and indeed several successful datasets have been described in the literature (7, 8, 35, 36).

Potential Applications

While it is clear that many sequences are shared among animals exposed to the same antigen, the data also show that the degree of sharing drops in a log-linear fashion as a function of the number of sharing individuals. This implies that it is unlikely that any one sequence motif can function as a marker for a large population of subjects. If this would be possible, a simple (proteomics) assay for the presence of such an amino acid motif could function as a proxy for a variety of immunological conditions, such as response to a vaccination (37), pathogen infection, auto-immune disease, or a response against tumor-associated antigens. While such an application based on a single peptide seems unlikely based on the current dataset, it is still conceivable that a panel of peptides or peptide homology to a conserved motif could fulfill such a purpose. A test based on composite markers would be more complex, and proper discovery and validation of such markers would require more extensive datasets that protect against false discovery and give sufficient confidence that a candidate motif indeed is predictive of an immune response in a larger population.

Other applications may include the identification of specific clonal expansions in immune sera for the production of monoclonal antibodies, or characterization of immune responses that target pathogens or in auto-immune conditions. In recent work, it has been shown that in the T-cell receptor, sequences are not only shared between individuals, but it is even feasible to predict epitope specificity based on new unseen sequence data (38, 39). Extension of such methods to immunoglobulin repertoires could accelerate the development of new applications of repertoire analysis.

An important aspect of future work is the location of sampling for repertoire analysis. Lymphoid organs and lymphocytes infiltrating at disease locations are very much of interest, but may be unrealistic sampling locations for many human applications. A comparison of such sampling locations with more readily accessible PBMCs, also in relation to time after antigen exposure, could help decision making on the best strategy in the development of new applications. Such an analysis could also include a comparison on which combination of cellular compartments best reflects immunoglobulin proteins that can be observed in serum, and which combination of compartments is most representative for the immunology of localized disease processes in an organism. It is also conceivable that the sharing of sequences that we observe in spleen and serum is more pronounced when considering a distinct niche, for example local to the disease. While modern NGS and proteomics techniques provide rapidly expanding views on the makeup of the adaptive immune response, the underlying processes and dynamics remain incompletely understood.

AUTHOR CONTRIBUTIONS

MV contributed to conception, NGS amplicon generation, proteomics sample preparation, data analysis, and drafting of the manuscript; LD contributed to conception and proteomics data acquisition; WI developed the customized library preparation and was responsible for NGS data acquisition; PS-S contributed to conception of the work; TL contributed to conception and drafting of the manuscript. All the authors critically read, contributed, and approved the manuscript.

ACKNOWLEDGMENTS

We wish to acknowledge the discussions with Dr. Pim French, Maurice de Wit, and Hanna IJspeert as well as the resources provided by the Erasmus MC Cancer Computational Biology Center.

FUNDING

This work was supported by Eurostars program E10054 (Effibody) and by the Open Technology Programme (project 14325) which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/fimmu.2017.01286/full#supplementary-material.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302(5909):575–81. doi:10.1038/302575a0
2. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol* (2017) 35(3):203–14. doi:10.1016/j.tibtech.2016.09.010
3. Galson JD, Clutterbuck EA, Truck J, Ramasamy MN, Munz M, Fowler A, et al. BCR repertoire sequencing: different patterns of B-cell activation after two meningococcal vaccines. *Immunol Cell Biol* (2015) 93(10):885–95. doi:10.1038/icb.2015.57
4. Obermeier B, Mentele R, Malotka J, Kellermann J, Kümpfel T, Wekerle H, et al. Matching of oligoclonal immunoglobulin transcriptomes and proteomes of cerebrospinal fluid in multiple sclerosis. *Nat Med* (2008) 14(6):688–93. doi:10.1038/nm1714
5. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32(2):158–68. doi:10.1038/nbt.2782
6. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Curr Opin Chem Biol* (2015) 24:112–20. doi:10.1016/j.cbpa.2014.11.007

7. Ogishi M, Yotsuyanagi H, Moriya K, Koike K. Delineation of autoantibody repertoire through differential proteogenomics in hepatitis C virus-induced cryoglobulinemia. *Sci Rep* (2016) 6:29532. doi:10.1038/srep29532
8. Chen J, Zheng Q, Hammers CM, Ellebrecht CT, Mukherjee EM, Tang HY, et al. Proteomic analysis of pemphigus autoantibodies indicates a larger, more diverse, and more dynamic repertoire than determined by B cell genetics. *Cell Rep* (2017) 18(1):237–47. doi:10.1016/j.celrep.2016.12.013
9. Zhou J, Lottenbach KR, Barenkamp SJ, Lucas AH, Reason DC. Recurrent variable region gene usage and somatic mutation in the human antibody response to the capsular polysaccharide of *Streptococcus pneumoniae* type 23F. *Infect Immun* (2002) 70(8):4083–91. doi:10.1128/IAI.70.8.4083-4091.2002
10. VanDuijn MM, Dekker LJ, Zeneyedpour L, Smitt PA, Luider TM. Immune responses are characterized by specific shared immunoglobulin peptides that can be detected by proteomic techniques. *J Biol Chem* (2010) 285(38):29247–53. doi:10.1074/jbc.M110.139071
11. Arentz G, Thurgood LA, Lindop R, Chataway TK, Gordon TP. Secreted human Ro52 autoantibody proteomes express a restricted set of public clonotypes. *J Autoimmun* (2012) 39(4):466–70. doi:10.1016/j.jaut.2012.07.003
12. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) 13(6):691–700. doi:10.1016/j.chom.2013.05.008
13. Truck J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J Immunol* (2015) 194(1):252–61. doi:10.4049/jimmunol.1401405
14. Wang C, Liu Y, Cavanagh MM, Le Saux S, Qi Q, Roskin KM, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc Natl Acad Sci U S A* (2015) 112(2):500–5. doi:10.1073/pnas.1415875112
15. Leypoldt F, Wandinger KP. Paraneoplastic neurological syndromes. *Clin Exp Immunol* (2014) 175(3):336–48. doi:10.1111/cei.12185
16. Linhart C, Shamir R. The degenerate primer design problem. *Bioinformatics* (2002) 18(Suppl 1):S172–81. doi:10.1093/bioinformatics/18.suppl_1.S172
17. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) 11(6):653–5. doi:10.1038/nmeth.2960
18. Brochet X, Lefranc MP, Giudicelli V. 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* (2008) 36(Web Server issue):W503–8. doi:10.1093/nar/gkn316
19. Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, et al. VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput Biol* (2015) 11(11):e1004503. doi:10.1371/journal.pcbi.1004503
20. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics* (2015) 16:175. doi:10.1186/s12859-015-0613-1
21. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* (2009) 10:356. doi:10.1186/1471-2105-10-356
22. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* (2010) 5(3):e9490. doi:10.1371/journal.pone.0009490
23. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* (2014) 32(3):223–6. doi:10.1038/nbt.2839
24. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* (2002) 30(1):207–10. doi:10.1093/nar/30.1.207
25. Yu L, Guan Y. Immunologic basis for long HCDR3s in broadly neutralizing antibodies against HIV-1. *Front Immunol* (2014) 5:250. doi:10.3389/fimmu.2014.00250
26. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* (2004) 14(6):1188–90. doi:10.1101/gr.849004
27. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) 324(5928):807–10. doi:10.1126/science.1170020
28. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110(33):13463–8. doi:10.1073/pnas.1312146110
29. Hoehn KB, Fowler A, Lunter G, Pybus OG. The diversity and molecular evolution of B-cell receptors during infection. *Mol Biol Evol* (2016) 33(5):1147–57. doi:10.1093/molbev/msw015
30. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* (2000) 13(1):37–45. doi:10.1016/S1074-7613(00)00006-6
31. Chen C, Roberts VA, Rittenberg MB. Generation and analysis of random point mutations in an antibody CDR2 sequence: many mutated antibodies lose their ability to bind antigen. *J Exp Med* (1992) 176(3):855–66. doi:10.1084/jem.176.3.855
32. Poulsen TR, Meijer PJ, Jensen A, Nielsen LS, Andersen PS. Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid. *J Immunol* (2007) 179(6):3841–50. doi:10.4049/jimmunol.179.6.3841
33. Gorny MK, Wang XH, Williams C, Volsky B, Revesz K, Witover B, et al. Preferential use of the VH5-51 gene segment by the human immune response to code for antibodies against the V3 domain of HIV-1. *Mol Immunol* (2009) 46(5):917–26. doi:10.1016/j.molimm.2008.09.005
34. Boutz DR, Horton AP, Wine Y, Lavinder JJ, Georgiou G, Marcotte EM. Proteomic identification of monoclonal antibodies from serum. *Anal Chem* (2014) 86(10):4758–66. doi:10.1021/ac4037679
35. Sato S, Beausoleil SA, Popova L, Beaudet JG, Ramenani RK, Zhang X, et al. Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nat Biotechnol* (2012) 30(11):1039–43. doi:10.1038/nbt.2406
36. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* (2014) 111(6):2259–64. doi:10.1073/pnas.1317793111
37. Galson JD, Pollard AJ, Truck J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol* (2014) 35(7):319–31. doi:10.1016/j.it.2014.04.005
38. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* (2017) 547(7661):89–93. doi:10.1038/nature22383
39. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* (2017) 547(7661):94–98. doi:10.1038/nature22976

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.

The reviewer AH and handling editor declared their shared affiliation.

Copyright © 2017 VanDuijn, Dekker, van IJcken, Sillevis Smitt and Luider. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

PERSPECTIVE
published: 01 November 2017
doi: 10.3389/fimmu.2017.01418

Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data

Felix Breden[1], Eline T. Luning Prak[2]*, Bjoern Peters[3], Florian Rubelt[4], Chaim A. Schramm[5], Christian E. Busse[6], Jason A. Vander Heiden[7], Scott Christley[8], Syed Ahmad Chan Bukhari[9], Adrian Thorogood[10], Frederick A. Matsen IV[11], Yariv Wine[12], Uri Laserson[13], David Klatzmann[14], Daniel C. Douek[5], Marie-Paule Lefranc[15], Andrew M. Collins[16], Tania Bubela[17], Steven H. Kleinstein[9], Corey T. Watson[18], Lindsay G. Cowell[8], Jamie K. Scott[19] and Thomas B. Kepler[20,21]*

[1] Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada, [2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [3] La Jolla Institute for Allergy and Immunology, La Jolla, CA, United States, [4] Department of Microbiology and Immunology, Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA, United States, [5] Vaccine Research Center, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Bethesda, MD, United States, [6] Division of B Cell Immunology, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany, [7] Department of Neurology, Yale University School of Medicine, New Haven, CT, United States, [8] Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, United States, [9] Department of Pathology, Yale University School of Medicine, New Haven, CT, United States, [10] Centre of Genomics and Policy, McGill University, Montreal, QC, Canada, [11] Public Health Sciences Division and Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, United States, [12] Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel, [13] Department of Genetics and Genome Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [14] Immunology-Immunopathology-Immunotherapy (i3 & i2B), Sorbonne Université, Paris, France, [15] IMGT, LIGM, Institut de Génétique Humaine IGH, CNRS, University of Montpellier, Montpellier, France, [16] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Kensington, NSW, Australia, [17] Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada, [18] Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, United States, [19] Faculty of Health Sciences, Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada, [20] Department of Microbiology, Boston University School of Medicine, Boston, MA, United States, [21] Department of Mathematics and Statistics, Boston University, Boston, MA, United States*

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Michael Zemlin, Universitätsklinikum des Saarlandes, Germany
Deborah K. Dunn-Walters, University of Surrey, United Kingdom

***Correspondence:**
Felix Breden
breden@sfu.ca;
Eline T. Luning Prak
luning@pennmedicine.upenn.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *06 September 2017*
Accepted: *12 October 2017*
Published: *01 November 2017*

Citation:
Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, Vander Heiden JA, Christley S, Bukhari SAC, Thorogood A, Matsen IV FA, Wine Y, Laserson U, Klatzmann D, Douek DC, Lefranc M-P, Collins AM, Bubela T, Kleinstein SH, Watson CT, Cowell LG, Scott JK and Kepler TB (2017) Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. Front. Immunol. 8:1418. doi: 10.3389/fimmu.2017.01418

High-throughput sequencing (HTS) of immunoglobulin (B-cell receptor, antibody) and T-cell receptor repertoires has increased dramatically since the technique was introduced in 2009 (1–3). This experimental approach explores the maturation of the adaptive immune system and its response to antigens, pathogens, and disease conditions in exquisite detail. It holds significant promise for diagnostic and therapy-guiding applications. New technology often spreads rapidly, sometimes more rapidly than the understanding of how to make the products of that technology reliable, reproducible, or usable by others. As complex technologies have developed, scientific communities have come together to adopt common standards, protocols, and policies for generating and sharing data sets, such as the MIAME protocols developed for microarray experiments. The Adaptive Immune Receptor Repertoire (AIRR) Community formed in 2015 to address similar issues for HTS data of immune repertoires. The purpose of this perspective is to provide an overview of the AIRR Community's founding principles and present the progress that the AIRR Community has made in developing standards of practice and data sharing protocols. Finally, and most important, we invite all interested parties to join this effort to facilitate sharing and use of these powerful data sets (join@airr-community.org).

Keywords: B-cell receptors, T-cell receptors, data sharing, immunogenetics, community standards, high-throughput sequencing, immunoglobulins, antibodies

INTRODUCTION

The adaptive immune system provides protection against disease without inducing harmful autoimmunity; it reacts against the vast and ever-changing array of pathogens that an individual will encounter over a lifetime, while tolerating self. The variable regions of the adaptive immune receptors on B cells and T cells arise through the rearrangement of germline variable, diversity, and joining gene segments (4, 5). Humans each express over 100 million unique immunoglobulins (6) and a similar number of T-cell receptors (1, 7). The lymphocytes that express these receptors arise, proliferate, and die on time scales of hours to years (1, 8). Thus, the collection of B-cell and T-cell receptor variable region genes expressed at any given time—the adaptive immune receptor repertoire (AIRR)—is dynamic.

Immunoglobulin and T-cell receptor sequences have been studied for decades and several established databases exist including Kabat–Wu and Vbase2 (9, 10). Furthermore, there are databases that incorporate or allow viewing of structural data, such as IMGT, IEDB-3D, AntigenDB, and SAbDab [reviewed in Ref. (11)]. These data sets provide important insights into immune receptor–antigen interactions and can inform antibody engineering efforts. However, a single immunoglobulin or T-cell receptor sequence is but a drop of water in the ocean that is the immune repertoire. While many immune repertoire studies have been performed using a variety of methods [reviewed in Ref. (12)], adequate analysis of the repertoire as a whole was virtually impossible prior to the advent of high-throughput sequencing (HTS). Here, we focus on HTS-based profiling of AIRR.

Since HTS was first applied to AIRR profiling in 2009 (1, 3, 6, 7), there has been rapid advancement of both experimental and computational techniques. HTS of AIRRs (AIRR-seq) is yielding valuable insights into how variation in the AIRR differs across lymphocyte subsets (13–16) and anatomic compartments (17–20), varies over the course of a disease or with therapy (21–27), and is influenced by age (28–32), genetic background (33, 34), health status (19, 29, 35–37), antigen exposure (27, 38–40), and other factors. AIRR-seq data are increasingly important in the development of vaccines, monoclonal antibodies, cancer immunotherapies, and other applications [reviewed in Ref. (41)]. As the number of datasets continues to grow, comparative analyses of hundreds or even thousands of individuals will soon be feasible. Ensuring the reliability of such integrative analyses, however, will require the establishment of and adherence to standards for reporting and sharing data across multiple laboratories and centers.

CHALLENGES FOR AIRR-SEQ DATA SHARING

Several challenges currently impede the effective sharing of AIRR-seq data. First, the storage and transport of such large datasets, which can comprise hundreds of millions of sequences (and hundreds of gigabytes) per study, require substantial time and resources. Second, deposition into public archives is not uniformly required by journals or funding agencies. As of September 4, 2017, a Wiki page on the B-T.CR forum[1] lists 82 AIRR-seq studies that report full HTS data to a public archive,[2] while 42 (34%) do not.[3] Third, the information required to ensure appropriate use of such data by secondary users requires delineation (42). These challenges are not unique to AIRR-seq data. Indeed, the need for shared standards has been recognized and addressed for previous high-throughput technologies (43), including microarray data (44).

Another significant challenge for AIRR-seq data is that the processing pipeline between the experiment and the ultimate analysis of the data is lengthy and specialized (45–59). Beyond the steps required to process any HTS data, the annotation required of AIRR-seq data is unique to these genes and subject to substantial uncertainty (52). Unlike other genes, the antigen receptors of adaptive immunity are assembled through the recombination of randomly chosen gene segments, with non-templated nucleotides added to the junctions and nucleotides nibbled away from the gene segments (60). In B cells, somatic hypermutation during affinity maturation results in further diversification of immunoglobulin genes (61, 62). In order for these data to be effectively shared and reanalyzed, the development of new metadata standards specific to the experimental and bioinformatic methods associated with AIRR-seq are required.

A BRIEF HISTORY OF THE AIRR COMMUNITY

The AIRR Community was established in 2015 at a meeting organized by Felix Breden, Jamie Scott, and Thomas Kepler in Vancouver, BC, USA to address these data sharing challenges. Membership in the AIRR Community is open and is intended to cover all aspects of AIRR-seq technology and its uses. Membership includes researchers expert in the generation of AIRR data; statisticians and bioinformaticians versed in their analysis; informaticians and data security experts experienced in their management; basic scientists and physicians who turn to such data for critical insights; and experts in the ethical, legal, and policy implications of sharing AIRR data.

In 2015, the AIRR Community formed three Working Groups. The Minimal Standards Working Group was tasked with the development of a set of metadata standards for the publication and sharing of AIRR-seq datasets. The Tools and Resources Working Group is focused on the development of standardized resources to facilitate the comparison of AIRR-seq datasets and analysis tools, including collection, validation, and nomenclature of germline alleles. Finally, the Common Repository Working Group is working to establish requirements for repositories that will store AIRR data. The Working Groups are dynamic and often collaborate with each other, as methods evolve and applications of standards in one area (for example, metadata standards) impact other areas (data repository requirements). Full recommendations and membership lists for the Working Groups as well as video recordings of the 2016 AIRR Community meeting are

[1] https://www.b-t.cr.
[2] https://www.b-t.cr/t/317.
[3] https://www.b-t.cr/t/426.

available at http://www.airr-community.org. At the June 2016 meeting held at the National Institutes of Health (NIH), the AIRR Community ratified an initial set of recommendations that are summarized herein.

DATA GENERATION

Due to the complexity and diversity of the data sets being generated, the AIRR Community is developing best practices for the generation of AIRR-seq data. Such best practice guidelines will include, at a minimum: standard operating procedures for cell isolation and purification, including panels and gating strategies for flow cytometry; primers and protocols for amplification and sequencing of BCR or TCR rearrangements; and a clear description of library preparation and sequencing. Nomenclature is particularly important when it comes to the multiple stages of sample processing and data analysis. For example, what is meant by "raw data" differs among investigators, compounded by the fact that there are multiple levels of data preprocessing.

At present, the AIRR Community recommends that: (1) experimental protocols should be made available through a public repository granting digital object identifiers; (2) the change history of the experimental protocols, including details of what was changed and when the changes were made, should be made publicly available through the same repository; and (3) biological materials (e.g., plasmids, cell lines) should be made available to interested researchers *via* public repositories (e.g., Addgene for vectors, ATCC for cell lines), whenever possible.

DATA SHARING

For transparency and reliable reuse, experiments need to be sufficiently well annotated to allow evaluation of the quality of individual datasets and comparability of different datasets. Therefore, the AIRR Community has developed experimental metadata standards for AIRR-seq data generation, processing, and quality control. The data consist of the raw sequences and the processed sequences, while metadata include clinical and demographic data on study subjects and protocols for cell phenotyping, nucleic acid purification, AIRR amplicon production, HTS library preparation and sequencing, as well as documentation of the computational pipelines used to process the data. In publications or other forms of data sharing, these metadata sets and their components should be described in sufficient detail such that a person skilled in the art of AIRR sequencing and data analysis will be able to reproduce the experiment and data analyses that were performed. A manuscript describing a complete model for AIRR-seq data and metadata, and standardized terminology will soon be submitted.

Data sharing is also premised on the user's ability to locate and access the data. The AIRR Community recommends that all published AIRR-seq data be deposited in designated public repositories that adhere to the AIRR Community minimal standards guidelines, namely, that the data should be made available under the least restrictive terms possible. Limited exceptions to respect commercial interests in intellectual property rights are under consideration by the AIRR Community. To facilitate data sharing, the AIRR Community is also establishing an AIRR Data Commons, comprised of multiple, distributed repositories optimized for storing and querying AIRR-seq data, and supported by a centralized Gateway. Under such an intermediate distributed model (43), interoperability and effective data sharing are ensured because participating repositories will be required to comply with the community-established data and metadata standards and certain technical requirements.

LEGAL AND ETHICAL CONSIDERATIONS

Adaptive immune receptor repertoire-seq data can be subject to regulations regarding informed consent, intellectual property, and ethical treatment of research subjects. During the process of making AIRR-seq data publicly available, researchers typically would attest that they have sought appropriate informed consent or other authorization for sharing, where applicable. To reduce the potential for a breach of privacy of research subjects, medical and demographic metadata should to be structured in such a way that individual research subjects are not identifiable. Access to health information is regulated by national and international laws, such as the Health Insurance Portability and Accountability Act in the United States or the EU Regulation 2016/679 in the European Union, which requires medical information and personal identifiers to be safeguarded. For studies using AIRR-seq data from human subjects, data must be collected following a protocol that has been approved by the researcher's Institutional Review Board, which oversees human subjects' protections and ensures that all studies are performed in a legal and ethical manner (63). Human subjects must provide informed consent, and there should be broad agreement in the consent language regarding the confidentiality of medical information and the use of AIRR-seq data and metadata for future research. Without such provisions, the data in the database may be too constrained, with respect to time or breadth of investigation, to be usable by investigators other than the initial data generators.

Whenever data or other items of potential commercial value are shared with others, the individuals who generated and deposited the data should to be given proper credit. Hence, users of the database should, at a minimum, credit the data depositors in any publication or grant application. One mechanism whereby these rules could be followed is to create an online form that must be completed before access to the database is granted. Such a form would essentially be a contract for using the data. Enforcement of the terms of the contract could include monitoring of data use and denying access to the database should the terms of the data-use agreement be violated.

To facilitate broad access to and use of AIRR-seq data, the data should be made available under the least-restrictive terms possible. The default data sharing policy should be to deposit data in a public domain database with no restrictions over deposit, access, storage, curation, or use. For data deposited in public domain databases/repositories, neither the depositors nor the repositories should be permitted to interfere with access to and use of the data by others, including through the assertion of any intellectual property rights. Exceptions to open data sharing may arise in circumstances in which open data sharing would come

into conflict with the law, such as those pertaining to personal privacy and protected health information, or into conflict with decisions made by an Institutional Review Board.

DATA ANALYSIS

The AIRR Community strongly advocates the use of statistical methods for data analysis and hypothesis testing. Statistical methods systematically characterize error, quantify uncertainty, and provide a measure of confidence for inferences. Statistical methods also form the basis for data analysis in all other realms of biomedical and scientific research and should be adopted for AIRR-seq data. Expanded production of AIRR-seq data has been supported by a proliferation of computational tools for their processing and analysis, including tools for variable region gene annotation, inference of clonal history and partitioning and visualization (16, 46–52, 55, 56, 64–69). To encourage broad and well-informed use of these tools, the AIRR Community recommends that analysis software be released under an Open Source Initiative approved license, hosted on a publicly available website or repository with versioning, and be designed for modularity and inter-operability with other software. The AIRR Community will promote best practices in AIRR-seq data analysis by: (1) developing and publishing common criteria for the evaluation of statistical methods; (2) providing common "gold-standard" datasets of multiple types for use in software development, testing, and calibration; and (3) establishing best practices for data sharing and analysis software platforms.

CONCLUSION

Members of the AIRR Community have worked together for over 2 years with enthusiasm, driven by the belief that optimizing the reproducibility and sharing of AIRR-seq data will have a profound and positive effect on biomedical research and patient care. To encourage widespread adoption, the AIRR Community recommends that journals and funding agencies require AIRR-seq data be made available through a public data repository after publication or as negotiated in data-sharing agreements for unpublished data. The success of this initiative is also critically dependent upon acceptance by the researchers who generate and use AIRR-seq data. While members of the AIRR Community have tried to be inclusive through developing contacts with researchers in the field and extensively advertising the annual meetings, there are likely to be many researchers who generate, analyze, and use AIRR-seq data, who are not aware of the AIRR Community initiative. Community "buy in" results from creating data standards that are transparently developed through public discussion, robustly evaluated, and periodically updated as the field advances. This Perspective represents an open invitation to the larger scientific community to participate in and adopt the AIRR initiative. To that end, we welcome feedback on this Perspective and on the AIRR Community's efforts to date. Individuals interested in working on any facet of this important initiative are invited to attend, in person or online, the 2017 Community Meeting hosted by the NIH in Rockville, MD, USA, December 3–6, 2017. Most of all, we encourage anyone who is interested to join the AIRR Community[4] and participate in the working groups.

AUTHOR CONTRIBUTIONS

FB, EP, JS, and TK conceived of and wrote the manuscript. All other authors contributed ideas and/or proposed revisions to the text. The AIRR Community Working Groups developed and wrote the recommendations described herein.

ACKNOWLEDGMENTS

Many of the ideas presented herein evolved over the course of AIRR Community meetings and Working Group meetings. The AIRR Community initiative and Community meetings were supported by CIHR, NIH (Jon Warren and Joe Breen), NIH R13-AI116349, P01-AI106697, R01-AI097403 and P30-CA016520, GenMab, The Antibody Society, CHAVI, the IRMACS Centre, Simon Fraser University, Illumina, Genentech, TTP Labtech, Grifols, and Amgen.

[4]join@airr-community.org.

REFERENCES

1. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* (2009) 19(10):1817–24. doi:10.1101/gr.092924.109
2. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) 114(19):4099–107. doi:10.1182/blood-2009-04-217604
3. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) 324(5928):807–10. doi:10.1126/science.1170020
4. Sakano H, Maki R, Kurosawa Y, Roeder W, Tonegawa S. Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. *Nature* (1980) 286(5774):676–83. doi:10.1038/286676a0
5. Sakano H, Kurosawa Y, Weigert M, Tonegawa S. Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. *Nature* (1981) 290(5807):562–5. doi:10.1038/290562a0
6. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106(48):20216–21. doi:10.1073/pnas.0909775106
7. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol* (2013) 25(5):646–52. doi:10.1016/j.coi.2013.09.017
8. McLean AR, Michie CA. In vivo estimates of division and death rates of human T lymphocytes. *Proc Natl Acad Sci U S A* (1995) 92(9):3707–11. doi:10.1073/pnas.92.9.3707
9. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) 132(2):211–50. doi:10.1084/jem.132.2.211
10. Retter I, Althaus HH, Münch R, Müller W. VBASE2, an integrative V gene database. *Nucleic Acids Res* (2005) 33(Database issue):D671–4. doi:10.1093/nar/gki088

11. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. *Nucleic Acids Res* (2014) 42(Database issue):D1140–6. doi:10.1093/nar/gkt1043
12. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, et al. The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis. *Front Immunol* (2013) 4:413. doi:10.3389/fimmu.2013.00413
13. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, et al. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front Immunol* (2014) 5:96. doi:10.3389/fimmu.2014.00096
14. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010) 116(7):1070–8. doi:10.1182/blood-2010-03-275859
15. Martin VG, Wu YB, Townsend CL, Lu GH, O'Hare JS, Mozeika A, et al. Transitional B cells in early human B cell development – time to revisit the paradigm? *Front Immunol* (2016) 7:546. doi:10.3389/fimmu.2016.00546
16. Bashford-Rogers RJ, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) 23(11):1874–84. doi:10.1101/gr.154815.113
17. Briney BS, Willis JR, Finn JA, McKinney BA, Crowe JE Jr. Tissue-specific expressed antibody variable gene repertoires. *PLoS One* (2014) 9(6):e100839. doi:10.1371/journal.pone.0100839
18. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, et al. An atlas of B-cell clonal distribution in the human body. *Nat Biotechnol* (2017) 35(9):879–84. doi:10.1038/nbt.3942
19. Stern JN, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med* (2014) 6(248):248ra107. doi:10.1126/scitranslmed.3008879
20. Sathaliyawala T, Kubota M, Yudanin N, Turner D, Camp P, Thome JJ, et al. Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. *Immunity* (2013) 38(1):187–97. doi:10.1016/j.immuni.2012.09.020
21. Heather JM, Best K, Oakes T, Gray ER, Roe JK, Thomas N, et al. Dynamic perturbations of the T-cell receptor repertoire in chronic HIV infection and following antiretroviral therapy. *Front Immunol* (2015) 6:644. doi:10.3389/fimmu.2015.00644
22. Racanelli V, Brunetti C, De Re V, Caggiari L, De Zorzi M, Leone P, et al. Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS One* (2011) 6(9):e25606. doi:10.1371/journal.pone.0025606
23. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* (2012) 120(26):5173–80. doi:10.1182/blood-2012-07-444042
24. Weng WK, Armstrong R, Arai S, Desmarais C, Hoppe R, Kim YH. Minimal residual disease monitoring with high-throughput sequencing of T cell receptors in cutaneous T cell lymphoma. *Sci Transl Med* (2013) 5(214):214ra171. doi:10.1126/scitranslmed.3007420
25. Kalos M, Levine BL, Porter DL, Katz S, Grupp SA, Bagg A, et al. T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. *Sci Transl Med* (2011) 3(95):95ra73. doi:10.1126/scitranslmed.3002842
26. Morris H, DeWolf S, Robins H, Sprangers B, LoCascio SA, Shonts BA, et al. Tracking donor-reactive T cells: evidence for clonal deletion in tolerant kidney transplant patients. *Sci Transl Med* (2015) 7(272):272ra10. doi:10.1126/scitranslmed.3010760
27. Havenar-Daughton C, Carnathan DG, Torrents de la Peña A, Pauthner M, Briney B, Reiss SM, et al. Direct probing of germinal center responses reveals immunological features and bottlenecks for neutralizing antibody responses to HIV env trimer. *Cell Rep* (2016) 17(9):2195–209. doi:10.1016/j.celrep.2016.10.085
28. Russell Knode LM, Naradikian MS, Myles A, Scholz JL, Hao Y, Liu D, et al. Age-associated B cells express a diverse repertoire of VH and Vkappa genes with somatic hypermutation. *J Immunol* (2017) 198(5):1921–7. doi:10.4049/jimmunol.1601106
29. Gibson KL, Wu YC, Barnett Y, Duggan O, Vaughan R, Kondeatis E, et al. B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* (2009) 8(1):18–25. doi:10.1111/j.1474-9726.2008.00443.x
30. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci U S A* (2014) 111(36):13139–44. doi:10.1073/pnas.1409155111
31. Rechavi E, Lev A, Lee YN, Simon AJ, Yinon Y, Lipitz S, et al. Timely and spatially regulated maturation of B and T cell repertoire during human fetal development. *Sci Transl Med* (2015) 7(276):276ra25. doi:10.1126/scitranslmed.aaa0072
32. Guo C, Wang Q, Cao X, Yang Y, Liu X, An L, et al. High-throughput sequencing reveals immunological characteristics of the TRB-/IgH-CDR3 region of umbilical cord blood. *J Pediatr* (2016) 176:69–78.e1. doi:10.1016/j.jpeds.2016.05.078
33. Notarangelo LD, Kim MS, Walter JE, Lee YN. Human RAG mutations: biochemistry and clinical implications. *Nat Rev Immunol* (2016) 16(4):234–46. doi:10.1038/nri.2016.28
34. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) 92(4):530–46. doi:10.1016/j.ajhg.2013.03.004
35. Tipton CM, Fucile CF, Darce J, Chida A, Ichikawa T, Gregoretti I, et al. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat Immunol* (2015) 16(7):755–65. doi:10.1038/ni.3175
36. Stamatopoulos K, Belessi C, Moreno C, Boudjograh M, Guida G, Smilevska T, et al. Over 20% of patients with chronic lymphocytic leukemia carry stereotyped receptors: pathogenetic implications and clinical correlations. *Blood* (2007) 109(1):259–70. doi:10.1182/blood-2006-03-012948
37. Rubelt F, Bolen CR, McGuire HM, Vander Heiden JA, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. *Nat Commun* (2016) 7:11112. doi:10.1038/ncomms11112
38. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci U S A* (2014) 111(13):4928–33. doi:10.1073/pnas.1323862111
39. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110(33):13463–8. doi:10.1073/pnas.1312146110
40. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* (2014) 111(6):2259–64. doi:10.1073/pnas.1317793111
41. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32(2):158–68. doi:10.1038/nbt.2782
42. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public for the future of immunology. *Immunol Res* (2014) 58(2–3):234–9. doi:10.1007/s12026-014-8516-1
43. Contreras JL, Reichman JH. DATA ACCESS. Sharing by design: data and decentralized commons. *Science* (2015) 350(6266):1312–4. doi:10.1126/science.aaa7485
44. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* (2001) 29(4):365–71. doi:10.1038/ng1201-365
45. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* (2015) 7:121. doi:10.1186/s13073-015-0243-2
46. Imkeller K, Arndt PF, Wardemann H, Busse CE. sciReptor: analysis of single-cell level immunoglobulin repertoires. *BMC Bioinformatics* (2016) 17:67. doi:10.1186/s12859-016-0920-1
47. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. ImmuneDB: a system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* (2017) 33(2):292–3. doi:10.1093/bioinformatics/btw593

48. Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, et al. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front Immunol* (2012) 3:176. doi:10.3389/fimmu.2012.00176
49. Ralph DK, Matsen FA IV. Likelihood-based inference of B cell clonal families. *PLoS Comput Biol* (2016) 12(10):e1005086. doi:10.1371/journal.pcbi.1005086
50. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* (2015) 12(5):380–1. doi:10.1038/nmeth.3364
51. Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, et al. VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput Biol* (2015) 11(11):e1004503. doi:10.1371/journal.pcbi.1004503
52. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res* (2013) 2:103. doi:10.12688/f1000research.2-103.v1
53. Kepler TB, Munshaw S, Wiehe K, Zhang R, Yu JS, Woods CW, et al. Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *Front Immunol* (2014) 5:170. doi:10.3389/fimmu.2014.00170
54. Liberman G, Benichou JI, Maman Y, Glanville J, Alter I, Louzoun Y. Estimate of within population incremental selection through branch imbalance in lineage trees. *Nucleic Acids Res* (2016) 44(5):e46. doi:10.1093/nar/gkv1198
55. Vincent B, et al. iWAS – a novel approach to analyzing next generation sequence data for immunology. *Cell Immunol* (2016) 299:6–13. doi:10.1016/j.cellimm.2015.10.012
56. Volpe JM, Cowell LG, Kepler TB. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics* (2006) 22(4):438–44. doi:10.1093/bioinformatics/btk004
57. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
58. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41(Web Server issue):W34–40. doi:10.1093/nar/gkt382
59. Zhang B, Meng W, Luning Prak ET, Hershberg U. Discrimination of germline V genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment. *J Immunol Methods* (2015) 427:105–16. doi:10.1016/j.jim.2015.10.009
60. Gellert M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem* (2002) 71:101–32. doi:10.1146/annurev.biochem.71.090501.150203
61. Weigert MG, Cesari IM, Yonkovich SJ, Cohn M. Variability in the lambda light chain sequences of mouse antibody. *Nature* (1970) 228(5276):1045–7. doi:10.1038/2281045a0
62. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. *Nature* (1991) 354(6352):389–92. doi:10.1038/354389a0
63. Freedman RS, Cantor SB, Merriman KW, Edgerton ME. 2013 HIPAA changes provide opportunities and challenges for researchers: perspectives from a cancer center. *Clin Cancer Res* (2016) 22(3):533–9. doi:10.1158/1078-0432.CCR-15-2155
64. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical clustering can identify B cell clones with high confidence in ig repertoire sequencing data. *J Immunol* (2017) 198(6):2489–99. doi:10.4049/jimmunol.1601850
65. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) 31(20):3356–8. doi:10.1093/bioinformatics/btv359
66. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30(13):1930–2. doi:10.1093/bioinformatics/btu138
67. Ostmeyer J, Christley S, Rounds WH, Toby I, Greenberg BM, Monson NL, et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics* (2017) 18(1):401. doi:10.1186/s12859-017-1814-6
68. Toby IT, Levin MK, Salinas EA, Christley S, Bhattacharya S, Breden F, et al. VDJML: a file format with tools for capturing the results of inferring immune receptor rearrangements. *BMC Bioinformatics* (2016) 17(Suppl 13):333. doi:10.1186/s12859-016-1214-3
69. Christley S, Levin MK, Toby I, Fonner J, Monson N, Rounds WH, et al. VDJPipe: a pipelined tool for pre-processing immune repertoire sequencing data. *BMC Bioinformatics* (2017) 18(1):448. doi:10.1186/s12859-017-1853-z

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.

Copyright © 2017 Breden, Luning Prak, Peters, Rubelt, Schramm, Busse, Vander Heiden, Christley, Bukhari, Thorogood, Matsen IV, Wine, Laserson, Klatzmann, Douek, Lefranc, Collins, Bubela, Kleinstein, Watson, Cowell, Scott and Kepler. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 13 November 2017
doi: 10.3389/fimmu.2017.01433

Antibody Heavy Chain Variable Domains of Different Germline Gene Origins Diversify through Different Paths

Ufuk Kirik[1†], Helena Persson[2], Fredrik Levander[1,3], Lennart Greiff[4,5] and Mats Ohlin[1,6,7]*

[1] Department of Immunotechnology, Lund University, Lund, Sweden, [2] Science for Life Laboratory, Drug Discovery and Development Platform, School of Biotechnology, KTH Royal Institute of Technology, Stockholm, Sweden, [3] National Bioinformatics Infrastructure Sweden (NBIS), Science for Life Laboratory, Department of Immunotechnology, Lund University, Lund, Sweden, [4] Department of Clinical Sciences, Lund University, Lund, Sweden, [5] Department of Otorhinolaryngology, Head and Neck Surgery, Skåne University Hospital, Lund, Sweden, [6] Science for Life Laboratory, Drug Discovery and Development Platform, Human Antibody Therapeutics, Lund University, Lund, Sweden, [7] U-READ, Lund School of Technology, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:
Jacob Glanville, Distributed Bio, Inc., United States

Reviewed by:
Claude-Agnes Reynaud, Institut national de la santé et de la recherche médicale, France
Michael Zemlin, Universitätsklinikum des Saarlandes, Germany

***Correspondence:**
Mats Ohlin
mats.ohlin@immun.lth.se

†Present address:
Ufuk Kirik, Disease Systems Biology Program, University of Copenhagen, Copenhagen, Denmark

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *28 June 2017*
Accepted: *16 October 2017*
Published: *13 November 2017*

Citation:
Kirik U, Persson H, Levander F, Greiff L and Ohlin M (2017) Antibody Heavy Chain Variable Domains of Different Germline Gene Origins Diversify through Different Paths. Front. Immunol. 8:1433. doi: 10.3389/fimmu.2017.01433

B cells produce antibodies, key effector molecules in health and disease. They mature their properties, including their affinity for antigen, through hypermutation events; processes that involve, e.g., base substitution, codon insertion and deletion, often in association with an isotype switch. Investigations of antibody evolution define modes whereby particular antibody responses are able to form, and such studies provide insight important for instance for development of efficient vaccines. Antibody evolution is also used *in vitro* for the design of antibodies with improved properties. To better understand the basic concepts of antibody evolution, we analyzed the mutational paths, both in terms of amino acid substitution and insertions and deletions, taken by antibodies of the IgG isotype. The analysis focused on the evolution of the heavy chain variable domain of sets of antibodies, each with an origin in 1 of 11 different germline genes representing six human heavy chain germline gene subgroups. Investigated genes were isolated from cells of human bone marrow, a major site of antibody production, and characterized by next-generation sequencing and an in-house bioinformatics pipeline. Apart from substitutions within the complementarity determining regions, multiple framework residues including those in protein cores were targets of extensive diversification. Diversity, both in terms of substitutions, and insertions and deletions, in antibodies is focused to different positions in the sequence in a germline gene-unique manner. Altogether, our findings create a framework for understanding patterns of evolution of antibodies from defined germline genes.

Keywords: antibody germline gene, antibody sequence, somatic hypermutation, immunoglobulin, insertion and deletion, substitution

INTRODUCTION

Antibodies, central components of humoral immunity, are crucial to our survival. The immune system allows antibodies to evolve in efforts to enhance their ability to mediate protection against disease. The biochemistry and mechanism of this complex evolution process has been extensively studied at the molecular level (1). Technological advances in sequencing and single cell analysis

Abbreviations: BM, bone marrow; CDR, complementarity determining region; FR, framework region; H, heavy; V, variable; VH, heavy chain variable; VL, light chain variable.

technology have recently allowed us to, at great depth, study antibody sequences as they develop *in vivo* (2). Indeed, various sequencing strategies and bioinformatics pipelines have been generated to allow such analysis (3, 4).

Studies of the development of humoral immune responses require knowledge of the repertoire of genes that are available in the genome. Such information allows us to properly analyze germline gene rearrangement events and hypermutation, as exemplified by extensive studies of the response against the envelope protein of HIV-1 (5, 6). Databases and associated analysis tools, like IMGT/IMGT V-QUEST/IMGT HighV-QUEST (7), have consequently been built to allow efficient analysis of antibody-encoding sequences, their genetic origin, and their evolution. Common concepts, like standardized framework and complementarity determining regions (FR and CDR, respectively), the latter of which is considered to represent the antigen-contacting part of the antibody, are commonly used in such analysis. However, numerous definitions of these regions exist in parallel (8–13), highlighting the difficulties associated with the establishment of a clear-cut definition of these regions. We hypothesized that a thorough understanding of the ways through which antibodies derived from different germline genes evolve as a consequence of somatic mutation processes will aid the establishment of such definitions. Such understanding will also aid a proper mutational analysis of clones that populate immune responses.

In this study, we have focused our attention to human IgG encoded by unsorted cells in bone marrow (BM) (14), a major site of antibody production, to define how evolution proceeds in antibody heavy (H) chains derived from 11 commonly used germline genes. The advent of high throughput next-generation sequencing methodology, and its application to studies of antibody gene sequences (2), allowed us to decipher the mutability of antibodies of different origins in ways not possible in the recent past. The analysis was highly enhanced by germline gene inference technology (15) that defined the germline gene/allele repertoires of the donors under study, thereby minimizing errors originating from inappropriate gene assignment. We now demonstrate how antibodies of different germline gene origins evolve residues and introduce insertions and deletions into CDRs and FRs. This information has implications for our understanding and interpretation of human immune responses.

MATERIALS AND METHODS

Antibody-Encoding Transcriptomes

Antibody-encoding transcripts were isolated from unsorted cells of BM of six subjects diagnosed with allergic rhinitis, examined out of season of most seasonal pollen allergens (14). Transcripts encoding H chain variable (V) domains of different antibody isotypes were individually amplified by PCR, barcoded, and sequenced using Illumina MiSeq technology (14). Sequences are available from the European Nucleotide Archive accession number PRJEB18926. Reads were processed by pRESTO (16) and transcripts encoding each isotype were analyzed by IMGT HighV-QUEST (17) as previously described (14). A summary of the number of sequences at different stages of the analysis pipeline is provided in Supplementary Table EIV in Levin et al. (14).

Germline Gene Repertoire

The germline gene repertoire of the donors have been inferred using IgDiscover (15) using the IgM-encoding transcriptomes of the donors' BM, and has, when possible, been quality controlled by haplotype-based analysis (18, 19). Eleven commonly expressed germline genes (IGHV1-8, IGHV1-18, IGHV2-5, IGHV3-7, IGHV3-11, IGHV3-21, IGHV3-23, IGHV4-39, IGHV4-59, IGHV5-51, and IGHV6-1) (Table 1 in Supplementary Material) mostly encoded by a single or a few highly related alleles, representing six germline gene subgroups, were selected for further analysis (**Table 1**). Full-length (all codons from 1 to 105) sequences of functional germline genes were downloaded from the IMGT database[1] (release 201718-0). Sequence similarity between these genes/alleles was determined after alignment using the ClustalW algorithm (20) as implemented in MacVector 15.5.0 (MacVector, Inc., Apex, NC, USA). Hot-spots for mutation of individual germline genes/alleles were identified by analysis through use of IMGT V-QUEST (21).

Analysis of Diversification of Residues Encoded Proteins

Data defining productive sequences with an origin in investigated germline genes were retrieved following IMGT HighV-QUEST-based analysis (17). Only sequences not showing evidence of

[1]http://www.imgt.org.

TABLE 1 | Examples of germline gene allele repertoire of the six lymphocyte donors as assessed using the IgM-encoding transcriptome.

Germline	Allele composition of each donor[a]					
	Donor 1	Donor 2	Donor 3	Donor 4	Donor 5	Donor 6
IGHV1-2[b]	*02, *p06	*02, *04	*02	*02	*04, *p06	*02, *p06
IGHV1-8	*01	*01	*01	*01	*01	*01
IGHV1-18	*01	*01	*01	*01	*01	*01
IGHV2-5	*02	*01, *02	*02	*01, *02	*02	*01, *02
IGHV3-7[c]	*01, *02	*01	*01 *02	*01, *02	*01, *03	*01
IGHV3-11	*01, *03	*01	*01, *03	*01	*01, *06	*01
IGHV3-21	*01	*01	*01	*01	*01	*01
IGHV3-23	*01	*01	*01	*01	*01	*01
IGHV4-39[d]	*01	*01	*01 *07	*01	*07	*01
IGHV4-59[e]	*01, *08	*01	*01, *08	*01	*01	*01, *08
IGHV5-51	*01	*01	*01	*01	*01	*01
IGHV6-1	*01	*01	*01	*01	*01	*01

Sequence sets not used in the analysis due to an inability to precisely define the germline origin of each protein sequence in at least three donors are showed with a gray background.

[a]Several alleles with identical nucleotide sequence in the assessed parts of the gene may exist, in which case only the lowest allele number is shown.

*[b]IGHV1-2*p06 is a sequence variant (T163C) of IGHV1-2*02 that is not present in the IMGT germline gene database.*

*[c]There is no difference in amino acid sequence in the analyzed part of the sequence of IGHV3-7*01, IGHV3-7*02, and IGHV3-7*03.*

*[d]There is no difference in amino acid sequence in the analyzed part of the sequence of IGHV4-39*01 and IGHV4-39*07.*

*[e]There is no difference in amino acid sequence the analyzed part of the sequence of IGHV4-59*01 and IGHV4-59*08.*

insertions and deletions were scored with respect to presence of substitutions. The frequency of each amino acid was calculated for each position within the range of codons from 27 to 104. Such analysis was performed only for donors that were homozygous for a given allele or heterozygous for alleles that encode identical protein products from the analyzed part of their unmutated sequence. Sequence variability was calculated as the number of amino acids encoded by more than 1% of all reads, divided by the fraction of reads encoding the most common residue. For comparison with real protein structures, examples of structures with an origin in IGHV1-18 (PDB: 3SDY) (22), IGHV1-8 (PDB 3X3G and 3U1S) (23, 24), IGHV2-5 (PDB: 3QRG), IGHV3 subgroup (PDB: 2R56 and 3FZU) (25, 26), IGHV4-39 (PDB: 5C6T) (27), IGHV4-59 (PDB: 3HI1) (28), and IGHV5-51 (PDB: 4BUH) (29) were identified using the IMGT/3Dstructure-DB web interface (30), and coordinates were downloaded from RCSB Protein Data Bank.[2] Structures were visualized using MacPyMOL v1.8.0.6. Sequence numbering and CDR and FR definitions are those defined by the IMGT nomenclature (13).

Insertions and Deletions

Somatic hypermutation not only involves base substitutions but also insertions and deletions in the coding sequence (31, 32). The positions of such productive (in-frame) modifications were scored in each read based on IMGT HighV-QUEST analysis (17).

Evidence of Selection

The 10 most highly expressed rearrangements (based on a defined CDRH3-encoding sequence) of six germline genes, IGHV1-18, IGHV2-5, IGHV3-23, IGHV4-39, IGHV5-51, and IGHV6-1, were investigated. Sequences (codons 27–104) were only retrieved from donors homozygous for a given allele to eliminate the risk of incorrect allele assignment that would contribute to perceived sequence diversification. The sequence with the highest number of counts was chosen so as to minimize the impact of random errors introduced by PCR and/or sequencing artifacts. Sequences showing evidence of insertion or deletion were not used, as the analysis pipeline is incompatible with such modes of antibody diversification. The resulting sequences of IGHV1-18 ($n = 60$ sequences), IGHV2-5 ($n = 30$ sequences), IGHV3-23 ($n = 60$ sequences), IGHV4-39 ($n = 50$ sequences), IGHV5-51 ($n = 60$ sequences), and IGHV6-1 ($n = 60$ sequences) were analyzed for evidence of positive and negative selection using Bayesian Estimation of Antigen-Driven Selection in Immunoglobulin Sequences (BASELINe, version 1.3) using a web-based interface.[3] Focused selection statistics and the Human S5F somatic hypermutation targeting model were used for this assessment (33, 34).

RESULTS

Individual Germline Repertoires

Bone marrow had been obtained from six individuals with different germline gene repertoires, a material that has previously been used for assessment of antibody repertoires in allergic subjects out of season of exposure to most environmental allergens (14). This dataset was now reanalyzed to assess antibody diversification. In any such material, allelic diversity will contribute to antibody diversity and will compromise computational analysis of antibody evolution unless correct allele assignment is made. In particular, as differences between alleles often are small, incorrect allele assignment of hypermutated genes cannot be avoided. Prior analysis of these donors' IgM repertoires, repertoires that carry large numbers of unmutated sequences, can be used to ensure proper downstream analysis of IgG repertoires. Such analysis (19) was performed using IgDiscover (15) to define the lymphocyte donors' IGHV germline gene and allele makeup. We furthermore used a haplotype quality-control approach to define the validity of many of the allele calls (18, 19). This approach also allowed us to validate novel alleles not present in the IMGT reference directory used for gene assignments (19). By using donors with defined germlines, we minimized the risk of introducing artifacts in our analysis of the mutational paths taken by antibodies of different germline gene origins. Importantly, this approach identified allele IGHV1-2*p06 (IGHV1-2*02 T163C) in three of six individuals (19) (**Table 1**), an allele that is not identified by standard IMGT HighV-QUEST or V-QUEST analysis. A failure to identify this allele would incorrectly have enhanced the perceived substitution frequency in one position of this gene by the approach taken in this study. As a consequence of the substantial, but difficult to detect, allelic diversity of IGHV1-2, it was not included in this study. We also made sure that the investigated germline genes were not extensively similar to alleles of other germline genes in the donors' repertoires, as such similarity may, following hypermutation, incorrectly relate products of other genes to the genes under investigation. IGHV3-23 and IGHV3-23D are in this context treated as one gene as they are identical in sequence. Among the other genes, only one investigated allele of IGHV4-59 had a highly similar allele defined by IMGT (>98% nucleotide identity) assigned to another gene location (Figure S1 in Supplementary Material), but this other allele (IGHV4-4*08) was not present in the repertoires of the investigated individuals.

Overall, several germline genes were highly diverse (different alleles used by different donors or presence of different alleles in a given individual), but others were not, some of which were used for this study (**Table 1**). In all, our analysis focused on a set of commonly used "core" genes (35) utilized in rearrangements obtained from individuals conceived to be homozygous for a given allele or heterozygous for alleles expected to encode identical protein sequences in their unmutated form. As the analysis focused on sequences (germline-encoded protein sequences are shown in **Figure 1**) from CDRH1 up to the end of FR3, alleles that encode identical protein products in this part of the H chain variable domain could be included in the study. Thus, for the purpose of analysis of selection associated with hypermutation, only sequences derived from individuals homozygous for a given gene sequence encoding CDRH1 to FR3 (codons 27–104) were used.

The IgG Population Encoded by BM Is Highly Somatically Evolved

We chose to study mutations in antibodies of the IgG isotype produced in BM as this is a major site of antibody production.

[2] http://www.rcsb.org/pdb/.

[3] http://selection.med.yale.edu/baseline/.

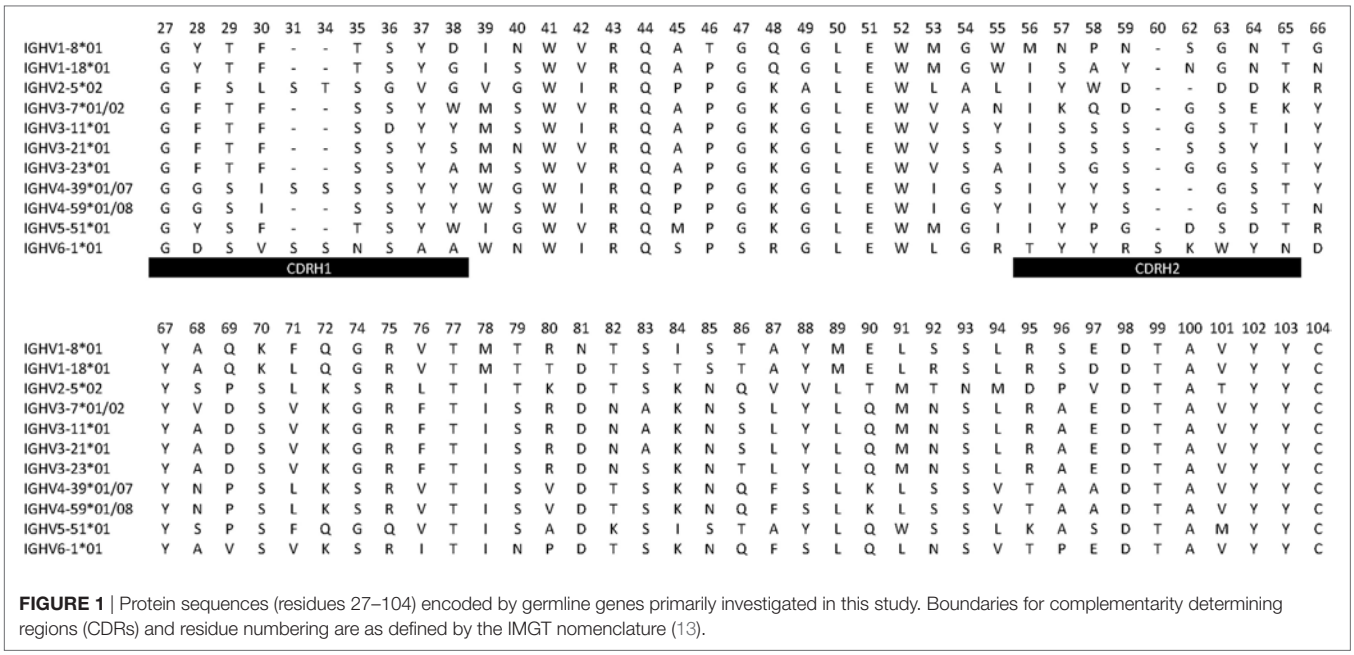


FIGURE 1 | Protein sequences (residues 27–104) encoded by germline genes primarily investigated in this study. Boundaries for complementarity determining regions (CDRs) and residue numbering are as defined by the IMGT nomenclature (13).

By focusing on the entire transcriptome and not sequences collapsed to individual clones, the analysis also focused on features of highly produced products. Samples (10 ml) obtained from BM, as analyzed in this study, are reproducible representations of antibodies produced at this site (14). This population of cells largely contains mutated transcripts as evidenced by the fact that only 1.4% (range 0.6–2.5%) of them showed a level of mutation below 2% at the nucleotide level (the corresponding IgM-encoding transcriptome contained 45% (range 35–55%) of sequences displaying a degree of mutation below 2%) (14), as determined using IMGT HighV-QUEST.

Germline Genes Differ Extensively in the Extent of Targeting by Substitutions

The degree of substitution in the part of VH encoding CDR1, FR2, CDR2, and FR3 with an origin in 11 well-defined germline genes from six human IGHV germline gene subgroups was analyzed. The average frequency of substitution of a residue from residues 27 to 104 ranged from 12.5% (IGHV2-5) to 18.4% (IGHV4-39). Different substitution patterns were seen with different residues being targeted by diversification depending on germline gene origin (**Figure 2**). Substitutions were, as expected, often located to some residues within CDRs, but they also occurred frequently in numerous residues in FRs. It was not uncommon for FR residues encoded by a particular germline gene to be substituted in >25% of all transcripts. Diversification of FRs is thus an important aspect of antibodies produced from cells in BM that have undergone a somatic hypermutation processes.

The degree of targeting of CDR1 and CDR2 differed substantially between different germline genes both in terms of the frequency of substitution and the degree of variability introduced (**Figure 2**; Figure S2 in Supplementary Material). For instance, substitutions were most frequently incorporated into CDRH2 of VH with an origin in IGHV3-11, while substitutions were incorporated more frequently into CDRH1 of VH with an origin in IGHV5-51 (Figure S2 in Supplementary Material). The precise codons targeted by successful diversification differed between germline genes. For instance, while residue 29 (mostly S or T in germline sequences) was targeted by substantial diversification in proteins derived from some genes (e.g., investigated genes of subgroups 1, 3, and 5) it was not targeted in genes of other germline origins (IGHV2-5, IGHV4-39, IGHV4-59, and IGHV6-1) (**Figure 3**) (additional examples are provided in **Figure 4**). In the case of S29-encoding germline genes, the extensive targeting of mutations to this residue in IGHV5-51 was associated with the presence of a mutational hotspot in the codon of this gene (Figure S3 in Supplementary Material). Similarly, residues in immediate proximity to CDR in the linear sequence were frequently diversified. The extent of diversification of some of these residues differed substantially depending on germline gene origin. For instance, while W55 of IGHV1-8 and IGHV1-18 and R55 of IGHV6-1 were rarely (<10%) substituted, A55 of IGHV3-23 was substituted in products encoded by 76% of the transcripts with an origin in this germline gene (**Figure 5**). Among these genes, IGHV6-1 carries an AA dinucleotide hotspot motif while IGHV3-23 carries a TA motif and a AGCT motif that may specifically target codon 55 with mutations.

In summary, VH domains encoded by transcriptomes found at a major site of antibody production, the BM, differ in the paths through which they evolve residues within or in the immediate vicinity to CDRs.

Evolution of Residues Belonging to the Cores of Variable Domains

Residues that make up the core regions of antibodies are important for protein stability, and may thus conceivably be less targeted by

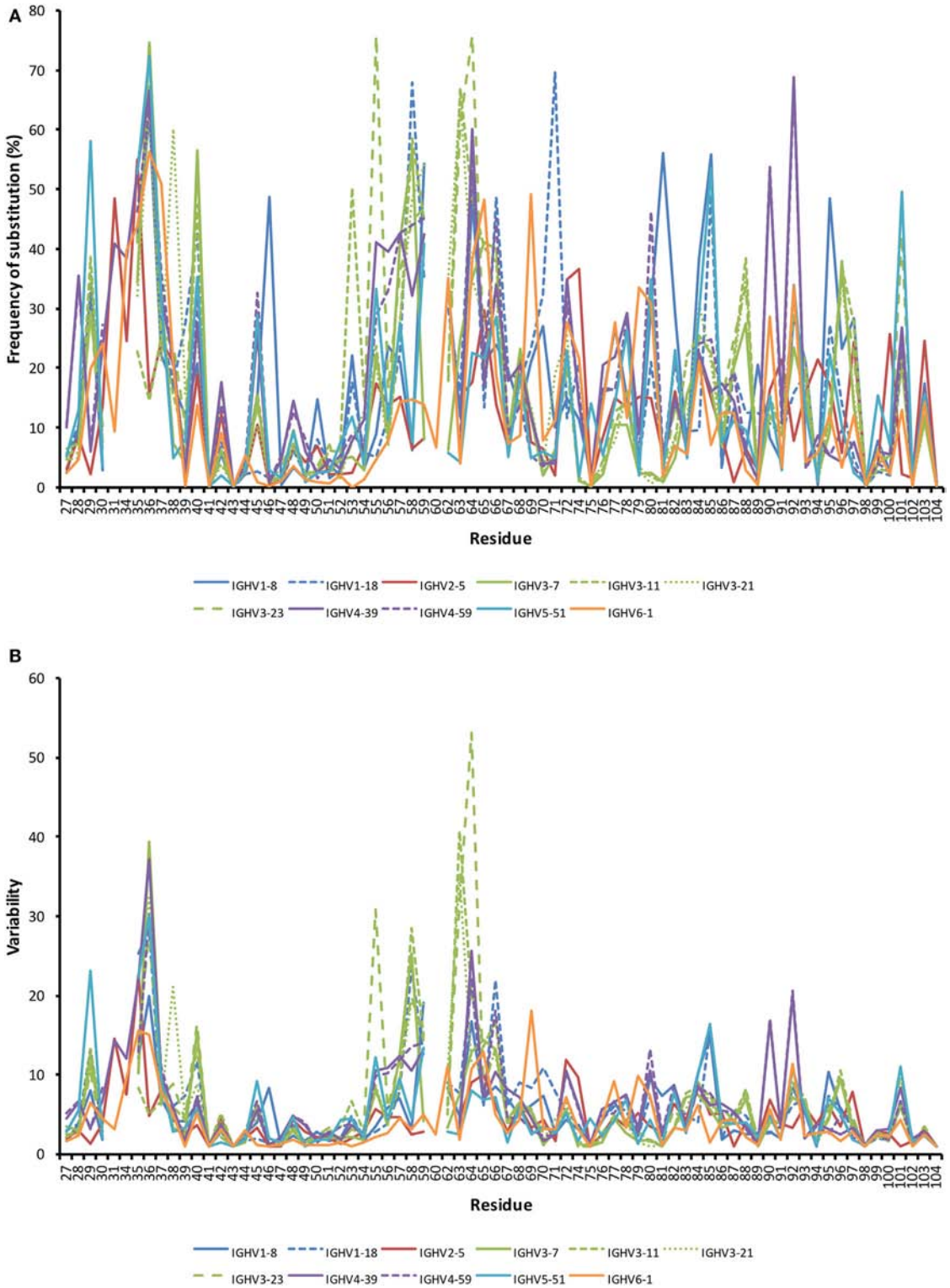


FIGURE 2 | Frequency of substitutions **(A)** and degree of variability **(B)** of residues encoded by transcripts in bone marrow (BM) with an origin in the 11 investigated germline genes. Residues 27–38 code for complementarity determining region (CDR) 1, while residues 56–65 code for CDR2. Variability was calculated as (the number of amino acids encoded by more than 1% of all reads)/(fraction of reads encoding the most common residue). Only sequences showing no evidence of insertions and deletions were included in the analysis. Numerous residues differed substantially between germline genes in the degree of targeting by substitution. Note the substantial degree of substitution also in some residues of framework region (FR). Frequency of substitution and variability for individual genes are shown in Figure S2 in Supplementary Material.

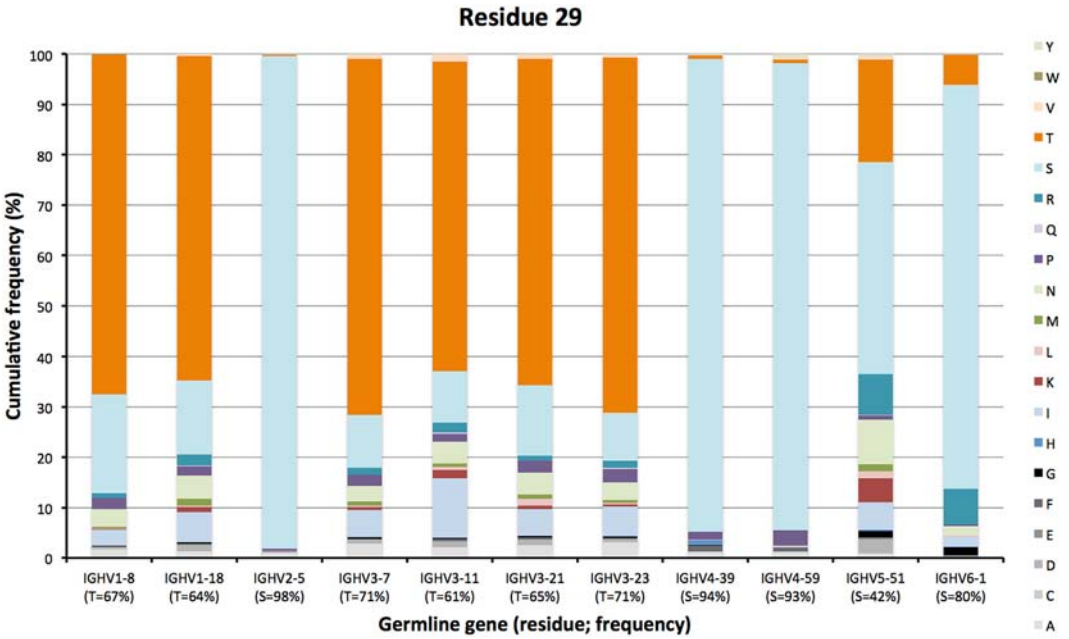


FIGURE 3 | Diversity of residues encoded by transcripts of different germline gene origins in position 29 of VH of IgG. The germline-encoded residues and their frequencies in IgG-encoding transcripts are shown at the bottom of the graph. Note the extensive diversification of S29 in IgG-encoding transcripts derived from IGHV5-51 but not from other genes, in particular IGHV2-5, IGHV4-39, and IGHV4-59.

mutation. Indeed, five residues within the region CDRH1-FR3 of VH were substituted in <2.5% of all human IgG transcripts independently of their germline gene origin. These included W41 and C104 in the domain's central core, residues R43 and D98 in the domain's charge cluster, and residue Y102 in the lower core of the domain (**Figure 6**). Nevertheless, as described below, we observed several other residues belonging to the core regions that are diversified more extensively during somatic antibody evolution.

The lower core is shielded from the upper core by the highly conserved central core (including, e.g., W41 and C104) and a more direct influence by substitutions in this core on the binding site may be limited. Instead, lower core residues may affect the biophysical properties of the domain (36). Several of these residues (of which residues 53, 54, 71, 76, 89, 91, 94, 100, and 102 were assessed in this study) were essentially untouched by somatic diversification, while others, depending on germline gene origin, were diversified. In particular, residue 53 of IGHV3-11 (but not IGHV3-7, IGHV3-21, and IGHV3-23) and residue 71 of IGHV1-18, were prone to diversification (**Figures 6** and **7**). IGHV1-18 encodes a L at position 71, while other genes of the IGHV1 subgroup encode F in this position. Mutation of this codon in IGHV1-18 introduced F in this position at a high frequency (55%). Mutation of residue 53 incorporated conserved hydrophobic substitutions in place of the germline-encoded residue. In the case of IGHV3-11, V was mainly substituted by L or I. Similarly, mutation of residue 76 largely introduced conservative, hydrophobic substitutions (**Figure 7**). In summary, some residues of the VH domain's lower core are targets for conservative hypermutation.

A cluster of charged residues is situated close to the lower core of VH (36). It involves residues at positions 43, 51, 75, 95, 97, and 98 (37). Of these, 0–2 residues, mostly residue 95 and 97, were targeted by substitutions at frequencies above 10% (**Figure 6**). Of note, 48% of all sequences with an origin in IGHV1-8 were targeted by substitution at residue R95 (a codon not associated with a mutational hotspot), while only 9% of sequences with an origin in IGHV3-7 was diversified in this position. Limited diversity (T or K) dominated the diversity introduced at this position. Similarly, substitution at position 97 (a codon not associated with a mutational hotspot in any of the investigated germline genes) in IGHV3 gene subgroup members was dominated by a conservative E → D mutation, while substitution of V97 in IGHV2-5, an unusual, hydrophobic side chain in this cluster, introduced a range of modifications although mainly to A (**Figure 8**). Altogether, there is room for diversification in the charge cluster in a germline-directed manner, modifications that for instance may affect the biophysical properties of the domain.

Substitutions of residues that belong to the VH/VL interface may affect binding site architecture. We investigated the tendency for substitution in five residues (40, 42, 50, 52, and 103), the surfaces (**Figure 6**) of which are substantially buried by formation of the VH/VL dimer (11). Among these, residues 50 and 52 were rarely mutated [each below 10% of transcripts, except in the case of residue 50 of IGHV1-18 (15%)] while, in particular residue 40 (G, N, or S) but also residues 42 (I or V) and 103 (Y) were frequently substituted (**Figure 6**), although mostly in a restricted manner (**Figure 5**; Figure S4 in Supplementary Material). Residue 40 of some germline gene origins showed substantial levels of substitution (even above 50%)

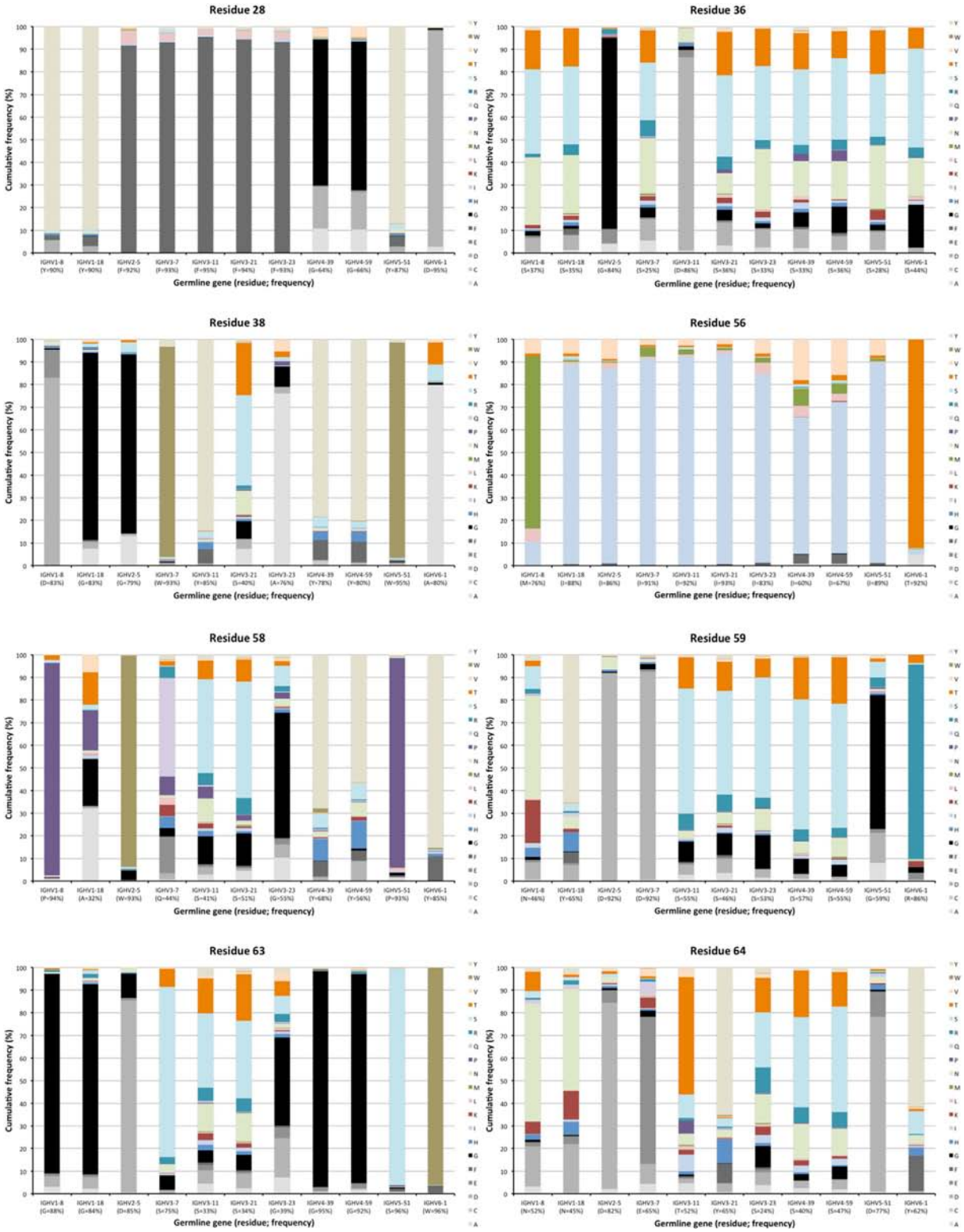


FIGURE 4 | Examples of differences in diversification of residues in complementarity determining region (CDR) encoded by transcripts of different germline gene origins. Residues shown include positions 28, 36, and 38 of CDRH1, and positions 56, 58, 59, 63, and 64 of CDRH2 of IgG. The germline-encoded residues and their frequencies in IgG-encoding transcripts are shown at the bottom of the graph. Substitutions are introduced in 5–36% (residue 28), 14–75% (residue 36), 5–60% (residue 38), 7–40% (residue 56), 6–68% (residue 58), 8–54% (residue 59), 4–67% (residue 63), and 18–76% (residue 64) of the transcripts depending on their different germline gene origins.

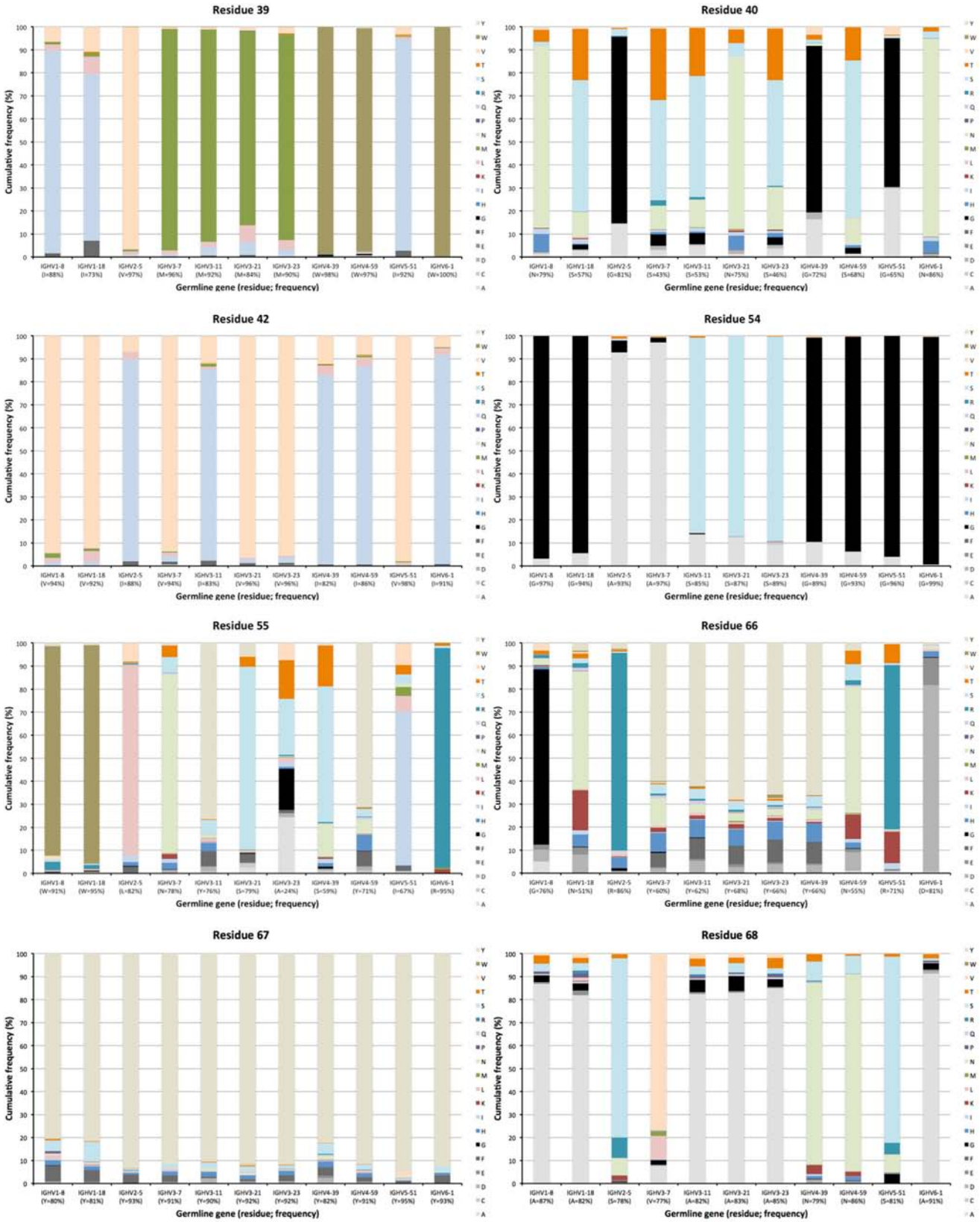


FIGURE 5 | Continued

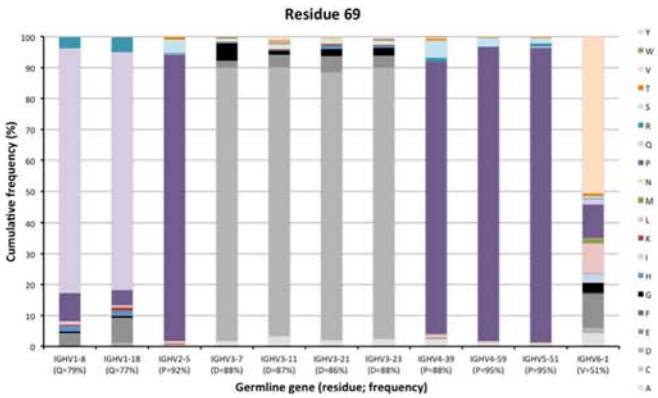


FIGURE 5 | Extent of diversification of residues close to complementarity determining region (CDR) in the linear protein sequence as encoded by transcripts of different germline gene origins. Residues shown include residues 39, 40, 42 immediately after CDRH1, and residues 54–55 immediately before and residues 66–69 immediately after CDRH2 of IgG. Residue 41, a conserved tryptophane belonging to the domain's core, is not diversified in products encoded by any germline gene. The germline-encoded residues and their frequencies in IgG-encoding transcripts are shown at the bottom of the graph.

but the diversification was largely limited in scope (such as S → N or T, or G → A) (**Figure 5**). In summary, there is room for diversification of some residues often buried in the VH-VL interface, modifications that may affect the binding site or the stability of the VH-VL pair.

The upper core of antibody H chain variable domains (37) (of which residues 28, 30, 39, 78, 80, and 87 have been assessed here) is located just beneath the paratope and diversification of its residues may have profound effects on the binding site (38). Several of the residues that constitute the upper core (36) are by definition part of the sequences that comprise CDR, although their side chains are not necessarily extensively exposed on the surface of the domain. The residues play different roles, depending on their biophysical nature (37). Several residues in this core of VH, depending on its germline gene origin, are prone to accept mutations. Many germline genes encode large aromatic residues at position 28 that were rarely mutated (mostly <10%) (**Figures 2** and **4**). However, germline genes IGHV4-39 and IGHV4-59 encode a G in this position, a residue that was frequently (approximately 35%) substituted (mostly to A, D, and V) in products of IgG-encoding transcripts. Similarly, genes that encode an aromatic side chain in position 30 of VH rarely substituted it (≤10%) while genes derived from germline genes like IGHV4-39, IGHV4-59, and IGHV6-1, which encode a hydrophobic amino acid in position 30, were more prone to substitute it, mostly for another hydrophobic residue (**Figure 9**). Residue 80 in the upper core is important for the positioning and conformation of CDR2 (38). IGHV3 germline genes incorporate R at this position, a side chain that was only very rarely (≤2% of reads) substituted by other residues. In contrast, R80 in antibody-encoding genes derived from IGHV1-8 underwent substitution at a high frequency. This ability for diversification is not associated with the presence of a mutational hotspot (*WA*/*TW* or *RGYW*/*WRCY*) in this codon in IGHV1-8 (Figure S3 in Supplementary Material). Other germline genes encode other residues in position 80 and these may also be substituted to a substantial extent (**Figure 9**). Altogether, there is tolerance for diversification of many residues of the upper core in a germline-origin-dependent manner.

Antibody Evolution Provides Diversity Beyond CDR and Domain Core Structures

Hypermutation may extend to surface residues beyond CDRs, even to residues that are not located in immediate proximity to those defined to make up the CDRs. Numerous residues, in particular in FR3 carried such diversity (**Figures 2** and **10**). Sequences around residue 85 have been considered as a fourth CDR (39). This residue frequently carried diversity, a feature particularly evident in transcripts with an origin in IGVH1-8, IGHV1-18, and IGHV5-51 germline genes, in which case about 50% of the transcripts carried substitutions. This side chain is localized immediately below CDR1 in the folded domain (**Figure 10**) and it is highly conceivable that mutations may affect binding affinity and/or specificity. Certainly, antibodies derived from some germline genes show extensive evolvability in this part of the domain.

Other residues that are located at a substantial distance from CDR, were also frequently mutated. For instance, residues 90 and 92 in FR3 showed evidence of extensive diversification in transcripts derived from some germline genes, in particular those of IGHV4-39 and IGHV4-59 (**Figure 10**). Similarly, residue 101 in many VH carried a substantial level of diversification (**Figure 10**). Only in the case of IGHV3-23 was this propensity for substitution in position 101 associated with the presence of a mutational hotspot. The side chain of residue 101 is also exposed on the domain's surface near the interface with VL far away from the binding site. Some residues, although not targeted to diversification in general, may be targeted extensively in antibodies derived from some germline genes. For instance, in similarity to residue 71 of IGHV1-18 (described above; **Figure 7**), residues 46 and 81 of IGHV1-8, and to some extent residue 75 of IGHV5-51 were frequently mutated

(**Figure 11**). The corresponding codons of the germline genes encode T, N, and Q, respectively, while most other germline genes encode P, D, and R, respectively (**Figure 1**). The evolution of in particular IGHV1-8-derived H chain variable domain often introduced precisely these residues into the products.

In summary, numerous FR residues, the side chains of which are found on the surface of VH, are diversified in a germline gene-defined manner through antibodies' evolution processes *in vivo*.

Insertions and Deletions

Antibody sequences can evolve not only by hypermutation but also by insertion and deletion of entire codons (31, 32). The present dataset allows for analysis of such processes in hypermutated antibody sequences of different germline gene origins. We identified the location of such modifications (as annotated by IMGT HighV-QUEST) in transcripts derived from a number of germline genes (**Figure 12**; Figure S5 in Supplementary Material). Insertions were on average longer than deletions

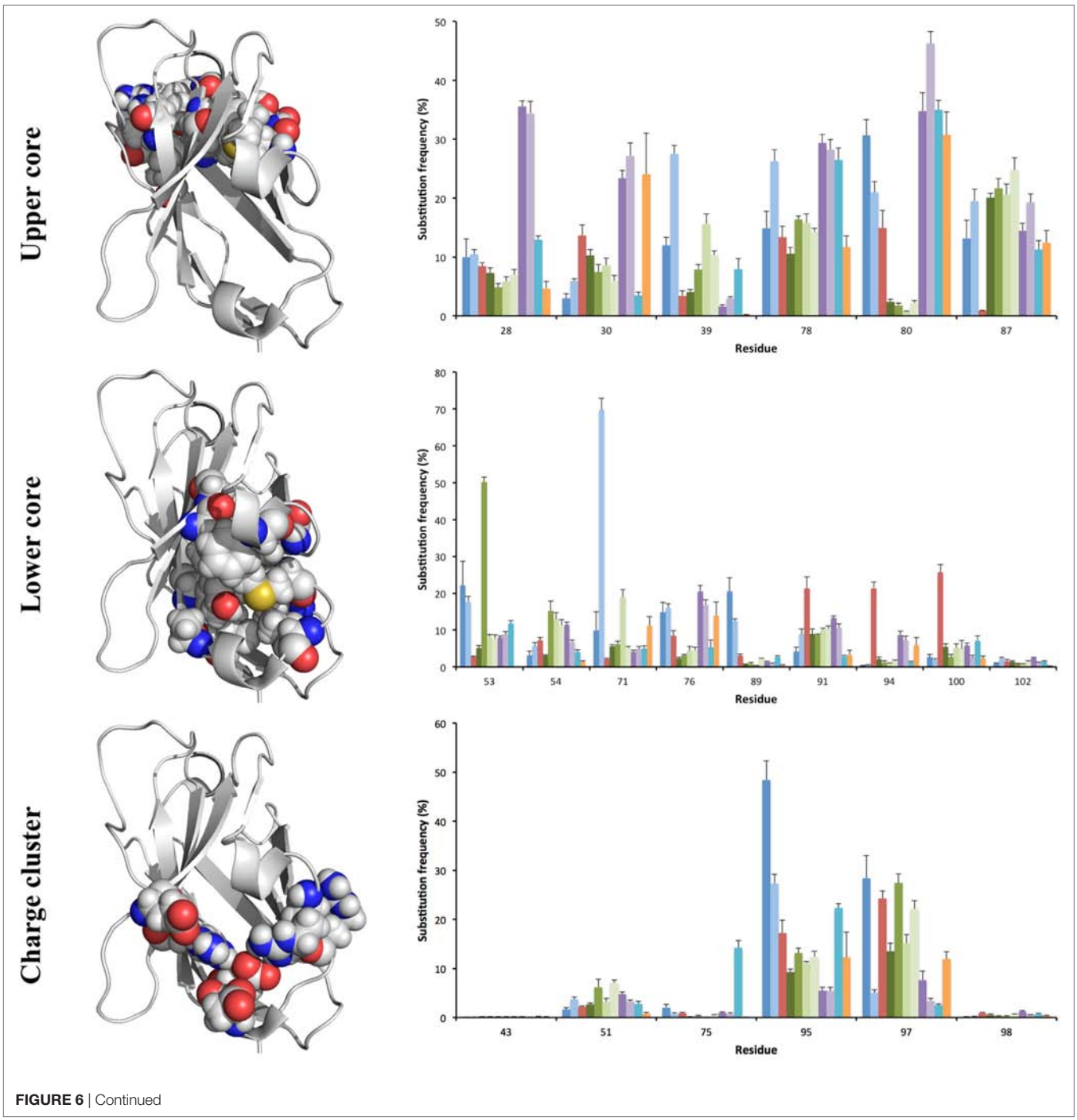


FIGURE 6 | Continued

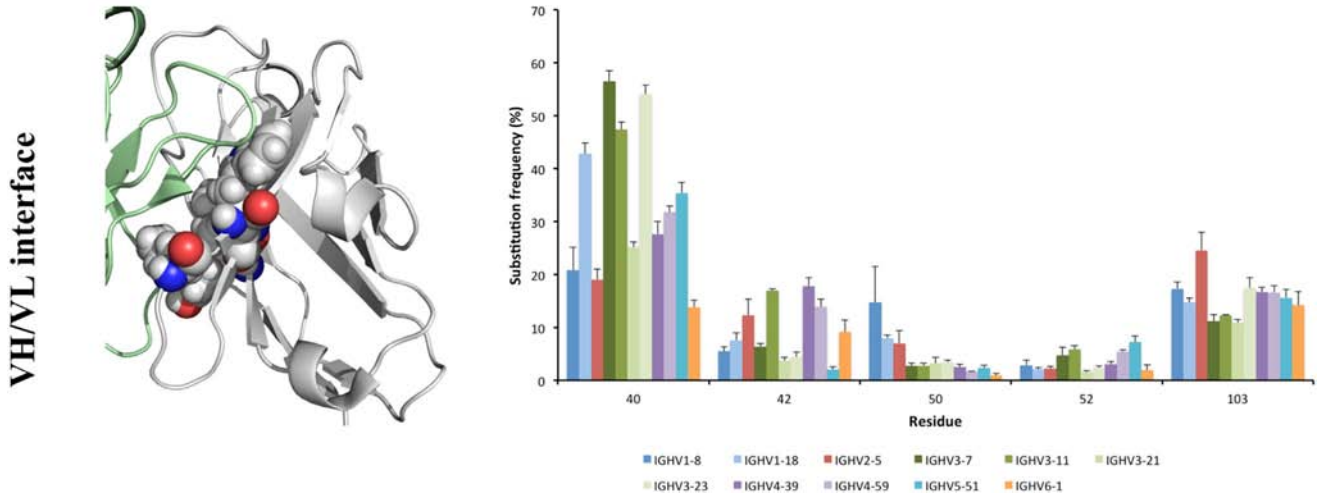


FIGURE 6 | Substitution frequency of the charge cluster, lower and upper cores, and major residues that participate in VH/VL interface of the antibody H chain. Error bars indicate SEM between repertoires of different donors. Residues belonging to these structures include those exemplified in the structure of the VH of a human antibody fragment with an origin in germline gene IGHV1-18 (PDB: 3SDY). The VH domain is shown with a gray backbone. Part of the VL domain in the illustration of the domain interface is shown in green. Achieved diversity of individual residues in the sequence populations are outlined in **Figures 7**–**9**, and Figure S4 in Supplementary Material.

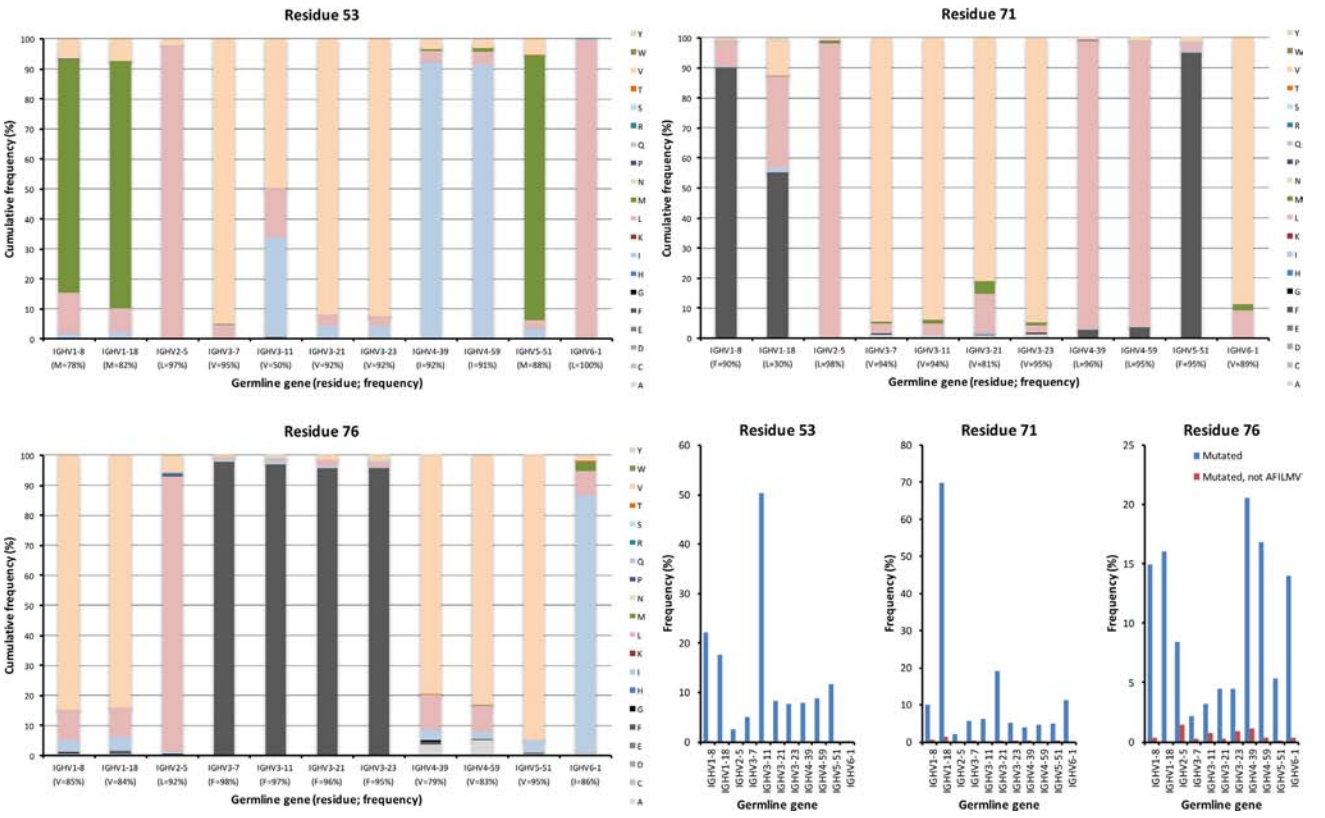


FIGURE 7 | Substitutions of lower core residues 53, 71, and 76 as introduced during somatic hypermutation processes of IgG-encoding genes derived from different germline genes. The diversification of these mostly hydrophobic residues largely results in introduction of other hydrophobic residues (panels to the bottom right).

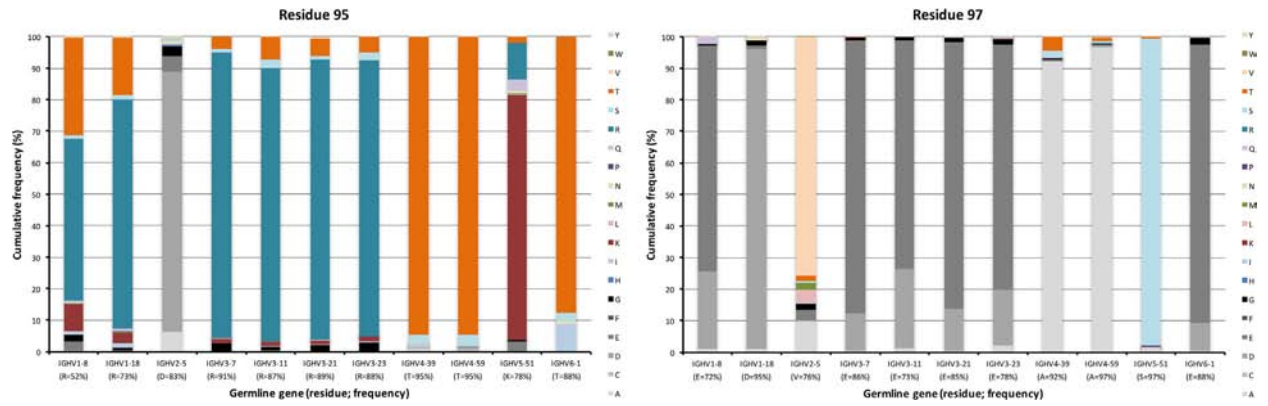


FIGURE 8 | Substitutions of charge cluster residues 95 and 97 as introduced during somatic hypermutation processes of IgG-encoding genes derived from different germline genes.

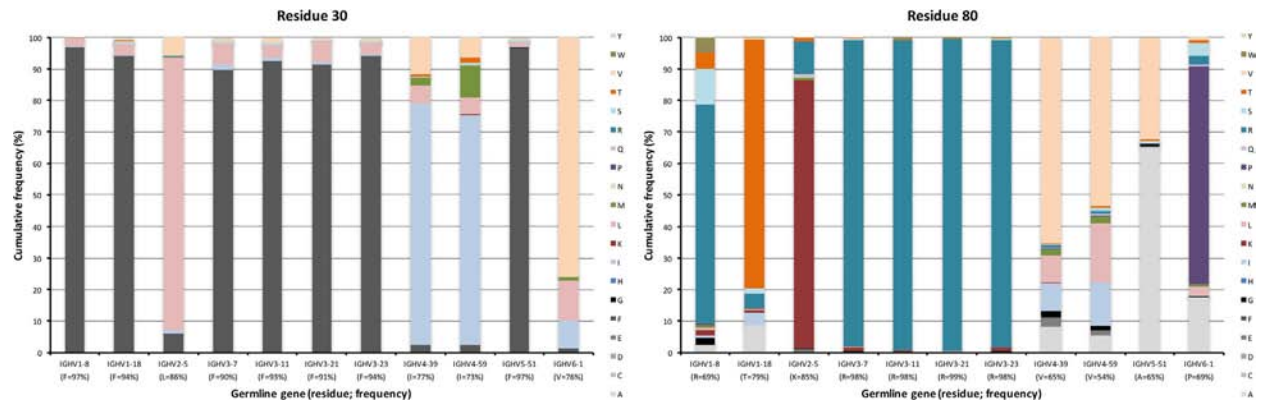


FIGURE 9 | Substitutions of upper core residues 30 and 80 as introduced during somatic hypermutation processes of IgG-encoding genes derived from different germline genes. Substitution of residue 28 is shown in **Figure 4**.

(6.7 and 4.6 bases, respectively; $p = 0.028$ using the Wilcoxon signed rank test) in in-frame transcripts with an origin in the 11 germline genes (irrespective of allelic origin) that are the focus of this study. Most such modifications occurred within/close to CDRs. Several genes (like IGHV3-23) were primarily targeted by insertions and deletions in CDRH2 while others (like IGHV2-5 and IGHV4-39) were targeted also by such modifications in CDRH1. Some genes [like IGHV3-7, IGHV5-51 (**Figure 12**), and IGHV1-69 (Figure S5 in Supplementary Material)] also extensively introduced insertions and deletions in CDRH4, i.e., in the loop situated in close proximity to other, conventional CDRs. In summary, it appears that rearranged sequences derived from different germline genes target insertions and deletions to different parts of their sequence.

Somatic Hypermutation and Evidence of Selection

In vitro evolution through hypermutation may contain objective evidence of selection as events likely to contribute to improved binding are favored over those with no or negative influence on antigen recognition. Such productive events are considered focused to the CDR and they may be detectable using computational approaches (33, 34, 40–42), although this possibility has also been questioned (43, 44). We investigated the evidence for such selection in only the most frequent, independent sequences of each germline in each donor to minimize the effect of random PCR and sequencing errors. Only donors that expressed a single allele of a gene were included in the analysis to minimize the risk of errors introduced by incorrect allele assignment. Such analysis demonstrated that there, despite the high degree of substitutions in FR of VH encoded by BM-derived transcripts, was a profound negative selection for mutations in FR. Although there was less selection against substitution of residues in CDR, it was not possible to identify positive selection in VH domains with an origin in any of the investigated germline genes (**Figure 13**).

DISCUSSION

Analysis of the information content of human antibody diversification holds promise for understanding antibody evolution

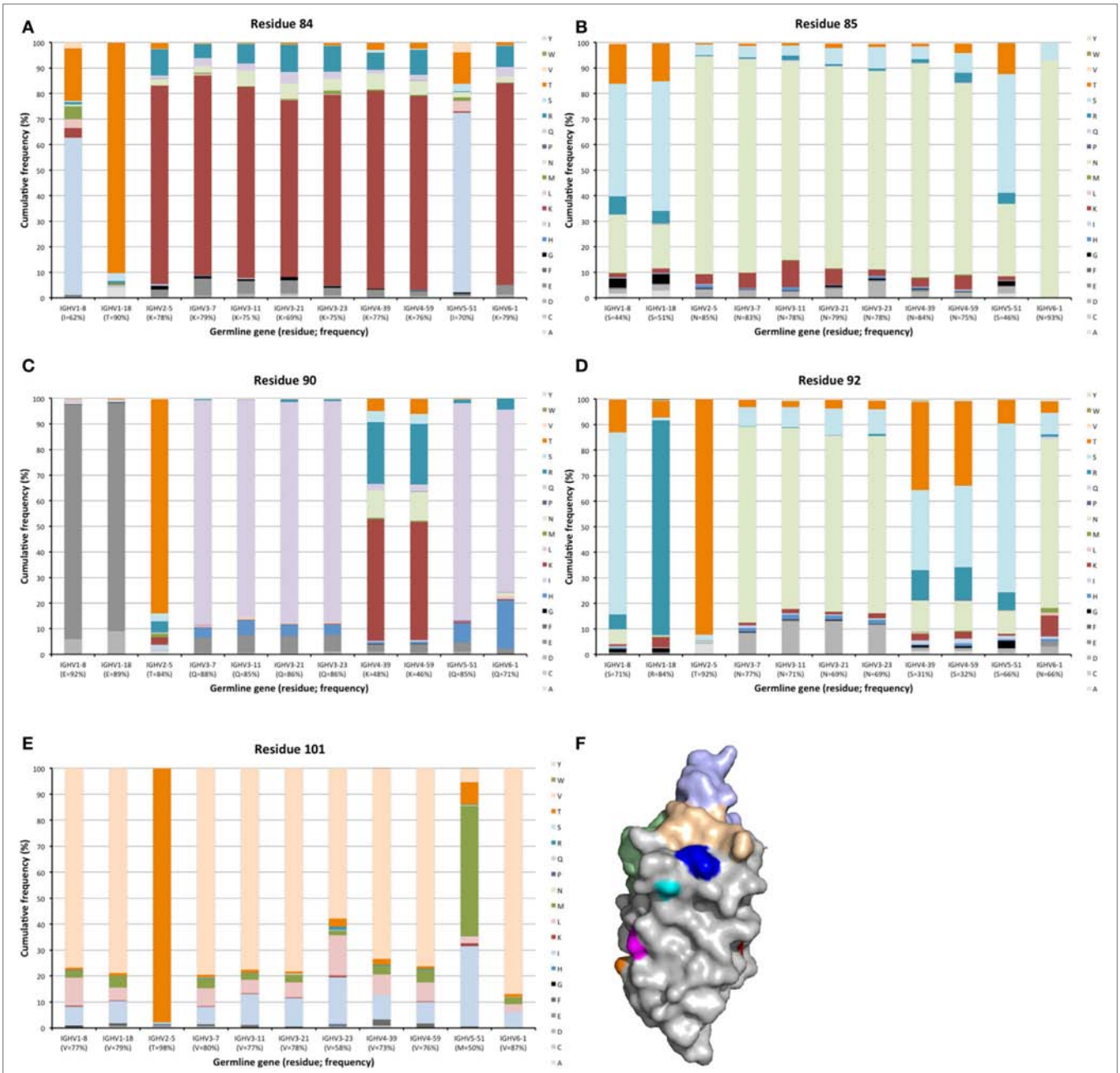


FIGURE 10 | Substitutions of residues 84 **(A)**, 85 **(B)**, 90 **(C)**, 92 **(D)**, and 101 **(E)** in FR3 as introduced during somatic hypermutation processes of IgG-encoding genes derived from 11 different germline genes. Structure **(F)** of an IGHV5-51-derived scFv (PDB: 4BUH) indicating the side chain atoms of residues 84 (cyan), 85 (dark blue) in close proximity to CDR1, and of residues 90 (magenta), 92 (orange), and 101 (red). CDR are shown in brown (CDR1), green (CDR2), and light blue (CDR3).

and affinity maturation as it occurs *in vivo*, and evolution processes as used *in vitro* to develop high affinity, highly biophysically stable antibodies. The advent of next-generation sequencing and the availability of much larger collections of antibody sequences allows for a very in-depth analysis of antibody diversity. We envisaged that such analysis would also define constraints on human repertoire development as a function of antibody germline gene origin and thus enhance the way we in the future analyze events involving somatic diversification. Such studies and concomitant studies of antigen-antibody complex structure have been used to understand in detail how human humoral immune repertoires develop, or fail to develop efficiently. Large-scale studies, as recently reviewed, have addressed the evolution of antibodies against highly functional epitopes on viral antigens like the envelope protein of HIV-1 with the intention to enable design of immunogens that more

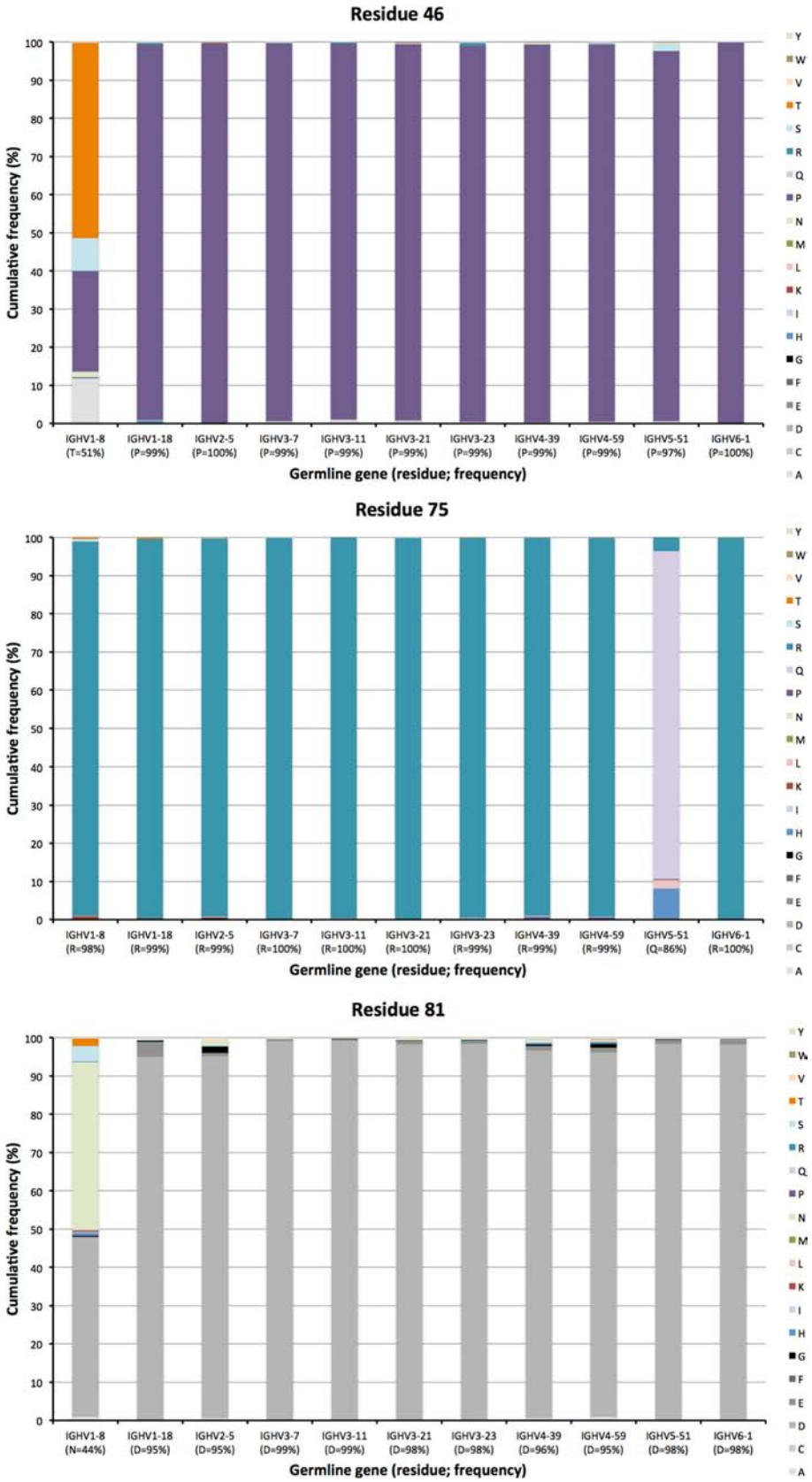


FIGURE 11 | Substitutions of residues 46, 75, and 81 as introduced during somatic hypermutation processes of IgG-encoding genes derived from 11 different germline genes.

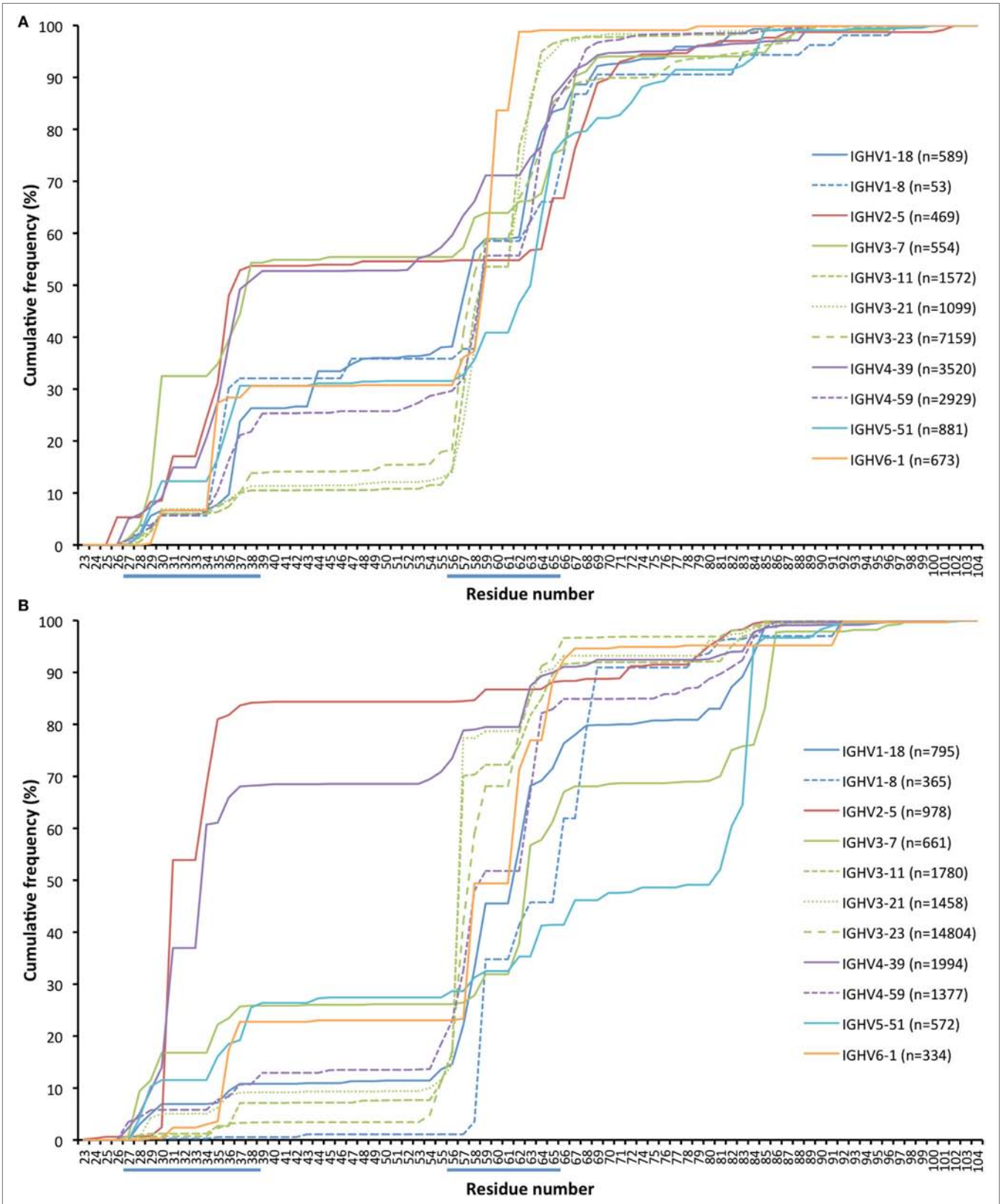


FIGURE 12 | Cumulative frequency of in-frame codon insertion **(A)** and deletion **(B)** (as calculated by IMGT HighV-QUEST) in rearranged genes derived from a set of germline genes (irrespective of allele origin). The occurrence of such events in genes derived from additional germline genes, not representing the core genes investigated in this study, are shown in Figure S5 in Supplementary Material.

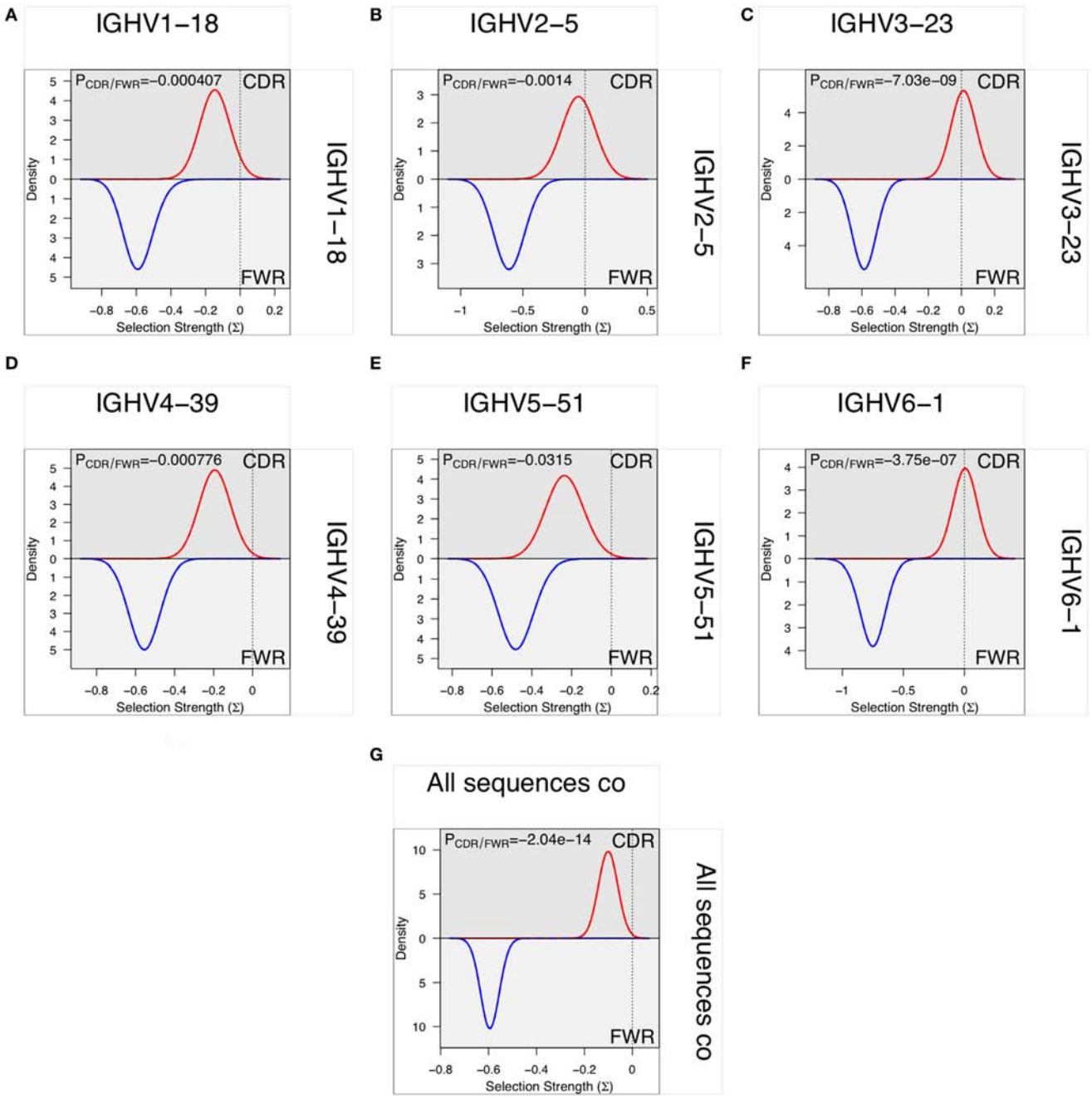


FIGURE 13 | Analysis, using BASELINe (33, 34), of selection pressure on the diversification of residues of the framework region (FR) and complementarity determining region (CDR) with an origin in six germline genes [IGHV1-18 **(A)**, IGHV2-5 **(B)**, IGHV3-23 **(C)**, IGHV4-39 **(D)**, IGHV5-51 **(E)**, and IGHV6-1 **(F)**]. There is a profoundly more negative selection of diversification of residues of FR2 and FR3 than of CDR1 and CDR2 but there is no evidence for positive selection in the substitution pattern of CDR of sequences derived from any of the individual germline genes. Indeed, overall there is evidence for negative selection for substitution also in the CDR **(G)**.

efficiently induce protective immunity in vaccinated subjects (5, 6, 45). In the present study, the aim was to deconvolute antibody diversification paths, not from a global perspective but with a focus on products of individual germline genes, to enable enhanced quality of future analysis of antibody evolution. To do so, we employed sequences that encode IgG in BM (14), a major site of long-term, sustained antibody production. The sequences were derived from subjects diagnosed with seasonal allergic rhinitis but they were obtained out-of season of most seasonal allergens. We consequently consider them not to be biased by an on-going allergic immune response. In any event, we do not consider that major features of evolution of the IgG

response of such subjects would be dramatically different from that of non-allergic, immunocompetent, subjects.

The present study focuses on the diversity found from CDR1 until the end of FR3 of the human H chain V domain. By focusing our attention on diversification of well-defined germline genes and alleles, our analysis is minimally confounded by differences in germline gene allelic makeup between individuals or between haplotypes of an individual. In preparation of the present study, we consequently inferred the germline VH repertoires of the lymphocyte donors (19) and analyzed genes of donors with well-established germline gene allele composition. The one exception to this rule is IGHV3-23 and its, in their mature peptide-coding sequences, exactly duplicated sequence IGHV3-23D, sequences that were treated as one entity in this study. Furthermore, as some germline genes are highly similar, confounding outcomes may occur as a consequence of mutational processes rendering sequences derived from one germline gene more similar to the nucleotide sequence of other germline genes. Such highly similar germline genes (>98% nucleotide identity) were, to avoid erroneous interpretation, not investigated in the present study.

Antibodies of different germline gene origins differed substantially in terms of diversified residues, in agreement with recent findings (46). There is thus a solid basis for defined, preferred germline-centric paths of antibody evolution. Certainly, driving forces that promote higher substitution frequencies may relate to affinity maturation, stability enhancement etc. It is likely that part of the observed differences relate to the presence or absence of sequences acting as hot-spots for the mutational machinery (47). However, structural analysis have previously demonstrated that amino acids that are in mutational hot-spots are not more likely to actually undergo substitution during somatic hypermutation, suggesting that such hot-spots "are not a major driving force in determining which residues are mutated" (48). In any event, IGHV2-5, IGHV4-39, IGHV4-59, IGHV5-51, and IGHV6-1 all encode S29, a residue capable of different interactions in different antibodies (**Figures 14A–D**). It is, however, only in sequences derived from IGHV5-51 that this residue is targeted by extensive substitutions. This gene is also the only one among the five that carries a mutational hot-spot motif affecting this codon. If this is a selected, germline-encoded feature preventing extensive, non-functional substitution of products derived from the other four genes, or not, is currently not known. However, if this is not the case, there is a capacity for antibody evolution in antibodies derived from some genes that is not efficiently explored by the human immune system.

Beyond hot-spot involvement in the orchestration of substitution, it is also likely that structural consideration in many cases guide the ability of antibodies of different germline gene origins to tolerate or even prefer substitutions. Certainly, loops belonging to different canonical classes (49, 50), positioning side chains with identical residue numbers in entirely differently orientations and environments, may affect their ability to structurally accept substitutions. Importantly, some germline genes encode unusual residues in some positions, the side chain of which may be suboptimal for its environment. We envisage that such unusual residue may be more commonly mutated even if they reside in FR, often attaining the more common residue following substitution, as illustrated by residues 53 of IGHV3-11 and residue 71 or IGHV1-18. Furthermore, residue 80, a residue in the upper core of the variable domain, with particular importance for the conformation of CDRH2 (38) provide interesting insight into germline-directed paths for antibody evolution. The common diversification of R80 in products with an origin in IGHV1-8 (in contrast to the lack of diversification of R80 in products derived from IGHV3 subgroup genes) (**Figures 6** and **9**) is not associated with the presence of a mutational hotspot in IGHV1-8 affecting codon 80 (Figure S3 in Supplementary Material). Possibly R80 is less important for maintenance of the integrity of the domain or the general architecture of the binding site in antibodies derived from this germline gene. Interestingly, reorientation of R80 has been demonstrated in one antibody (24) with a likely origin in IGHV1-8. This was also associated with a reorientation of CDRH2 as determined by X-ray crystallography (**Figures 14E-H**). We hypothesize that substitution of R80 in products originating from genes like IGHV1-8 may be part of an efficient route to evolve antibody functionality while still being tolerated in terms of structural stability. We hypothesize that some germline genes may even have an inherent need, or, if one so-prefer, capacity, for evolution that is not present in other germline genes. In all, the reason for such high substitution frequency *in vivo* may differ between antibodies of different germline gene origins. Future studies will have to address the difficulties encountered by, or alternatively extended opportunity of, the cells producing antibodies derived from these germline genes to gain an advantage in the race for selection through the affinity maturation processes occurring in germinal centers.

Residues beyond CDR may interact with antigen or contribute indirectly to the architecture of the binding site. There are, however, many definitions of CDRs (8–12) apart from the one used in this study (13), definitions that accommodate different viewpoints of what constitutes an antigen binding site. Indeed, substantial diversity is apparently tolerated, or even selected for, in residues in immediate proximity to CDRs as defined by IMGT (**Figures 1** and **5**). For instance, codon 55 of different germline genes encodes very different side chains. Is the observed difference in mutability solely a result of the presence or absence of a mutational hotspot, or is it also directed by a difference in importance of this residue for establishment of a core antigen-interacting surface (51) in products encoded by the different germline genes? It is furthermore evident that many other residues in FRs, even those belonging to the cores of the antibody fold and to the VH/VL interface, may harbor extensive diversity. We envisage that such diversity, when structurally tolerated may contribute substantially to improved biophysical properties of the encoded antibody or even to affinity maturation, for instance through stabilization of the binding site during affinity maturation (52). In particular, an area with CDR-like potential, CDR4, that resides in a loop adjacent to the

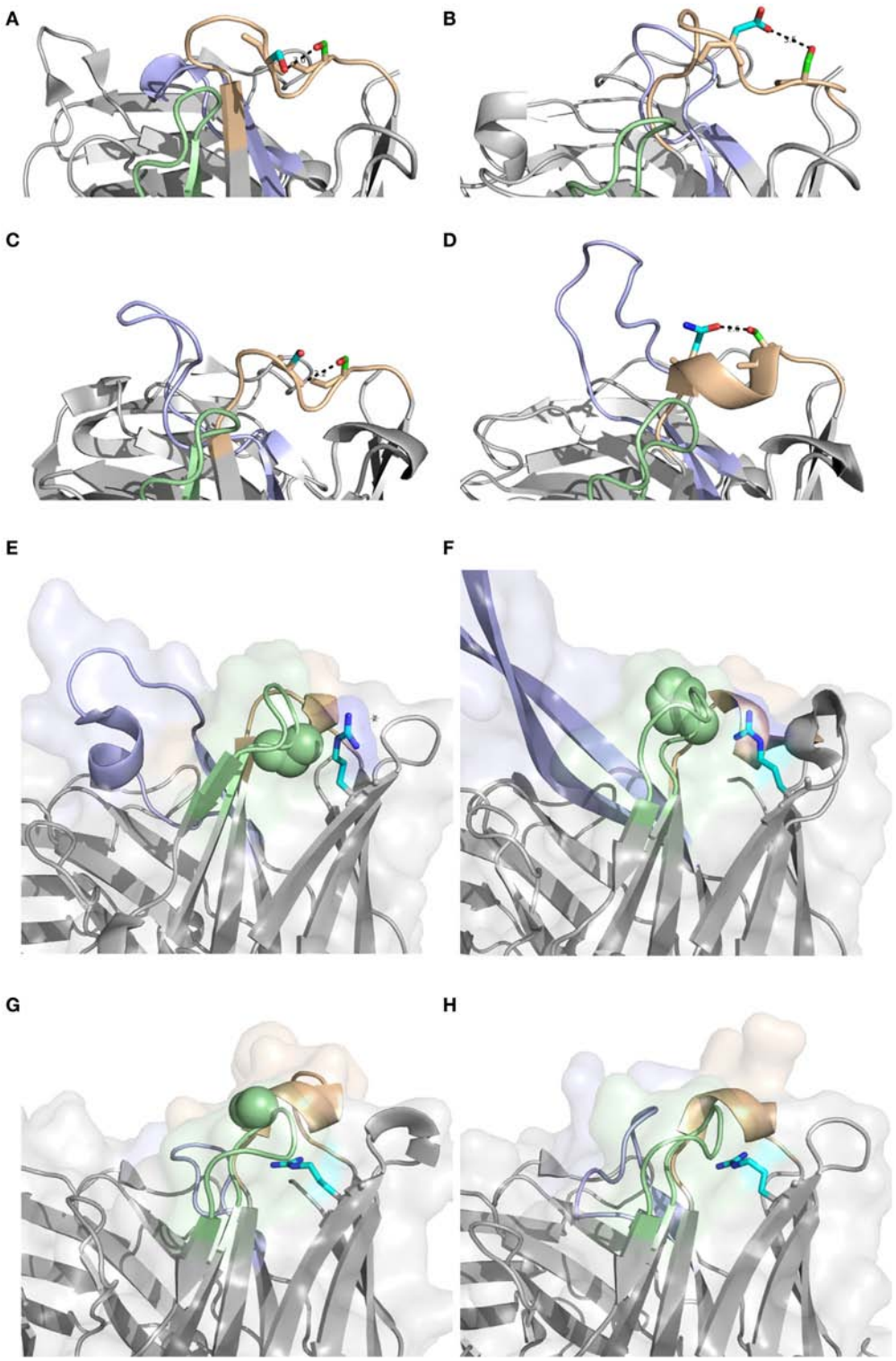


FIGURE 14 | Position 29 of rearranged sequences of different germline gene origins are to different extents targeted by mutagenesis resulting in amino acid substitution (**Figure 3**). A diversity of potential polar interactions of the side chain of S29 have been implicated such as those to Oγ of S31 (IGHV2-5; PDB: 3QRG) **(A)**, Oδ2 of D31 (IGHV4-39; PDB: 5C6T) **(B)**, backbone O of S31 (IGHV4-59; PDB: 3HI1) **(C)**, and Oδ1 of N36 (IGHV5-51; PDB: 4BUH) **(D)**. H chain CDR1, CDR2, and CDR3, are colored in brown, green, and blue, respectively. Side chain atoms are colored in red (oxygen), blue (nitrogen), green (carbon of residue 29) and cyan (carbon of residue 31 or 36). Position of R80 (carbon of side chain shown in cyan) and residue 58 (residues of the side chain as spheres). Structures PDB: 3×3G (with P58) **(E)** and PDB: 3U1S (with H58) **(F)** both derived from IGHV1-8, and PDB: 3FZU (with S58) **(G)** and PDB: 2R56 (with G58) **(H)**, both derived from the IGHV3 subgroup, are shown. H chain CDR1, CDR2, and CDR3, are colored in brown, green, and blue, respectively.

classical CDRs in the folded structure have been defined and exploited (11, 39, 48, 53). It has been suggested to be able to accommodate extensive diversity (39). We have now identified that some, but not all, germline genes introduce diversity in this loop. Such diversity may contribute to functional evolution of antibodies of such germline gene origins. In all, the preferred paths of evolution of antibody V domains extend substantially beyond conventional CDRs in ways directed by an antibody's germline gene origin.

Antibody variable domains, indeed, diversify not only by substitution but also through insertion and deletion of residues into the variable domain sequence (31, 32). In this study, we observed germline gene-inherited patterns that differently target genes with such insertions and deletions, not only in conventional CDRs but also in CDR4 (**Figure 12**; Figure S5 in Supplementary Material). In the past, we hypothesized that the presence of repetitive codons might be one feature that targets such modifications to a particular part of a gene (54). It is also conceivable that parts of VH domains of different germline gene origins are able to structurally harbor such diversity to different extents, a factor that certainly needs further assessment. Nevertheless, it is conceivable that any immune response that relies on introduction of sequence insertion and/or deletion *in vivo* will only recruit members derived from those germline genes that introduce such diversification with ease in critical parts of the sequence. Important immune responses that require such modification have been reported (5, 55) and other responses relying on such evolution, such as those targeting occluded sites, will likely be described in the future. By understanding the ability of particular germline genes to diversify by insertion and deletion, it will be possible to develop our understanding of selection of germline genes made by the immune system in the generation of these particularly difficult immune responses.

Overall, the diverse pattern of diversification even beyond conventional CDR likely complicates computational efforts to assess the involvement of selection during antibody development. We employed one such analysis technology on highly expressed IgG H chain V domain sequences encoded in BM but found no evidence of a positive selection force in the mutational pattern targeting CDR. Our findings are in line with past studies demonstrating a failure to identify evidence of positive selection in CDR, while evidence of negative selection of modifications in FR is detected (43, 44). In agreement with a recent study (46), and given the diversity of paths through which antibodies of different origins evolve, we suggest that any approach to assess selection ought to take germline gene-specific mutational patterns as found in selected and non-selected repertoires into account and not rely entirely on an analysis of mutations based on current global CDR definitions. Processes to facilitate analysis of selection in a germline gene origin-centric fashion have been initiated elsewhere (46). We foresee that such a development will be required if computational approaches are to accurately address the impact of selection on antibody repertoire development. In some situations, such as IgE responses, this aspect is a matter of substantial biological controversy (56, 57) and certainly need further investigations, the outcomes of which will impact our understanding of fundamental biological processes associated with disease.

In summary, we identified germline gene-unique patterns of evolution that occur during hypermutation of antibodies of diverse IGHV germline gene origins. Our findings extend the findings of a recent study, published during preparation of the present manuscript, that identified gene-specific substitution profiles of antibodies of different germline gene origins (46). Collectively, we have demonstrated a diversity of paths taken by antibodies of different germline gene origins to evolve by somatic hypermutation, including not only base substitution but also processes of codon insertion and deletion. Our study forms the basis for improved understanding of molecular evolution as it proceeds in immune responses *in vivo* and establishes a foundation for future germline gene origin-centered analysis approaches.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Regionala etikprövningsnämnden (Lund). All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Regionala etikprövningsnämnden (Lund).

AUTHOR CONTRIBUTIONS

UK: bioinformatic pipeline development and bioinformatic analysis, manuscript preparation, and approved the final manuscript. HP: conceived the study, manuscript preparation, and approved the final manuscript. FL: initial bioinformatic pipeline development, manuscript preparation, and approved the final manuscript. LG: patient management, manuscript preparation, and approved the final manuscript. MO: conceived the study, bioinformatic analysis, main responsibility for manuscript preparation, and approved the final manuscript.

FUNDING

This study was supported by grants from the Swedish Research Council (grant number 2016-01720), and Lund University's Avtal om Läkarutbildning och Forskning (ALF). We acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with NGS and access to the UPPMAX computational infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/fimmu.2017.01433/full#supplementary-material.

REFERENCES

1. Methot SP, Di Noia JM. Molecular mechanisms of somatic hypermutation and class switch recombination. *Adv Immunol* (2017) 133:37–87. doi:10.1016/bs.ai.2016.11.002
2. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32:158–68. doi:10.1038/nbt.2782
3. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* (2015) 7:121. doi:10.1186/s13073-015-0243-2
4. Boyd SD, Crowe JE Jr. Deep sequencing and human antibody repertoire analysis. *Curr Opin Immunol* (2016) 40:103–9. doi:10.1016/j.coi.2016.03.008
5. Kepler TB, Wiehe K. Genetic and structural analyses of affinity maturation in the humoral response to HIV-1. *Immunol Rev* (2017) 275:129–44. doi:10.1111/imr.12513
6. Kwong PD, Chuang GY, Dekosky BJ, Gindin T, Georgiev IS, Lemmin T, et al. Antibodyomics: bioinformatics technologies for understanding B-cell immunity to HIV-1. *Immunol Rev* (2017) 275:108–28. doi:10.1111/imr.12480
7. Lefranc MP. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* (2011) 2011:595–603. doi:10.1101/pdb.top115
8. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) 196:901–17. doi:10.1016/0022-2836(87)90412-8
9. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature* (1989) 342:877–83. doi:10.1038/342877a0
10. Kabat EA, Wu TT, Perry HM, Gottesmann KS, Foeller C. *Sequences of Proteins of Immunological Interest*. Bethesda, MD: U.S. Department of Health and Human Services (1991).
11. Honegger A, Plückthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* (2001) 309:657–70. doi:10.1006/jmbi.2001.4662
12. Abhinandan KR, Martin AC. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol* (2008) 45:3832–9. doi:10.1016/j.molimm.2008.05.022
13. Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* (2011) 2011:633–42. doi:10.1101/pdb.ip85
14. Levin M, Levander F, Palmason R, Greiff L, Ohlin M. Antibody-encoding repertoires of bone marrow and peripheral blood-a focus on IgE. *J Allergy Clin Immunol* (2017) 139:1026–30. doi:10.1016/j.jaci.2016.06.040
15. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
16. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30:1930–2. doi:10.1093/bioinformatics/btu138
17. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
18. Kirik U, Greiff L, Levander F, Ohlin M. Data on haplotype-supproted immunoglobulin germline gene inference. *Data Brief* (2017) 13:620–40. doi:10.1016/j.dib.2017.06.031
19. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol Immunol* (2017) 87:12–22. doi:10.1016/j.molimm.2017.03.012
20. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, Mcwilliam H, et al. Clustal W and clustal X version 2.0. *Bioinformatics* (2007) 23:2947–8. doi:10.1093/bioinformatics/btm404
21. Giudicelli V, Brochet X, Lefranc MP. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* (2011) 2011:695–715. doi:10.1101/pdb.prot5633
22. Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, et al. A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science* (2011) 333:843–50. doi:10.1126/science.1204839
23. McLellan JS, Pancera M, Carrico C, Gorman J, Julien JP, Khayat R, et al. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* (2011) 480:336–43. doi:10.1038/nature10696
24. Tamada T, Shinmi D, Ikeda M, Yonezawa Y, Kataoka S, Kuroki R, et al. TRAIL-R2 superoligomerization induced by human monoclonal agonistic antibody KMTR2. *Sci Rep* (2015) 5:17936. doi:10.1038/srep17936
25. Niemi M, Jylha S, Laukkanen ML, Söderlund H, Makinen-Kiljunen S, Kallio JM, et al. Molecular interactions between a recombinant IgE antibody and the beta-lactoglobulin allergen. *Structure* (2007) 15:1413–21. doi:10.1016/j.str.2007.09.012
26. Houde D, Arndt J, Domeier W, Berkowitz S, Engen JR. Characterization of IgG1 conformation and conformational dynamics by hydrogen/deuterium exchange mass spectrometry. *Anal Chem* (2009) 81:2644–51. doi:10.1021/ac802575y
27. Chandramouli S, Ciferri C, Nikitin PA, Calo S, Gerrein R, Balabanis K, et al. Structure of HCMV glycoprotein B in the postfusion conformation bound to a neutralizing human antibody. *Nat Commun* (2015) 6:8176. doi:10.1038/ncomms9176
28. Chen L, Kwon YD, Zhou TQ, Wu XL, O'dell S, Cavacini L, et al. Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. *Science* (2009) 326:1123–7. doi:10.1126/science.1175868
29. Levin M, Davies AM, Liljekvist M, Carlsson F, Gould HJ, Sutton BJ, et al. Human IgE against the major allergen Bet v 1 – defining an epitope with limited cross-reactivity between different PR-10 family proteins. *Clin Exp Allergy* (2014) 44:288–99. doi:10.1111/cea.12230
30. Ehrenmann F, Lefranc MP. IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc* (2011) 2011:750–61. doi:10.1101/pdb.prot5637
31. Ohlin M, Borrebaeck CAK. Insertions and deletions in hypervariable loops of antibody heavy chains contribute to molecular diversity. *Mol Immunol* (1998) 35:233–8. doi:10.1016/S0161-5890(98)00030-3
32. Wilson PC, De Bouteiller O, Liu YJ, Potter K, Banchereau J, Capra JD, et al. Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *J Exp Med* (1998) 187:59–70. doi:10.1084/jem.187.1.59
33. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* (2011) 39:W499–504. doi:10.1093/nar/gkr413
34. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) 40:e134. doi:10.1093/nar/gks457
35. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) 184:6986–92. doi:10.4049/jimmunol.1000445
36. Honegger A, Malebranche AD, Rothlisberger D, Plückthun A. The influence of the framework core residues on the biophysical properties of immunoglobulin heavy chain variable domains. *Protein Eng Des Sel* (2009) 22:121–34. doi:10.1093/protein/gzn077
37. Ewert S, Huber T, Honegger A, Plückthun A. Biophysical properties of human antibody variable domains. *J Mol Biol* (2003) 325:531–53. doi:10.1016/S0022-2836(02)01237-8
38. Tramontano A, Chothia C, Lesk AM. Framework residue-71 is a major determinant of the position and conformation of the 2nd hypervariable region in the Vh domains of immunoglobulins. *J Mol Biol* (1990) 215:175–82. doi:10.1016/S0022-2836(05)80102-0
39. Bond CJ, Wiesmann C, Marsters JC Jr, Sidhu SS. A structure-based database of antibody variable domain diversity. *J Mol Biol* (2005) 348:699–709. doi:10.1016/j.jmb.2005.02.063
40. Lossos IS, Tibshirani R, Narasimhan B, Levy R. The inference of antigen selection on Ig genes. *J Immunol* (2000) 165:5122–6. doi:10.4049/jimmunol.165.9.5122
41. Dahlke I, Nott DJ, Ruhno J, Sewell WA, Collins AM. Antigen selection in the IgE response of allergic and nonallergic individuals. *J Allergy Clin Immunol* (2006) 117:1477–83. doi:10.1016/j.jaci.2005.12.1359
42. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) 20:683–94. doi:10.1093/intimm/dxn026

43. MacDonald CM, Boursier L, D'cruz DP, Dunn-Walters DK, Spencer J. Mathematical analysis of antigen selection in somatically mutated immunoglobulin genes associated with autoimmunity. *Lupus* (2010) 19:1161–70. doi:10.1177/0961203310367657
44. Levin M, Ohlin M. Inconclusive evidence for or against positive antigen selection in the shaping of human immunoglobulin E repertoires: a call for new approaches. *Int Arch Allergy Immunol* (2013) 161:122–6. doi:10.1159/000345421
45. Ward AB, Wilson IA. The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunol Rev* (2017) 275:21–32. doi:10.1111/imr.12507
46. Sheng Z, Schramm CA, Kong R, Program NCS, Mullikin JC, Mascola JR, et al. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front Immunol* (2017) 8:537. doi:10.3389/fimmu.2017.00537
47. Rogozin IB, Diaz M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G: C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol* (2004) 172:3382–4. doi:10.4049/jimmunol.172.6.3382
48. Burkovitz A, Sela-Culang I, Ofran Y. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J* (2014) 281:306–19. doi:10.1111/febs.12597
49. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* (1997) 273:927–48. doi:10.1006/jmbi.1997.1354
50. Nowak J, Baker T, Georges G, Kelm S, Klostermann S, Shi J, et al. Length-independent structural similarities enrich the antibody CDR canonical class model. *MAbs* (2016) 8:751–60. doi:10.1080/19420862.2016.1158370
51. Tomlinson IM, Walter G, Jones PT, Dear PH, Sonnhammer EL, Winter G. The imprint of somatic hypermutation on the repertoire of human germline V genes. *J Mol Biol* (1996) 256:813–7. doi:10.1006/jmbi.1996.0127
52. Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC. Structural insights into the evolution of an antibody combining site. *Science* (1997) 276:1665–9. doi:10.1126/science.276.5319.1665
53. Young NM, Watson DC, Cunningham AM, Mackenzie CR. The intrinsic cysteine and histidine residues of the anti-*Salmonella* antibody Se155-4: a model for the introduction of new functions into antibody-binding sites. *Protein Eng Des Sel* (2014) 27:383–90. doi:10.1093/protein/gzu018
54. Lantto J, Ohlin M. Uneven distribution of repetitive trinucleotide motifs in human immunoglobulin heavy variable genes. *J Mol Evol* (2002) 54:346–53. doi:10.1007/s00239-001-0049-2
55. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* (2013) 39:245–58. doi:10.1016/j.immuni.2013.04.012
56. Davies JM, Platts-Mills TA, Aalberse RC. The enigma of IgE+ B-cell memory in human subjects. *J Allergy Clin Immunol* (2013) 131:972–6. doi:10.1016/j.jaci.2012.12.1569
57. Gadermaier E, Levin M, Flicker S, Ohlin M. The human IgE repertoire. *Int Arch Allergy Immunol* (2014) 163:77–91. doi:10.1159/000355947

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.

Copyright © 2017 Kirik, Persson, Levander, Greiff and Ohlin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

REVIEW
published: 08 December 2017
doi: 10.3389/fimmu.2017.01753

How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data

Aleksandr Kovaltsuk[1], Konrad Krawczyk[1], Jacob D. Galson[2], Dominic F. Kelly[3], Charlotte M. Deane[1]† and Johannes Trück[2]*†*

[1] Department of Statistics, University of Oxford, Oxford, United Kingdom, [2] Division of Immunology and the Children's Research Center, University Children's Hospital, University of Zürich, Zürich, Switzerland, [3] Oxford Vaccine Group, Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Center, Oxford, United Kingdom

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Christopher Vollmers, University of California, Santa Cruz, United States
Jeffrey J. Gray, Johns Hopkins University, United States
Jeliazko R. Jeliazkov, Johns Hopkins University, United States

***Correspondence:**
Charlotte M. Deane
deane@stats.ox.ac.uk;
Johannes Trück
johannes.trueck@kispi.uzh.ch

†Joint senior authors.

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *11 October 2017*
Accepted: *27 November 2017*
Published: *08 December 2017*

Citation:
Kovaltsuk A, Krawczyk K, Galson JD, Kelly DF, Deane CM and Trück J (2017) How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. Front. Immunol. 8:1753. doi: 10.3389/fimmu.2017.01753

Next-generation sequencing of immunoglobulin gene repertoires (Ig-seq) allows the investigation of large-scale antibody dynamics at a sequence level. However, structural information, a crucial descriptor of antibody binding capability, is not collected in Ig-seq protocols. Developing systematic relationships between the antibody sequence information gathered from Ig-seq and low-throughput techniques such as X-ray crystallography could radically improve our understanding of antibodies. The mapping of Ig-seq datasets to known antibody structures can indicate structurally, and perhaps functionally, uncharted areas. Furthermore, contrasting naïve and antigenically challenged datasets using structural antibody descriptors should provide insights into antibody maturation. As the number of antibody structures steadily increases and more and more Ig-seq datasets become available, the opportunities that arise from combining the two types of information increase as well. Here, we review how these data types enrich one another and show potential for advancing our knowledge of the immune system and improving antibody engineering.

Keywords: Ig-seq, antibody modeling, B cell, Antibodies, Developability, computational modeling, Next-generation sequencing

INTRODUCTION

Antibodies are proteins produced by the B cells of jawed vertebrates. Their primary function is to recognize structural sequence motifs (epitopes) within molecules (antigens) usually related to pathogens, which may lead to direct neutralization of those pathogens or their toxins. Further functions of antibodies are activation of the complement system or tagging of antigens for elimination by other immune pathways. Antibodies have the capacity for binding an extraordinary variety of epitopes as a result of their sequence diversity, which is estimated at 10^{13} unique molecules in the human antibody repertoire (1). An antibody is a large complex molecule (~150 kDa). It can be divided into two parts, the crystallizable fragment (Fc) and the antigen binding fragment (Fab). The Fab fragment is further split into constant and variable regions. There are five possible main Fc portions in humans, and which one is used on a particular antibody is governed by the process of class switching (2). The variable region (Fv) is composed of two domains called the heavy (VH) and light (VL) chains. Within each B cell, the antibody Fv domains are built by somatic recombination between V(D)J segments (3, 4). Upon antigen recognition, somatic hypermutation introduces further diversification into the naïve Fv domains (5). Within each of the VL and VH chains lie three hypervariable loops, the complementarity determining regions (CDRs), which are the most

diverse parts of the antibody (**Figure 1**). These loops form the majority of chemical interactions with antigens, thus defining the antigen-binding region, the paratope (6). The CDR3 of the heavy chain (CDR-H3) is the most diverse of the CDRs as it is being formed at the join between the V, D, and J gene segments and subject to high levels of hypermutation. As a result of this diversity, CDR-H3 plays a key role in antigen recognition and binding (7). The non-CDR sections of the variable domain are called the framework. Framework positions next to CDRs along with CDR sequence govern the structural shape of the loops (8).

The properties of antibodies, in particular designable antigen recognition specificity and binding affinity, have made them useful as diagnostics and research agents as well as the most successful class of biopharmaceuticals (10). Although small molecules constitute the largest proportion of potential therapeutics in clinical trials, the antibody market is steadily growing, with new antibody approvals at a rate of about four per year. As of 2016, five out of the 10 best-selling drugs worldwide were recombinant monoclonal antibodies (11).

Successful exploitation of antibodies relies on our ability to interrogate their diversity and function. Application of next-generation sequencing of immunoglobulin gene repertoire (Ig-seq) to antibody profiling is able to produce comprehensive snapshots of the repertoire diversity (12). However, most Ig-seq techniques are currently unable to perform sequencing of paired heavy–light antibody sequences or to obtain an immunoglobulin gene repertoire solely from antibody-secreting B cells (13–15). Advances in liquid chromatography tandem-mass spectroscopy (LC-MS/MS) now allow high-throughput analysis of serum antibodies at the amino-acid sequence level (16, 17). Previously transcriptomics and Ig-seq datasets have been used to deconvolute MS spectra of serum antibodies into constituent full-length entities (18). Such combined Ig-seq and LC-MS/MS techniques have provided new insights in vaccination and autoimmunity studies (19, 20). Recent advances in computational tools that integrate *de novo* antibody sequencing, error correction data, and sequence homology databases now permit an accurate assembly of full-length antibodies based on the remit of LC-MS/MS spectra alone (21).

The biggest advantage of Ig-seq and LC-MS/MS techniques is their high-throughput nature. This means that the methods provide a broad-brush description and quantification of antibodies in the repertoire. However, this will often include inaccurate data caused by PCR or sequencing errors. The limitation of Ig-seq and LC-MS/MS methods is that they provide sequence information only, whereas it is the shape/structure of an antibody that determines its exact biological function. For instance, antibody CDRs with low-sequence identities can adopt structurally close shapes,

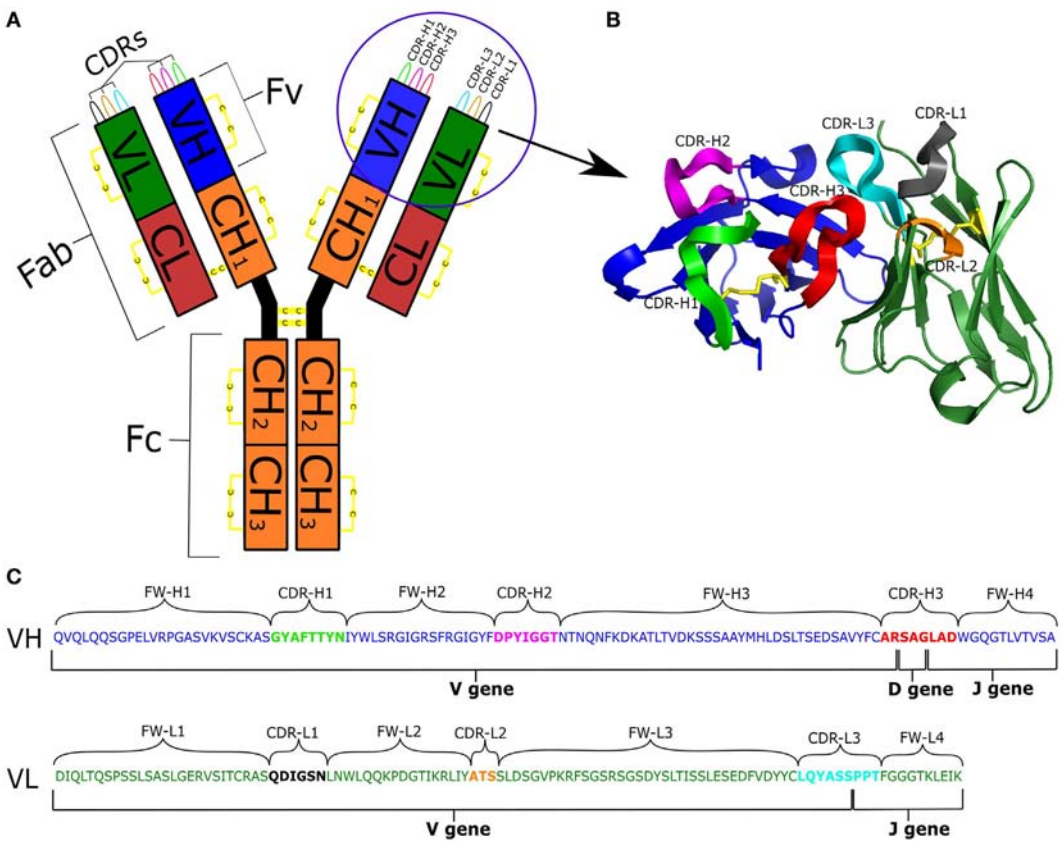


FIGURE 1 | **(A)** Schematic representation of an antibody IgG structure. **(B)** Structure of the Fv region. **(C)** Genetic composition of VH and VL chains [IMGT numbering (9)]: VH is colored blue; VL is green; CDRs are labeled and depicted in different colors; and disulfide bonds are in yellow.

and hence present conformationally similar, though perhaps chemically different, binding sites (22). Knowledge of antibody structure is vital for inferring chemistry of antigen recognition as well as allowing binding site comparison between antibodies. Current experimental determination of antibody structures is achieved by X-ray crystallography or NMR spectroscopy. However, collecting such detailed experimental information limits the rate of analysis to the level of individual or a small number of antibodies (23).

To help tackle the rising costs and time required for engineering and characterization of antibodies, a number of computational tools have been developed that can facilitate experimental efforts. Computational methods are used to profile the physico-chemical properties of antibodies, predict antibody–antigen contacts, and redesign antibody–antigen complexes (24, 25). The tools can be broadly divided into those that require only the sequence of an antibody as input and those that require the structure of the antibody. The inclusion of structural information where available has been shown to improve prediction of most properties over sequence-based methods (26). These improved predictions are only possible if a native structure or an accurate model of the antibody is available.

Since the structure of an antibody is key to its function and high-throughput crystallographic determination of the structures of every antibody is currently not feasible, computational modeling techniques may aid to reduce attrition in the biopharmaceutical industry and to accelerate drug discovery (27). The development of systematic relationships between the antibody information gathered from Ig-seq and techniques such as X-ray crystallography, NMR spectroscopy, and tandem LC-MS/MS could radically improve our understanding of antibody biology. As the number of antibody structures steadily increases and more Ig-seq datasets become available, the opportunities that arise from combining them increase as well. As of October 9, 2017, more than 2,860 antibody structures were available in the Protein Data Bank (PDB) (28) as identified by the Structural Antibody Database (29). The publically available volume of sequences produced from Ig-seq experiments is now in the hundreds of millions (30). In this manuscript, we consider the information obtained from high-throughput sequencing experiments and antibody structures. We review how these datasets can enrich one another and with the help of computational techniques, advance our knowledge of antibody diversity, maturation, and selection and pave the way for improved antibody engineering.

IMMUNOGLOBULIN GENE REPERTOIRE SEQUENCING TECHNOLOGIES

Ig-seq offers high-throughput characterization of immunoglobulin gene sequences at great depth and typically includes several B-cell samples in a single-sequencing run. By controlling the number of samples that are combined and the number of B cells contained therein, it is possible to obtain a large fraction of an immunoglobulin repertoire from a sample. The potential applications of Ig-seq include vaccine and drug development as well as immunodiagnostics (12, 31, 32). Such applications rely on our ability to efficiently identify the population of antibodies responding to an antigen challenge. Ig-seq has already been successfully applied to isolate antigen-specific antibodies from immunized animals in conjunction with common laboratory screening platforms such as phage display (33) or hybridoma (34) or even when the screening step was omitted (35). Furthermore, amino-acid sequence convergences in the CDR-H3 have been observed in the response to a variety of antigens, and may serve as an additional way to isolate antigen-specific antibodies through identifying sequences common among several individuals exposed to the same antigen (30, 36–39).

Heavy and light chains are products of two independent mRNA transcripts that co-assemble into full-length immunoglobulin molecules in the endoplasmic reticulum of the B cell. However, cognate pairing is lost after B-cell bulk lysis prior to Ig-seq and most Ig-seq studies therefore only consider heavy chains (12). However, for human and mouse native pairing is crucial for antibody folding, stability, expression, and antigen binding (40–42). Furthermore, information on the heavy/light chain dimer is required to create an accurate three-dimensional (3D) model of the Fv region and of its antigen-binding pocket which is essential for rational antibody engineering (43). Such models can map antibody sequences to structural space (44), identify the paratope and its physico-chemical properties (45), interrogate the mode of interaction with antigens (46), and predict antibody developability properties (47, 48). Predicting or experimentally obtaining the native VH/VL pairing of the antibody is therefore crucial for our understanding of antibody biology and our ability to engineer these molecules (49).

Several approaches have been devised to circumvent the loss of native pairing in current Ig-seq experiments. Reddy et al. (35) assigned VH/VL pairs based on relative variable chain frequencies in VH and VL chain Ig-seq datasets. This methodology required an accompanying VL Ig-seq dataset and does not always produce antibodies with good pharmacodynamics properties, indicating that it is not always accurate (35). Researchers have also used protein expression platforms, such as recombinant cell lines or phage display, to assign VL to VH chains in a combinatorial fashion followed by experimental screening to identify productive VH/VL combinations (20, 50). Dekosky et al. (15, 51) published the first high-throughput paired VH/VL gene sequencing approach by using single-cell linkage PCR to physically join the VH and VL chains prior to Illumina sequencing. The limitation of this approach is that the current Illumina read length cannot cover the entire paired sequence, so the analysis is restricted to only CDR-H3, CDR-L3, and neighboring framework 4 and proximal positions of framework 3 of respective chains. Once sufficient paired datasets are available, these can potentially act as a reference for guiding computational pairing when VH-only Ig-seq is performed (52). Paired Ig-seq techniques currently yield smaller dataset sizes than unpaired sequencing––for instance, there were 200k sequences for the paired dataset from Dekosky et al. (15) as opposed to 40-m unpaired VH sequences in a recent study (53). The unprecedented speed and depth of Ig-seq techniques both paired and unpaired is unfortunately accompanied by high-sequencing error rates as discussed below.

The four main high-throughput sequencing platforms used to interrogate the immunoglobulin gene repertoire are Illumina, Roche 454, PacBio, and IonTorrent (39, 54–57). Earlier studies often used the Roche 454 technology as it offered greater read lengths than the Illumina methodology. In recent years, Illumina sequencing platforms are usually preferred as they have increasing read length, higher read depth, lower error rates, and lower costs per base (57, 58). Employment of unique molecular identifiers (UIDs) now permits sequencing of the entire antibody chain together with a fragment of a constant domain which holds antibody isotype information (59, 60). Unfortunately, any high-throughput Ig-seq technique suffers from significant error rates (61). Sequencing error can be introduced into Ig-seq datasets from incorrect base calling and sequencing primer artifacts, and has distinct features depending on the sequencing platform used. Error and biases can also originate from the process of preparing sequencing material including reverse transcriptase and polymerase error, amplification of nonproductive V(D)J variable domains during DNA sequencing and multiplex PCR amplification biases (62, 63). Such error may result in the overestimation of the actual number of unique clones in an Ig-seq dataset (62).

Several computational and experimental approaches have been developed to identify and remove or correct erroneous reads (58, 63), though no single-error correction strategy is currently widely used in Ig-seq repertoire analysis (30, 58). In particular, the recent application of UID to Ig-seq can help to correct errors in sequenced transcripts by generating a consensus of reads originating from the same mRNA molecule. As many studies are confined to CDR-H3 analysis, erroneous reads may also be corrected for by using a consensus CDR-H3 sequence for analysis following CDR-H3 clustering (39, 51, 64).

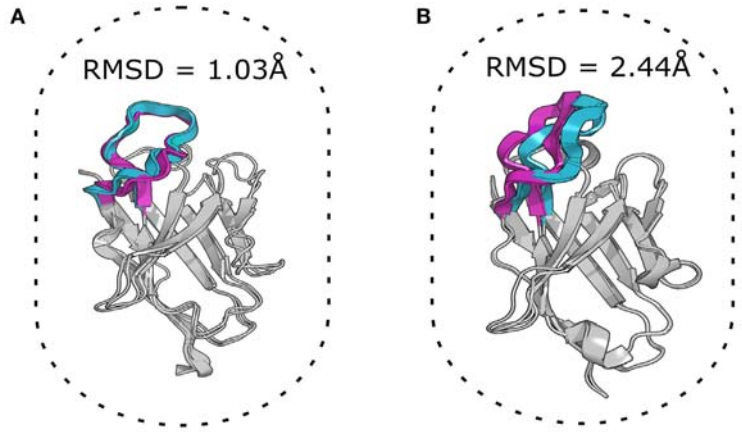
ANTIBODY STRUCTURAL PROPERTIES

The structure of an antibody is crucial in order to understand its function. Antibody–antigen recognition relies on the 3D conformation of the antibody binding site, the paratope, in relation to the cognate epitope on the antigen. In their 3D form, antibodies adopt a Y-shape conformation which can exist in monomer (IgG, IgD, and IgE), dimer (IgA) or pentamer (IgM) forms in humans (65). Several disulfide bonds help to maintain the immunoglobulin fold (**Figure 1**). One set of disulfide bonds hold the heavy constant domains together in the hinge region and another set connects the light and heavy chains (66). Intravariable domain cysteine pairs play a crucial part in shaping the antibody Fv region and artificial disruption of these bonds leads to impaired stability, folding and antigen recognition (67). These cysteines therefore have a crucial role in delineating the structural features of an antibody.

Equivalent residue positions across immunoglobulin sequences and structures can be identified by applying an antibody numbering scheme. Several numbering schemes have been developed to confer consistency and standardization on antibody sequence annotation (9, 22, 68–71). The most commonly used scheme in Ig-seq analysis is the IMGT scheme (12, 39). This numbering was built considering both structural and sequence information (9). The IMGT scheme supports symmetrical amino-acid insertions inside CDRs which ensures that structurally equivalent resides will be annotated the same regardless of CDR length. In contrast, Chothia numbering is often used by structural biologists for its simple CDR loop indel management and inherently structural focus (69, 71).

One of the principal differences between numbering schemes is how they define CDRs. Wu and Kabat (68) were the first to discover and define CDRs as portions of Fv chains that display high-sequence entropy, but as with numbering schemes, there is not a single widely adopted CDR definition and different schemes are used for legacy reasons or for specific features (such as insertion management in IMGT). The different numbering schemes define antibody CDR positions very consistently with the exception of CDR-H1 and CDR-H2 (70). Structural analysis of CDR loops has suggested that all CDRs, except for CDR-H3, adopt a restricted number of conformations, termed canonical classes (22, 72). The canonical classes link sequence patterns to a defined structure (22, 44). This enables the prediction of canonical class structure from sequence. Over the last 30 years, there have been several attempts to cluster CDR sequences/structures (22, 44, 69, 70, 72, 73). On the sequence level, the presence of certain cluster defining key residues indicates the shape the loop can adopt (22, 69, 73). Hence, some changes to the canonical CDRs can be tolerated with no explicit change to loop conformations. The different clustering methods tend to recapitulate previously found groups and find new canonical forms as a result of new data. Most algorithms incorporate CDR loops into clusters with the same number of residues (note that the number of residues varies with different CDR definitions). More recently, Nowak et al. (44) created a novel method of defining length-independent canonical classes based on findings that loops of mismatching lengths can be structurally related. This method allowed fast and accurate structural assignment of a far wider spectrum of canonical CDRs from Ig-seq datasets into fewer canonical clusters (44).

Complementarity determining region-3 of the heavy chain shows a high degree of sequence, length, and structure variation. Due to this diversity, it has so far proved impossible to classify CDR-H3 loops into canonical classes in the manner of the other CDRs. It has been proposed that CDR-H3 can be categorized into "bulged" or "extended" conformations based on the presence of asparagine at position 116 (IMGT numbering) (74, 75). However, increasing knowledge of CDR-H3 structural diversity has shown that the CDR-H3 bulged/extended configuration is difficult to predict solely from sequence (76). The relationship between sequence and structure in CDR-H3 can be important in Ig-seq as current approaches of clonotype assignment are based on CDR-H3 similarity. In this review, we define clonotypes by the presence of identical V, J genes, matching CDR-H3 lengths and CDR-H3 sequence identities greater than 85% (77). However, structural data show that CDR-H3 sequences within distinct clonotypes (sequence-dissimilar) can adopt similar 3D conformations, while those in the same clonotype (similar sequences) can adopt different 3D conformations (**Figure 2**). This suggests that the sequence alone is not a reliable indicator of similarity/difference between structures and therefore cannot



Pair	PDB	V gene	J gene	CDRH3	Sequence identity	RMSD
A	4NZU	IGHV3-30*11	IGHJ4*01	ARAPDCADADCHKGAFGY	27.7%	1.03Å
	4S1S	IGHV1-2*04	IGHJ1*01	VRTADCERDPCKGWVFPH		
B	3U7W	IGHV1-2*02	IGHJ1*01	TRGKYCTARDYYNWDFEH	88%	2.44Å
	4JDV	IGHV1-2*02	IGHJ1*01	ARGKYCTARDYYNWDFQH		

FIGURE 2 | Two aligned pairs of VH chains extracted from SAbDab, the antibody structural database (29). Complementarity determining region-3 of the heavy chain (CDR-H3) sequences in pair **(A)** belong to different CDR-H3 clonotypes but adopt very similar structural configurations with a root mean square deviation (RMSD) of ~1 Å. Pair **(B)** includes germline precursor (4JDV) and matured (3U7W) anti-gp120 antibodies (78, 79). Although CDR-H3 sequences of pair **(B)** are members of the same clonotype, the RMSD shows that their CDR-H3 shapes are structurally distinct (RMSD > 2 Å). CDR-H3 loops and their amino-acid sequences are in purple and cyan colors, mismatched amino acid are in bold. The RMSD of the backbone atom positions of proteins provides a pairwise measurement of the three-dimensional dissimilarity between two sets of coordinates where solved or predicted structures are available. Sub-Angstrom RMSD indicates structurally identical shapes, while an RMSD value greater than 2 Å for a short segment indicates structurally distinct configurations (80).

reliably indicate similar/different binding sites, functional properties and clonotype assignment.

The discrepancy between traditional clonotype assignments and native structure only illustrates how 3D information could be used to draw much more meaningful comparisons between antibodies in an Ig-seq dataset. Such comparisons should not be confined to CDR-H3 alone, but can be extended to the canonical CDRs and the entire Fv region, allowing for much more accurate grouping of functionally related antibodies.

COMPUTATIONAL TOOLS LEVERAGING ANTIBODY STRUCTURE INFORMATION

The increasing number of potential applications of antibodies as therapeutics has led to the development of computational tools which aim to streamline discovery pipelines. Some groups have already demonstrated the viability of *in silico* antibody engineering methodologies in conjunction with experimental workflows (81–84). Computational methods can be broadly divided into those that require a sequence as input and those that require a structure. Methods that require a structure as input accept experimental as well as computational models of the antibody. The large number of experimentally determined antibody structures has enabled researchers to rapidly and accurately model antibodies by leveraging homology methods (8, 85). Below we review current antibody modeling approaches and their applications.

Computational Antibody Modeling

The standard antibody modeling workflow includes four steps (**Figure 3**) (8, 86, 87). The first step is homology modeling of the VH and VL frameworks. The framework template can either be selected by sequence identity to the full-length chain (87) or to individual framework regions (8). Due to framework structure and sequence invariance, current computational tools can model framework structures very accurately (sub-Angstrom precision) (80). The second step is determining the VH/VL orientation, which can be achieved by copying the orientation angle from structures with high Fv sequence identity using VH/VL orientation methods such as AbAngle (88), analytical estimation of the angle using energy functions (89), tailored protein–protein docking (49) or structure-trained machine learning (90). Once the VH/VL orientation is set, it constrains the geometry of the binding site, allowing for the third step, which is modeling of non-H3 CDRs. At this stage, either the canonical classes are used (91) or template-based modeling such as FREAD (92) or ABGEN (93). In the final step, CDR-H3 is modeled using either homology or *ab initio* techniques (94). The resultant antibody model is refined for feasibility of dihedral angles from Ramachandran distribution, side chain orientations and side-chain clashes (89).

Homology modeling approaches can be fast at generating models if a template structure is available. Models can be created using online services: PIGSpro (86), Kotai Antibody Builder (95), and ABodyBuilder (8). Homology modeling is highly dependent on the availability of a similar template structure in

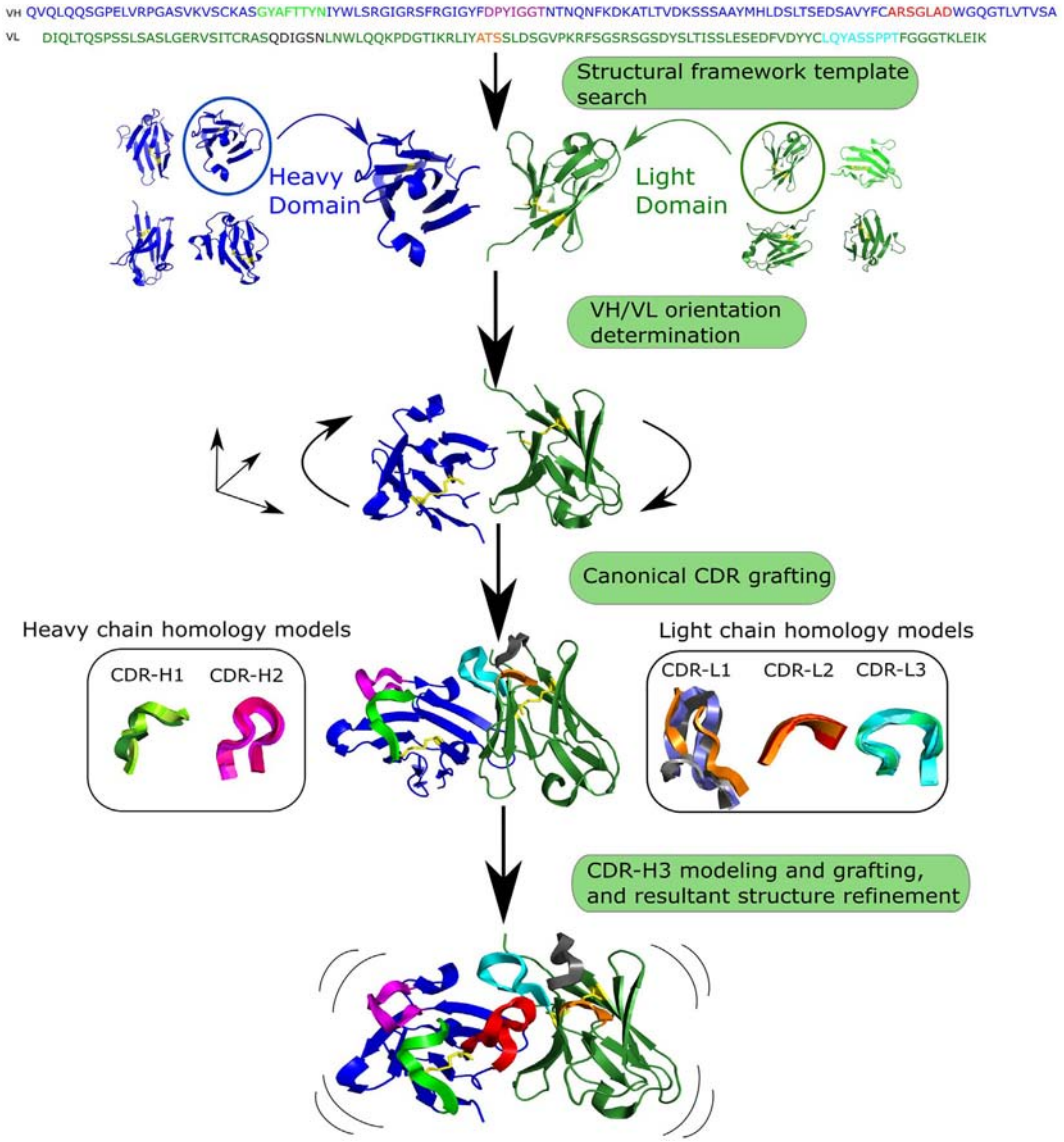


FIGURE 3 | Generalized workflow of antibody modeling. First, heavy and light chain frameworks are determined by homology modeling using templates from known structures. Next, the VH/VL orientation is calculated. The third step is modeling non-H3 complementarity determining regions (CDRs), followed by modeling and grafting of CDR-H3 onto the pre-assembled scaffold. Finally, sidechains are added to the resultant structure and it is refined.

current databases, which can be a problem for CDR-H3 where templates for longer loop length are often unavailable (94). This lack of templates is primarily due to the huge diversity of CDR-H3 shapes (96). An alternative to homology methods in such cases is *ab initio* modeling which does not rely on knowledge of already solved structures. These modeling methods create a large number of potential conformations, often referred to as decoys (97), which makes them computationally expensive compared with homology methods. *Ab initio* approaches include RosettaAntibody (98) and PLOP (99). RosettaAntibody is accessible online *via* the ROSIE (100) website, where a quick antibody modeling option is available which omits the step of intensive searching for low-energy CDR-H3 conformations. Hybrid loop modeling methodologies leverage the advantages of both modeling paradigms. For instance, Accelrys creates an initial loop model with a knowledge-based approach followed by *ab initio* loop refinement (101). More recently, a novel CDR-H3 modeling tool, Sphinx, was developed (102), inspired by the length-independent canonical CDR clustering of Nowak et al. (44). Sphinx outperformed all modeling tools on CDR-H3 structure prediction in an *ex post facto* comparison to the antibody modeling assessment (80). Despite development of different approaches, no single tool currently exists that is able to reliably model native CDR-H3 configurations. Accurate predictions of the CDR-H3 specifically and other CDRs in general are crucial to structurally characterize the antibody–antigen complex.

Performance of antibody modeling tools has been assessed in two blind studies, AMA-I and AMA-II (80, 103), where several

computational tools were benchmarked against a small number of X-ray solved but unpublished antibody crystal structures. Models of frameworks and canonical CDRs are usually accurate within 1–1.5 Å root mean square deviation (RMSD), respectively (see **Figure 2** for description of RMSD), which is very close to native structure. However, CDR-H3 prediction remains the biggest hurdle for computational antibody modeling as average accuracies for this step ranged between 2 and 3 Å RMSD, indicating a decidedly different structure to the native fold. Predictions of this quality are usually not suitable for rational design applications (80, 104).

AMA-II suggested that antibody modeling tools on average produce models of approximately similar accuracies with higher RMSD for longer loop lengths. However, the time required is radically different between homology and *ab initio* approaches (80). Homology modeling can produce a model on average in under a minute [ABodyBuilder (8)], whereas *ab initio* approaches may require up to tens of CPU hours per model [RosettaAntibody takes 482 CPU hours on average per model (100)]. To be able to use a fast homology method a suitable template is needed. Such templates are becoming more frequently available as the number of solved antibody structures increases (29). In order to model millions of sequences in a typical Ig-seq dataset, speed is crucial. Modeling at such high throughput can currently only be achieved by tools such as ABodyBuilder, which is able to generate a model within ~30 s (8). However, further increasing the rate and accuracy of antibody modeling, and developing new ways of speeding up CDR-H3 prediction, are needed if we are to structurally characterize complete Ig-seq datasets.

The accuracy and speed of some computational tools mean that thousands of sequences from Ig-seq datasets can be modeled. Such structurally annotated Ig-seq datasets allow more relevant comparisons of CDRs, binding sites and thus a more accurate grouping of molecules (**Figure 2**). The improved capacity to compare and group antibodies allows us to better visualize the antibody structure space and to investigate structural convergences of paratopes, which can be important for vaccine development (36, 37). In addition, modeled Ig-seq data can be used as input for several computational tools which annotate structure-derived antibody properties, such as therapeutic viability of the molecule (105).

Computational Prediction of Developability

Developing an antibody of high specificity and affinity against a target is only the initial step in engineering a therapeutic molecule. The resulting antibody can carry an array of risks, collectively described as developability, which includes low-expression yields, high-aggregation propensity, and off-target effects (106, 107). In the process of identifying therapeutic candidates, structurally mapped Ig-seq data can be computationally further refined for entities that pass developability criteria (45).

High-aggregation propensity is one of the most undesirable features of antibody therapeutics. Since aggregation is related to the hydrophobicity of the molecule, knowledge of structure is crucial as it allows the calculation of solvent accessible surface area. Structure-based aggregation propensity prediction tools operate by either locating surface-exposed aggregation hot spots and/or leveraging physico-chemical properties of the structure (105, 108). AGGRESCAN3D, a tool inspired by identification of hot spots in the beta amyloid peptide, distinguishes between buried, conformation engaged, and solvent-exposed aggregation prone hydrophobic patches (48). The drawback of this method was that it was not initially designed for antibodies. The Developability Index (DI) was designed for antibodies and is a structure based computational tool that quantitatively assess antibody's propensity to aggregate (105). The DI function considers the net charge of the full-length antibody and hydrophobicity of solvent-exposed sidechains of CDRs.

Such computational tools can be employed early in drug development pipelines to either isolate therapeutically viable drug candidates from the entirety of Ig-seq-derived antibody repertoire (47). Application of such structurally oriented tools requires large-scale modeling of Ig-seq datasets. Nevertheless, to date, there have not been many attempts to combine Ig-seq with structural and computational methods systematically.

COMBINING Ig-seq, STRUCTURAL, AND COMPUTATIONAL APPROACHES

Current approaches to delineate immune repertoires usually employ Ig-seq methodology only, remaining firmly within the remit of information that can be derived from sequences (31, 109, 110). The only study which has attempted to combine paired Ig-seq and structural information to characterize antibody 3D space was that of Dekosky et al. (45). Using high-throughput RosettaAntibody modeling, more than 2,000 models in naïve and antigen-experienced Ig-seq datasets were analyzed. These models helped to obtain a set of structural descriptors such as net charge, surface hydrophobicity of solvent accessible surface area for computationally determined paratopes. However, the choice of methodologies for this study imposed several limitations. Paired VH/VL data did not contain information about the full-length Fv region. Hence, all paired reads had to be completed using respective V germline gene sequences. Moreover, RosettaAntibody modeling speed only permitted the prediction of structure of 1% of the total Ig-seq dataset (2,000 sequences) in 570k CPU hours. Finally, the paired reads with CDR-H3 sequences longer than 16 amino acids were not included in the structural analysis as the modeling accuracy of such loops is currently low. This emphasizes the challenges of modeling longer CDR-H3 configurations (94, 96). Hence, novel fast and reliable CDR-H3 *ab initio* prediction as well as technologically optimized paired VH/VL gene Ig-seq are urgently needed for improved Ig-seq data modeling and interrogation.

RosettaAntibody (98) is a well-established antibody modeling tool and is able to structurally model sequence data; however, its run times make it difficult to structurally characterize the millions of sequences that are gathered during a typical Ig-seq experiment. For this reason, streamlined approaches are being developed to tackle the structural annotation of Ig-seq datasets. For instance, Nowak et al. (44) performed the structural clustering analysis of

TABLE 1 | Summary of currently available resources for computational/structural annotation of antibody sequences.

Tool type	Tool name and reference	Short tool description
ANTIBODY NUMBERING	ANARCI (113)	Variety of schemes (North, Chothia, Kabat, IMGT, AHo). Both online and command line versions are available
ANTIBODY NUMBERING	Abnum (71)	Online numbering tool that operates with Kabat and Chothia schemes
SEQUENCE ANALYSIS	IgBLAST (114)	Nucleotide and amino-acid antibody sequence analysis in IMGT and KABAT schemes
SEQUENCE ANALYSIS	IMGT/HighV-QUEST (115)	Online antibody nucleotide sequence analysis in IMGT numbering scheme
STRUCTURE DATABASE	SabDab (29)	Weekly updating database of all publically available antibody structures.
STRUCTURE/SEQUENCE DATABASE	abYsis (116)	Database of antibody structures and sequences
SEQUENCE DATABASE	DIGIT (111)	Database of antibody sequences
ANTIBODY MODELING	ABodyBuilder (8)	Homology modeling (30 s per model)
ANTIBODY MODELING	PIGSPro (86)	Homology modeling
ANTIBODY MODELING	Kotai Antibody Builder (95)	Homology modeling (90 min per model)
ANTIBODY MODELING	Accelrys (101)	Hybrid modeling (30 min per model)
ANTIBODY MODELING	RosettaAntibody (87)	*Ab initio* modeling (482 CPU hours per model)
ANTIBODY MODELING (COMMERCIAL)	Chemical Computing group (80)	Homology modeling tool combined with molecular dynamics (30 min per model)
CDR-H3 MODELING	Sphinx (102)	Length-independent hybrid modeling (30 min per model)
CDR-H3 MODELING	PLOP (99)	*Ab initio* modeling
CDR-H3 MODELING	FREAD (85)	Homology modeling (2 min per model)
PARATOPE PREDICTION	Paratome (117)	Structural consensus to identify additional antigen recognizing regions outside the CDRs
PARATOPE PREDICTION	i-Patch (118)	Statistical inference to devise a likelihood for a position to form a potential contact
PARATOPE PREDICTION	proABC (119)	Sequence-based method that leverages machine learning to predict residues that form interactions

Many of these tools have online presence and links to these are available on our website http://antibodystructure.org.

CDR-L3 of two large Ig-seq datasets: 200k paired Ig-seq sample from Dekosky et al. (15) and 9-m in-house UCB Pharma Ltd sequences as well as a database of 71k antibody sequences [DIGIT (111)]. Every CDR-L3 sequence was submitted to HMMER (112) to assign it to a length-independent cluster. This is the first instance of structurally mapping the entirety of an Ig-seq dataset. The method can be extrapolated to any non-H3 CDR to provide structural annotation of sampling of loop shapes as well as to identify yet uncharacterized loop configurations.

Structural characterization of large sequence sets can be extended to the entire Fv region. The modeling method, ABodyBuilder, was used to predict structures of 6,000 paired antibody sequences from public repositories (8). The average modeling time per 1,000 antibody sequences was 567 CPU hours compared with 285,000 CPU hours using RosettaAntibody (45). ABodyBuilder produces model accuracies that are in line with the AMA-II values (80). Using tools such as ABodyBuilder, one can perform large-scale structural modeling of Ig-seq data. Such structural characterization of Ig-seq similarity/difference would allow more accurate inter-molecule comparisons and assessment of developability. The structural software outlined in this manuscript together with other tools that are often employed in computational/structural annotation of antibody sequences is summarized in **Table 1**.

CONCLUSION

The ability to engineer better antibody-based therapeutics relies on our knowledge of the exact sequence and the 3D shape of individual molecules within the antibody repertoire. Next-generation sequencing methodologies that can yield millions of immunoglobulin gene sequences in a single sequencing run have already given insights into the steady-state and antigen-stimulated B-cell receptor repertoire (12, 32). On the other hand, low-throughput techniques such as X-ray crystallography can provide detailed information about individual antibody structures. Computational methodologies can offer a bridge between the two fields by allowing structural annotation of Ig-seq experiments (8, 44, 45). Availability of antibody structures and maturity of modeling techniques means it is now possible to perform large-scale structural characterizations of Ig-seq samples. This enriched structural content can be used to perform more precise characterization of antibodies allowing inter-antibody comparisons and grouping of structurally similar sequences (that may not be possible on the sequence level) as well as annotation of developability information. Large-scale Ig-seq datasets can also direct computational tools for targeted interrogation of antibody structural space. Statistical knowledge of the distribution of the antibody structures and sequences can offer crystallographers an idea of the common but currently unknown antibody variants. The Ig-seq and structural communities will benefit from cross-fertilization of ideas and methodologies. Together they will advance our knowledge of the antibodies in health and disease and pave the way for more advanced antibody-based therapeutics.

AUTHOR CONTRIBUTIONS

All authors contributed to the development of writing of the manuscript.

FUNDING

This work was supported by funding from Biotechnology and Biological Sciences Research Council (BBSRC) [BB/M011224/1]

and UCB Pharma Ltd awarded to AK. DK receives salary support from the NIHR Oxford Biomedical Research Centre. JT is funded by the Swiss National Science Foundation through an Ambizione-SCORE grant and has received further funding from the Olga Mayenfisch Foundation Zurich and the Bangerter-Rhyner Foundation Basel.

REFERENCES

1. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol* (2015) 36:738–49. doi:10.1016/j.it.2015.09.006
2. Vidarsson G, Dekkers G, Rispens T. IgG subclasses and allotypes: from structure to effector functions. *Front Immunol* (2014) 5:520. doi:10.3389/fimmu.2014.00520
3. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302:575–81. doi:10.1038/302575a0
4. Lefranc MP. IMGT, the international ImMunoGeneTics database®. *Nucleic Acids Res* (2003) 31:307–10. doi:10.1093/nar/gkg085
5. French D, Laskov R, Scharff M. The role of somatic hypermutation in the generation of antibody diversity. *Science* (1989) 244:1152–7. doi:10.1126/science.2658060
6. Collis AVJ, Brouwer AP, Martin ACR. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol* (2003) 325:337–54. doi:10.1016/S0022-2836(02)01222-6
7. Xu JL, Davis MM. Diversity in the CDR3 region of V H is sufficient for most antibody specificities. *Immunity* (2000) 13:37–45. doi:10.1016/S1074-7613(00)00006-6
8. Leem J, Dunbar J, Georges G, Shi J, Deane CM. ABodyBuilder: automated antibody structure prediction with data-driven accuracy estimation. *MAbs* (2016) 8:1259–68. doi:10.1080/19420862.2016.1205773
9. Lefranc M-P, Pommié C, Ruiz M, Giuducelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev Comp Immunol* (2003) 27:55–77. doi:10.1016/S0145-305X(02)00039-3
10. Reichert JM. Antibodies to watch in 2017. *MAbs* (2017) 9:167–81. doi:10.1080/19420862.2016.1269580
11. Strohl WR. Current progress in innovative engineered antibodies. *Protein Cell* (2017):1–35. doi:10.1007/s13238-017-0457-8
12. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotech* (2014) 32:158–68. doi:10.1038/nbt.2782
13. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol* (2013) 25:646–52. doi:10.1016/j.coi.2013.09.017
14. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Curr Opin Chem Biol* (2015) 24:112–20. doi:10.1016/j.cbpa.2014.11.007
15. Dekosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2014) 21:1–8. doi:10.1038/nm.3743
16. Obermeier B, Mentele R, Malotka J, Kellermann J, Kümpfel T, Wekerle H, et al. Matching of oligoclonal immunoglobulin transcriptomes and proteomes of cerebrospinal fluid in multiple sclerosis. *Nat Med* (2008) 14:688–93. doi:10.1038/nm1714
17. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* (2014) 111:2259–64. doi:10.1073/pnas.1317793111
18. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM. Proteogenomics: integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Annu Rev Anal Chem (Palo Alto Calif)* (2016) 9:521–45. doi:10.1146/annurev-anchem-071015-041722
19. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat Med* (2016) 22:1456–64. doi:10.1038/nm.4224
20. Chen J, Zheng Q, Hammers CM, Ellebrecht CT, Mukherjee EM, Tang HY, et al. Proteomic analysis of pemphigus autoantibodies indicates a larger, more diverse, and more dynamic repertoire than determined by B cell genetics. *Cell Rep* (2017) 18:237–47. doi:10.1016/j.celrep.2016.12.013
21. Tran NH, Rahman MZ, He L, Xin L, Shan B, Li M. Complete de novo assembly of monoclonal antibody sequences. *Sci Rep* (2016) 6:31730. doi:10.1038/srep31730
22. North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. *J Mol Biol* (2011) 406:228–56. doi:10.1016/j.jmb.2010.10.030
23. Li Y, Li H, Yang F, Smith-Gill SJ, Mariuzza RA. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Mol Biol* (2003) 10:482–8. doi:10.1038/nsb930
24. Sela-Culang I, Benhnia MREI, Matho MH, Kaever T, Maybeno M, Schlossman A, et al. Using a combined computational-experimental approach to predict antibody-specific B cell epitopes. *Structure* (2014) 22:646–57. doi:10.1016/j.str.2014.02.003
25. Sircar A, Gray JJ. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput Biol* (2010) 6:e1000644. doi:10.1371/journal.pcbi.1000644
26. Krawczyk K, Liu X, Baker T, Shi J, Deane CM. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics* (2014) 30:2288–94. doi:10.1093/bioinformatics/btu190
27. Ecker DM, Jones SD, Levine HL. The therapeutic monoclonal antibody market. *MAbs* (2015) 7:9–14. doi:10.4161/19420862.2015.989042
28. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* (2007) 35:D301–3. doi:10.1093/nar/gkl971
29. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. *Nucleic Acids Res* (2014) 42:D1140–6. doi:10.1093/nar/gkt1043
30. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep* (2017) 19:1467–78. doi:10.1016/j.celrep.2017.04.054
31. Galson JD, Pollard AJ, Trück J, Kelly DF. Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol* (2014) 35:319–31. doi:10.1016/j.it.2014.04.005
32. Parola C, Neumeier D, Reddy ST. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. *Immunology* (2017). doi:10.1111/imm.12838
33. Yang W, Yoon A, Lee S, Kim S, Han J, Chung J. Next-generation sequencing enables the discovery of more diverse positive clones from a phage-displayed antibody library. *Exp Mol Med* (2017) 49:e308. doi:10.1038/emm.2017.22
34. Krause JC, Tsibane T, Tumpey TM, Huffman CJ, Briney BS, Smith SA, et al. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *J Immunol* (2011) 187:3704–11. doi:10.4049/jimmunol.1101823
35. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* (2010) 28:965–9. doi:10.1038/nbt.1673
36. Trück J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J Immunol* (2015) 194:252–61. doi:10.4049/jimmunol.1401405
37. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) 13:691–700. doi:10.1016/j.chom.2013.05.008
38. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* (2014) 16:105–14. doi:10.1016/j.chom.2014.05.013
39. Galson JD, Trück J, Fowler A, Münz M, Cerundolo V, Pollard AJ, et al. In-depth assessment of within-individual and inter-individual variation

in the B cell receptor repertoire. *Front Immunol* (2015) 6:531. doi:10.3389/fimmu.2015.00531
40. Lowe D, Dudgeon K, Rouet R, Schofield P, Jermutus L, Christ D. Aggregation, stability, and formulation of human antibody therapeutics. *Adv Protein Chem Struct Biol* (2011) 84:41–61. doi:10.1016/B978-0-12-386483-3.00004-5
41. Tiller T, Schuster I, Deppe D, Siegers K, Strohner R, Herrmann T, et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs* (2013) 5:445–70. doi:10.4161/mabs.24218
42. Rouet R, Lowe D, Christ D. Stability engineering of the human antibody repertoire. *FEBS Lett* (2014) 588:269–77. doi:10.1016/j.febslet.2013.11.029
43. Krawczyk K, Dunbar J, Deane CM. Computational tools for aiding rational antibody design. In: Samish I, editor. *Methods in Molecular Biology*. Clifton, NJ: Palgrave Macmillan (2016). p. 399–416.
44. Nowak J, Baker T, Georges G, Kelm S, Klostermann S, Shi J, et al. Length-independent structural similarities enrich the antibody CDR canonical class model. *MAbs* (2016) 8:751–60. doi:10.1080/19420862.2016.1158370
45. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113:E2636–45. doi:10.1073/pnas.1525510113
46. Brenke R, Hall DR, Chuang GY, Comeau SR, Bohnuud T, Beglov D, et al. Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics* (2012) 28:2608–14. doi:10.1093/bioinformatics/bts493
47. Kumar S, Plotnikov NV, Rouse JC, Singh SK. Biopharmaceutical informatics: supporting biologic drug development via molecular modelling and informatics. *J Pharm Pharmacol* (2017). doi:10.1111/jphp.12700
48. Zambrano R, Jamroz M, Szczasiuk A, Pujols J, Kmiecik S, Ventura S. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res* (2015) 43:W306–13. doi:10.1093/nar/gkv359
49. Marze NA, Lyskov S, Gray JJ. Improved prediction of antibody VL-VH orientation. *Protein Eng Des Sel* (2016) 29:409–18. doi:10.1093/protein/gzw013
50. Sato S, Beausoleil SA, Popova L, Beaudet JG, Ramenani RK, Zhang X, et al. Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nat Biotechnol* (2012) 30:1039–43. doi:10.1038/nbt.2406
51. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31:166–9. doi:10.1038/nbt.2492
52. Laffy JMJ, Dodev T, Macpherson JA, Townsend C, Lu HC, Dunn-Walters D, et al. Promiscuous antibodies characterised by their physico-chemical properties: from sequence to structure and back. *Prog Biophys Mol Biol* (2017) 128:47–56. doi:10.1016/j.pbiomolbio.2016.09.002
53. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A public database of memory and naive B-cell receptor sequences. *PLoS One* (2016) 11:e0160853. doi:10.1371/journal.pone.0160853
54. Rounds WH, Ligocki AJ, Levin MK, Greenberg BM, Bigwood DW, Eastman EM, et al. The antibody genetics of multiple sclerosis: comparing next-generation sequencing to sanger sequencing. *Front Neurol* (2014) 5:166. doi:10.3389/fneur.2014.00166
55. Larsen PA, Smith TPL. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunol* (2012) 13:52. doi:10.1186/1471-2172-13-52
56. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci Rep* (2014) 4:6778. doi:10.1038/srep06778
57. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* (2012) 13:341. doi:10.1186/1471-2164-13-341
58. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol* (2017) 35:203–14. doi:10.1016/j.tibtech.2016.09.010
59. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) 11:1599–616. doi:10.1038/nprot.2016.093
60. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier––guided amplicon assembly. *J Immunol* (2016) 196:2902–7. doi:10.4049/jimmunol.1502563
61. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* (2012) 30:434–9. doi:10.1038/nbt.2198
62. Khan TA, Friedensohn S, de Vries ARG, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) 2:e1501371. doi:10.1126/sciadv.1501371
63. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) 11:653–5. doi:10.1038/nmeth.2960
64. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nat Methods* (2013) 10:57–9. doi:10.1038/nmeth.2276
65. Charles A, Janeway J, Travers P, Walport M, Shlomchik MJ. The distribution and functions of immunoglobulin isotypes. *Immunobiology: The Immune System in Health and Disease*. (2001). p. 1–9. Available from: http://www.ncbi.nlm.nih.gov/books/NBK27162/
66. Wang W, Singh S, Zeng DL, King K, Nema S. Antibody structure, instability, and formulation. *J Pharm Sci* (2007) 96:1–26. doi:10.1002/jps.20727
67. Glockshuber R, Schmidt T, Plückthun A. The disulfide bonds in antibody variable domains: effects on stability, folding in vitro, and functional expression in *Escherichia coli*. *Biochemistry* (1992) 31:1270–9. doi:10.1021/bi00120a002
68. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* (1970) 132:211–50. doi:10.1084/jem.132.2.211
69. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* (1997) 273:927–48. doi:10.1006/jmbi.1997.1354
70. Honegger A, Pluckthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* (2001) 309:657–70. doi:10.1006/jmbi.2001.4662
71. Abhinandan KR, Martin ACR. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol* (2008) 45:3832–9. doi:10.1016/j.molimm.2008.05.022
72. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) 196:901–17. doi:10.1016/0022-2836(87)90412-8
73. Martin ACR, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol* (1996) 263:800–15. doi:10.1006/jmbi.1996.0617
74. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* (1998) 275:269–94. doi:10.1006/jmbi.1997.1442
75. Kuroda D, Shirai H, Kobori M, Nakamura H. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* (2008) 73:608–20. doi:10.1002/prot.22087
76. Weitzner BD, Dunbrack RL, Gray JJ. The origin of CDR H3 structural diversity. *Structure* (2015) 23:302–11. doi:10.1016/j.str.2014.11.010
77. Chen Z, Collins AM, Wang Y, Gata BA. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res* (2010) 6:S4. doi:10.1186/1745-7580-6-S1-S4
78. Scharf L, West AP, Gao H, Lee T, Scheid JF, Nussenzweig MC, et al. Structural basis for HIV-1 gp120 recognition by a germ-line version of a broadly neutralizing antibody. *Proc Natl Acad Sci U S A* (2013) 110:6049–54. doi:10.1073/pnas.1303682110
79. Diskin R, Scheid JF, Marcovecchio PM, West AP, Klein F, Gao H, et al. Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science* (2011) 334:1289–93. doi:10.1126/science.1213782
80. Teplyakov A, Luo J, Obmolova G, Malia TJ, Sweet R, Stanfield RL, et al. Antibody modeling assessment II. Structures and models. *Proteins* (2014) 82:1563–82. doi:10.1002/prot.24554

81. Clark LA, Boriack-Sjodin PA, Eldredge J, Fitch C, Friedman B, Hanf KJM, et al. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Sci* (2006) 15:949–60. doi:10.1110/ps.052030506
82. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol* (2007) 25:1171–6. doi:10.1038/nbt1336
83. Thakkar S, Nanaware-Kharade N, Celikel R, Peterson EC, Varughese KI. Affinity improvement of a therapeutic antibody to methamphetamine and amphetamine through structure-based antibody engineering. *Sci Rep* (2014) 4:3673. doi:10.1038/srep03673
84. Choi Y, Hua C, Sentman CL, Ackerman ME, Bailey-Kellogg C. Antibody humanization by structure-based computational protein design. *MAbs* (2015) 7:1045–57. doi:10.1080/19420862.2015.1076600
85. Choi Y, Deane CM. FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins* (2010) 78:1431–40. doi:10.1002/prot.22658
86. Lepore R, Olimpieri PP, Messih MA, Tramontano A. PIGSPro: prediction of immunoGlobulin structures v2. *Nucleic Acids Res* (2017) 45:W17–23. doi:10.1093/nar/gkx334
87. Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, et al. Modeling and docking of antibody structures with Rosetta. *Nat Protoc* (2017) 12:401–16. doi:10.1038/nprot.2016.180
88. Dunbar J, Fuchs A, Shi J, Deane CM. ABangle: characterising the VH-VL orientation in antibodies. *Protein Eng Des Sel* (2013) 26:611–20. doi:10.1093/protein/gzt020
89. Zhu K, Day T, Warshaviak D, Murrett C, Friesner R, Pearlman D. Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. *Proteins* (2014) 82:1646–55. doi:10.1002/prot.24551
90. Bujotzek A, Dunbar J, Lipsmeier F, Schäfer W, Antes I, Deane CM, et al. Prediction of VH-VL domain orientation for antibody variable domain modeling. *Proteins* (2015) 83:681–95. doi:10.1002/prot.24756
91. Marcatili P, Rosi A, Tramontano A. PIGS: automatic prediction of antibody structures. *Bioinformatics* (2008) 24:1953–4. doi:10.1093/bioinformatics/btn341
92. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci* (2001) 10:599–612. doi:10.1110/ps.37601
93. Mandal C, Kingery BD, Anchin JM, Subramaniam S, Linthicum DS. ABGEN: a knowledge-based automated approach for antibody structure modeling. *Nat Biotechnol* (1996) 14:323–8. doi:10.1038/nbt0396-323
94. Marks C, Deane CM. Antibody H3 structure prediction. *Comput Struct Biotechnol J* (2017) 15:222–31. doi:10.1016/j.csbj.2017.01.010
95. Yamashita K, Ikeda K, Amada K, Liang S, Tsuchiya Y, Nakamura H, et al. Kotai antibody builder: automated high-resolution structural modeling of antibodies. *Bioinformatics* (2014) 30:3279–80. doi:10.1093/bioinformatics/btu510
96. Regep C, Georges G, Shi J, Popovic B, Deane CM. The H3 loop of antibodies shows unique structural characteristics. *Proteins* (2017) 85:1311–8. doi:10.1002/prot.25291
97. Zhu K, Day T. Ab initio structure prediction of the antibody hypervariable H3 loop. *Proteins* (2013) 81:1081–9. doi:10.1002/prot.24240
98. Sircar A, Kim ET, Gray JJ. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res* (2009) 37:W474–9. doi:10.1093/nar/gkp387
99. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, et al. A hierarchical approach to all-atom protein loop prediction. *Proteins* (2004) 55:351–67. doi:10.1002/prot.10613
100. Lyskov S, Chou FC, Conchúir SÓ, Der BS, Drew K, Kuroda D, et al. Serverification of molecular modeling applications: the Rosetta online server that includes everyone (ROSIE). *PLoS One* (2013) 8:e63906. doi:10.1371/journal.pone.0063906
101. Fasnacht M, Butenhof K, Goupil-Lamy A, Hernandez-Guzman F, Huang H, Yan L. Automated antibody structure prediction using accelrys tools: results and best practices. *Proteins* (2014) 82:1583–98. doi:10.1002/prot.24604
102. Marks C, Nowak J, Klostermann S, Georges G, Dunbar J, Shi J, et al. Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics* (2017) 33:1346–53. doi:10.1093/bioinformatics/btw823
103. Almagro JC, Beavers MP, Hernandez-Guzman F, Maier J, Shaulsky J, Butenhof K, et al. Antibody modeling assessment. *Proteins* (2011) 79:3050–66. doi:10.1002/prot.23130
104. Kuroda D, Shirai H, Jacobson MP, Nakamura H. Computer-aided antibody design. *Protein Eng Des Sel* (2012) 25:507–21. doi:10.1093/protein/gzs024
105. Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J Pharm Sci* (2012) 101:102–15. doi:10.1002/jps.22758
106. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL. Aggregation in protein-based biotherapeutics: computational studies and tools to identify aggregation-prone regions. *J Pharm Sci* (2011) 100:5081–95. doi:10.1002/jps.22705
107. Trainor K, Broom A, Meiering EM. Exploring the relationships between protein sequence, structure and solubility. *Curr Opin Struct Biol* (2017) 42:136–46. doi:10.1016/j.sbi.2017.01.004
108. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* (2008) 380:425–36. doi:10.1016/j.jmb.2008.05.013
109. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110:13463–8. doi:10.1073/pnas.1312146110
110. Galson JD, Clutterbuck EA, Trück J, Ramasamy MN, Münz M, Fowler A, et al. BCR repertoire sequencing: different patterns of B-cell activation after two meningococcal vaccines. *Immunol Cell Biol* (2015) 93:885–95. doi:10.1038/icb.2015.57
111. Chailyan A, Tramontano A, Marcatili P. A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res* (2012) 40:D1230–4. doi:10.1093/nar/gkr806
112. Eddy SR. Profile hidden Markov models. *Bioinformatics* (1998) 14:755–63. doi:10.1093/bioinformatics/14.9.755
113. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* (2015) 32:298–300. doi:10.1093/bioinformatics/btv552
114. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41:W34–40. doi:10.1093/nar/gkt382
115. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
116. Swindells MB, Porter CT, Couch M, Hurst J, Abhinandan KR, Nielsen JH, et al. abYsis: integrated antibody sequence and structure—management, analysis, and prediction. *J Mol Biol* (2017) 429:356–64. doi:10.1016/j.jmb.2016.08.019
117. Kunik V, Ashkenazi S, Ofran Y. Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res* (2012) 40:W521–4. doi:10.1093/nar/gks480
118. Krawczyk K, Baker T, Shi J, Deane CM. Antibody i-patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng Des Sel* (2013) 26:621–9. doi:10.1093/protein/gzt043
119. Olimpieri PP, Chailyan A, Tramontano A, Marcatili P. Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. *Bioinformatics* (2013) 29:2285–91. doi:10.1093/bioinformatics/btt369

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.

Copyright © 2017 Kovaltsuk, Krawczyk, Galson, Kelly, Deane and Trück. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

METHODS
published: 20 December 2017
doi: 10.3389/fimmu.2017.01796

Pacific Biosciences Sequencing and IMGT/HighV-QUEST Analysis of Full-Length Single Chain Fragment Variable from an *In Vivo* Selected Phage-Display Combinatorial Library

Audrey Hemadou[1], Véronique Giudicelli[2], Melissa Laird Smith[3†], Marie-Paule Lefranc[2], Patrice Duroux[2], Sofia Kossida[2], Cheryl Heiner[3], N. Lance Hepler[3], John Kuijpers[3], Alexis Groppi[4], Jonas Korlach[3], Philippe Mondon[5], Florence Ottones[1], Marie-Josée Jacobin-Valat[1], Jeanny Laroche-Traineau[1‡] and Gisèle Clofent-Sanchez[1*‡]*

[1] CRIMSB, UMR 5536, CNRS, Bordeaux, France, [2] IMGT®, The International ImMunoGeneTics Information System®, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Institut de Génétique Humaine, IGH, UMR 9002, CNRS, Montpellier University, Montpellier, France, [3] Pacific Biosciences, Menlo Park, CA, United States, [4] Université de Bordeaux, Centre de Bioinformatique de Bordeaux (CBiB), Bordeaux, France, [5] LFB Biotechnologies, Lille, France

Phage-display selection of immunoglobulin (IG) or antibody single chain Fragment variable (scFv) from combinatorial libraries is widely used for identifying new antibodies for novel targets. Next-generation sequencing (NGS) has recently emerged as a new method for the high throughput characterization of IG and T cell receptor (TR) immune repertoires both *in vivo* and *in vitro*. However, challenges remain for the NGS sequencing of scFv from combinatorial libraries owing to the scFv length (>800 bp) and the presence of two variable domains [variable heavy (VH) and variable light (VL) for IG] associated by a peptide linker in a single chain. Here, we show that single-molecule real-time (SMRT) sequencing with the Pacific Biosciences RS II platform allows for the generation of full-length scFv reads obtained from an *in vivo* selection of scFv-phages in an animal model of atherosclerosis. We first amplified the DNA of the phagemid inserts from scFv-phages eluted from an aortic section at the third round of the *in vivo* selection. From this amplified DNA, 450,558 reads were obtained from 15 SMRT cells. Highly accurate circular consensus sequences from these reads were generated, filtered by quality and then analyzed by IMGT/HighV-QUEST with the functionality for scFv. Full-length scFv were identified and characterized in 348,659 reads. Full-length scFv sequencing is an absolute requirement for analyzing the associated VH and VL domains enriched during the *in vivo* panning rounds. In order to further validate the ability of SMRT sequencing to provide high quality, full-length scFv sequences, we tracked the reads of an scFv-phage clone P3 previously identified by biological assays and Sanger sequencing. Sixty P3 reads showed 100% identity with the full-length scFv of 767 bp, 53 of them covering the whole insert of 977 bp, which encompassed the primer sequences. The remaining seven reads were identical over a shortened length of 939 bp that excludes the vicinity of primers at

OPEN ACCESS

Edited by:
Prabakaran Ponraj, Intrexon, United States

Reviewed by:
Lei Xu, National Cancer Institute (NIH), United States
Kwan-Ki Hwang, Duke University, United States

***Correspondence:**
Marie-Paule Lefranc marie-paule.lefranc@igh.cnrs.fr; Gisèle Clofent-Sanchez gisele.clofent-sanchez@rmsb.u-bordeaux2.fr

†Present address:
Melissa Laird Smith, Department of Genetics and Genomic Sciences, Icahn School of Medicine Mount Sinai, New York, United States

‡*Equivalent position of authors.*

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *30 October 2017*
Accepted: *30 November 2017*
Published: *20 December 2017*

Citation:
Hemadou A, Giudicelli V, Smith ML, Lefranc M-P, Duroux P, Kossida S, Heiner C, Hepler NL, Kuijpers J, Groppi A, Korlach J, Mondon P, Ottones F, Jacobin-Valat M-J, Laroche-Traineau J and Clofent-Sanchez G (2017) Pacific Biosciences Sequencing and IMGT/HighV-QUEST Analysis of Full-Length Single Chain Fragment Variable from an In Vivo Selected Phage-Display Combinatorial Library. Front. Immunol. 8:1796. doi: 10.3389/fimmu.2017.01796

both ends. Interestingly these reads were obtained from each of the 15 SMRT cells. Thus, the SMRT sequencing method and the IMGT/HighV-QUEST functionality for scFv provides a straightforward protocol for characterization of full-length scFv from combinatorial phage libraries.

Keywords: human antibody, IMGT/HighV-QUEST, immunoinformatics, immunoglobulin, Pacific Biosciences sequencing, phage combinatorial library, single chain fragment variable, next-generation sequencing

INTRODUCTION

Immunoglobulin (IG) or antibody fragments displayed as single chain Fragment variable (scFv) on filamentous phages (scFv-phages) are classically selected from scFv-phage combinatorial libraries to obtain human antibodies specific for a given target (1–3). This selection from scFv-phage display libraries is widely used for the discovery of novel specificities for therapeutic antibodies in cancer, cardiovascular, autoimmune, infectious or neurodegenerative pathologies, with many of them at various stages of clinical or research development (4–10). Classical *in vitro* phage display approaches involve multiple rounds of selection (or panning) for the enrichment of scFv-phages that demonstrate the desired specificity against a target followed, at the last selection round, by functional screening and characterization of selected candidates using appropriate assays. At this very last step, analysis of the selected scFv *via* Sanger sequencing is commonly used to identify sequences of interest, taking advantage of the genotype–phenotype linkage inherent to the display system. A critical limitation of using biological assays followed by Sanger sequencing is that only a minute fraction of the selected library is actually sampled, a few hundred at best, whereas the selected library may usually contain up to 10^5 to 10^6 variants. This limitation is further enhanced when scFv-phage selection is performed *in vivo* (biopanning) in different pathological models in which scFv-phages can encounter a very large panel of unknown biomarkers (11–13). Currently available next-generation sequencing (NGS) platforms allow the simultaneous sequencing of millions of reads. However, a main challenge for the NGS sequencing of scFv from combinatorial libraries remains the scFv length >800 bp, which is too long for most NGS platforms. Up to now, NGS methods have only provided reads encompassing one variable (V) domain (400 bp), therefore losing a critical piece of information found in scFv sequences, that of the association of two specific V domains [variable heavy (VH) and variable light (VL) for the IG] by the peptide linker. Although a few approaches have been proposed, retrieving information regarding V domain association has still not been solved (14–16).

The analysis of antibody scFv sequences is a difficult exercise because not only are scFv composed of two V domains, but these two V domains are different from each other and each can potentially be extremely diverse. Indeed, the huge diversity of IG or antibodies results from complex *in vivo* mechanisms that occur during the synthesis of the VH and VL domains, which include the molecular rearrangements at the DNA level of the variable (V), diversity (D) (only for VH), and joining (J) genes with nucleotide deletions and insertions (N-diversity) at the V-(D)-J junctions in the bone marrow pre-B and immature B cells (17, 18). In spleen and lymph nodes, somatic hypermutations accumulate in the mature B cell VH and VL, creating a huge diversity of the B cell membrane IG for the recognition of foreign antigens. Following a specific antibody-antigen interaction the B cell proliferates and generates clones engaged in *in vivo* selection and affinity maturation. The specificity of the V domains is conferred by the complementarity determining regions (CDR) and more particularly the CDR3 (19–21). The same features are observed in *in vitro* combinatorial libraries, which mimic the natural *in vivo* diversity, selection and affinity maturation (1–3).

In order to manage, analyze and compare the extraordinary diversity of the immune repertoires, IMGT®, the international ImMunoGeneTics information system®[1] (22), was created in 1989 in Montpellier by Marie-Paule Lefranc (Montpellier University, CNRS), which marked the birth of immunoinformatics (18), a new science at the interface between immunogenetics and bioinformatics. IMGT® has developed online tools that provide a detailed and accurate analysis of the V domains and which, in the case of nucleotide sequences, include IMGT/V-QUEST (23–25) for the analysis of the rearranged V-(D)-J sequences of the IG or antibodies and T cell receptors (TR), and IMGT/JunctionAnalysis (26, 27) for the analysis of the V-(D)-J junctions and of the included CDR3. The algorithms and IMGT reference directories of these tools have been implemented in IMGT/HighV-QUEST (28–31), the first and only web portal for NGS sequence analysis of IG and TR, begun in 2010. IMGT/HighV-QUEST analyses up to 500,000 NGS reads per batch and includes a statistical module for IMGT clonotype identification and comparison (analyses are performed on the results of up to one million reads, from one or several batches) (30). IMGT/StatClonotype (32, 33), a stand-alone tool and R package, allows for the comparison of IMGT clonotype diversity and expression between two NGS data sets, using the IMGT/HighV-QUEST statistical results output. In order to overcome the analysis challenge of the NGS scFv, the IMGT/V-QUEST functionality "Analysis of single chain Fragment variable (scFv)" which includes the search and characterization of two V domains in a single sequence [IMGT/V-QUEST Documentation[2] (34)] has recently been integrated in IMGT/HighV-QUEST (IMGT/HighV-QUEST Documentation[3]).

Abbreviations: CCS, circular consensus sequencing; IG, immunoglobulin; IMGT, IMGT®, the International ImMunoGeneTics Information System®; NGS, next-generation sequencing; scFv, single chain Fragment variable; SMRT, single-molecule real-time; TR, T cell receptor.

[1] http://www.imgt.org.

[2] http://www.imgt.org/IMGT_vquest/share/textes/imgtvquest.html.

[3] http://www.imgt.org/HighV-QUEST/doc.action.

The main challenge addressed in this study was to obtain high quality NGS reads of the scFv long enough to encompass the two VH and VL domains and to analyze and characterize the association of the two V domains in the NGS scFv reads using the scFv functionality in IMGT/HighV-QUEST. Full-length scFv reads are expected to have a length of around 800 bp with a whole insert of around 1,000 bp (including the 5′ and 3′ sequences of the vector and primer sequences at both ends). We used the Pacific Biosciences (PacBio) third-generation NGS technology, which provides long sequencing reads and the highest consensus accuracy available today (35–38). For this project, the high accuracy is a result of the generation of circular consensus sequencing (CCS) reads, by which the long sequencing reads allow for multiple passes of the same insert and thus removal of random sequencing errors upon consensus construction. Practically, one single molecule, real-time (SMRT) cell has 150,000 zero-mode waveguides (ZMWs), of which 50,000–75,000 are loaded with single molecules during sequencing, resulting in the production of ~50,000–75,000 unique consensus reads per run and per SMRT cell. PacBio NGS sequencing was performed from amplified DNA of scFv-phages isolated from the third round of an *in vivo* biopanning in an animal model of atherosclerosis (12, 13). The PacBio scFv CCS (version 2) reads were first analyzed using IMGT/HighV-QUEST with the scFv functionality. In a second step, the sequence quality was evaluated by tracking the NGS reads of a scFv-phage clone P3, identified by biological assays and Sanger sequencing.

MATERIALS AND METHODS

In Vivo Selection of scFv-Phages Specific to Atheroma Plaque

A fully human-recombinant scFv antibody library (scFv cloned in the pMG72 phagemid vector, containing the ampicillin-resistant gene for the selection and maintenance of the phagemid) with a diversity of 3.4×10^9 clones (full description in patent WO2007137616) was expressed by phage display and selected *in vivo*, as previously described (13). The scFv-phages were obtained from the scFv-phagemid combinatorial library by expression of the scFv on the phage surface, following the addition of a helper filamentous phage to recombinant phagemid infected bacteria in the exponential phase (39). Three rounds of biopanning were performed in atheromatous injured rabbits (12). All animal experiments were performed in conformity with the Guide for the Care and Use of Laboratory Animals (NIH Publication No. 85–23, revised 1996) and were accredited by the local ethical committee (Animal Care and Use Committee of Bordeaux, France under the No. 50120192). Briefly, the procedure was the following (**Figure 1**): 2.4×10^{12} colony-forming units (cfu) of scFv-phages were injected into an atheromatous rabbit. After 1 h in circulation, the animal was sacrificed, the aorta was retrieved and scFv-phages binding to the aorta were eluted in different fractions. The eluted scFv-phages were reamplified in XL1-Blue bacteria and following scFv expression at the phage surface as above (39), the amplified

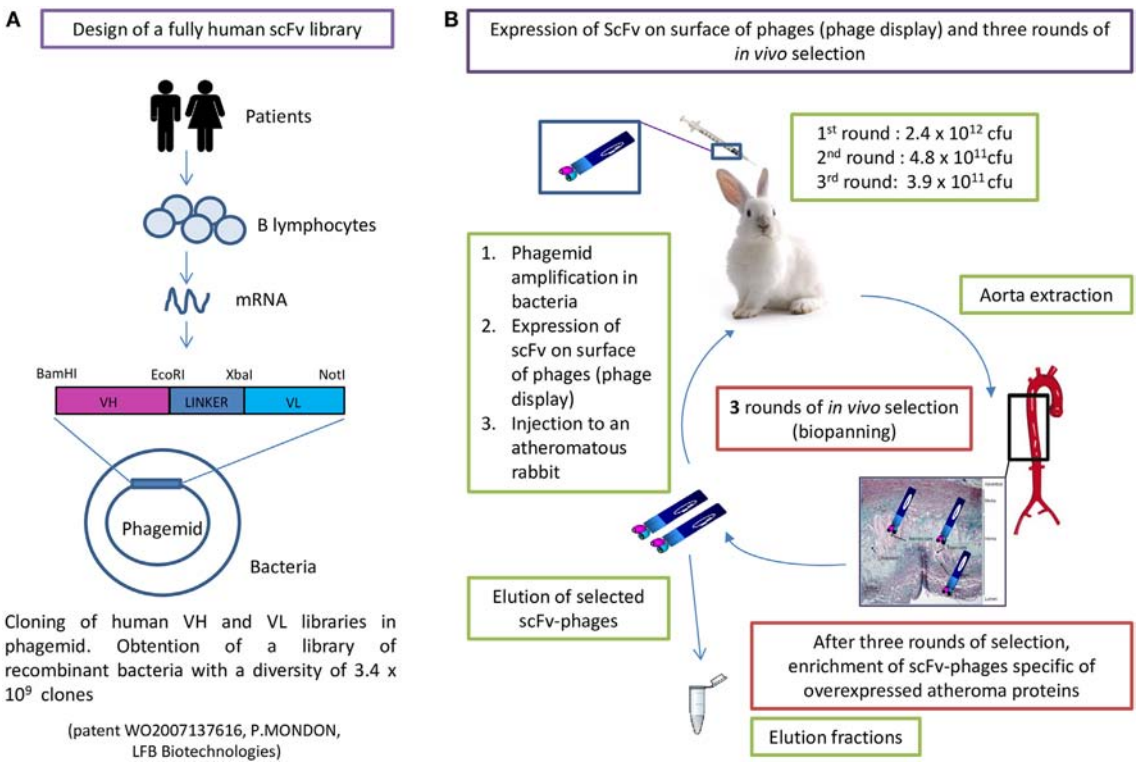


FIGURE 1 | Human single chain fragment variable-phagemid combinatorial library construction **(A)** and *in vivo* phage display selection **(B)**.

scFv-phages were reinjected in another atheromatous animal. Rounds 2 and 3 were conducted following the same procedure. The number of reinjected colony-forming units were 4.8×10^{11} in round 2 and 3.9×10^{11} in round 3. The total number of eluted scFv-phages from the third round was 1.5×10^{7} cfu (total for seven fractions corresponding to different areas of the aorta).

In this study, the analyzed fraction is the one recovered after the third round of selection from the endothelial cells of the damaged abdominal aorta vessel wall (named AAR3 for abdominal aorta round 3) (**Figure 2**). The scFv-phages were amplified in XL1-Blue bacteria and plated on 145 mm Petri dishes for storage of the whole AAR3 fraction before sequencing, and on 80 mm Petri dishes for limiting dilution (quantification of recombinant bacteria and picking of individual clones) (**Figure 2**). The recombinant bacteria plated on 145 mm Petri dishes were scratched and stored at −80°C in 50% v/v glycerol. Around 3.5×10^{5} clones issuing from the whole AAR3 fraction were counted by limiting dilution of recombinant bacteria on 80 mm Petri dishes. Ninety-six recombinant bacteria clones were individually picked and grown in selective medium on a 96-well MASTERBLOCK® polypropylene storage plate (Greiner Bio-One, France) and stored for *in vitro* bioreactivity assays (*in vitro* screening of scFv-phages on atheromatous proteins) and for Sanger sequencing.

Each of the 96 individual recombinant bacteria clones from the MASTERBLOCK® was repicked on an agarose 96-Well plate (Beckman Coulter, France). After incubation at 37°C overnight, plasmid DNA extraction, PCR amplification and sequencing were then performed by Beckman Coulter (France). One of these sequenced clones, P3 (767 bp), was used in this study for tracking the PacBio reads that were identical or related to it.

PacBio RS II Sequencing of the Whole AAR3 Fraction

Generating High Quality PCR Products

Single-molecule real-time sequencing requires high-quality, doubled-stranded DNA as input. To ensure this, plagemid DNA of recombinant bacteria from the whole AAR3 fraction was extracted directly from the frozen extract just before PCR amplification, using the QIAprep spin miniprep kit (Qiagen, France) according to manufacturer's instructions. To generate clean, undamaged and non-chimeric amplicons, the highest fidelity polymerase was used. All PCR reactions were performed in volumes of 50 µL using 25 µL of the KAPA HiFi™ HotStart ready Mix (Kapa Biosystems, France) and 20 ng of DNA template. Each primer was used at a final concentration of 0.3 µM. PCR reactions were performed with the forward primer (Primer 1, 23-mer FWD) 5′-TGCAAATTCTATTTCAAGGAGAC-3′ and the reverse primer (Primer 2, 20-mer REV) 5′-TCACGTG CAAAAGCAGCGGC-3′. These primers were designed based on the phagemid vector; for primer 1 from position −96 to −74 upstream of the BamHI site and for primer 2 from positions 95 to 114 downstream of the NotI site (**Figure 3A**).

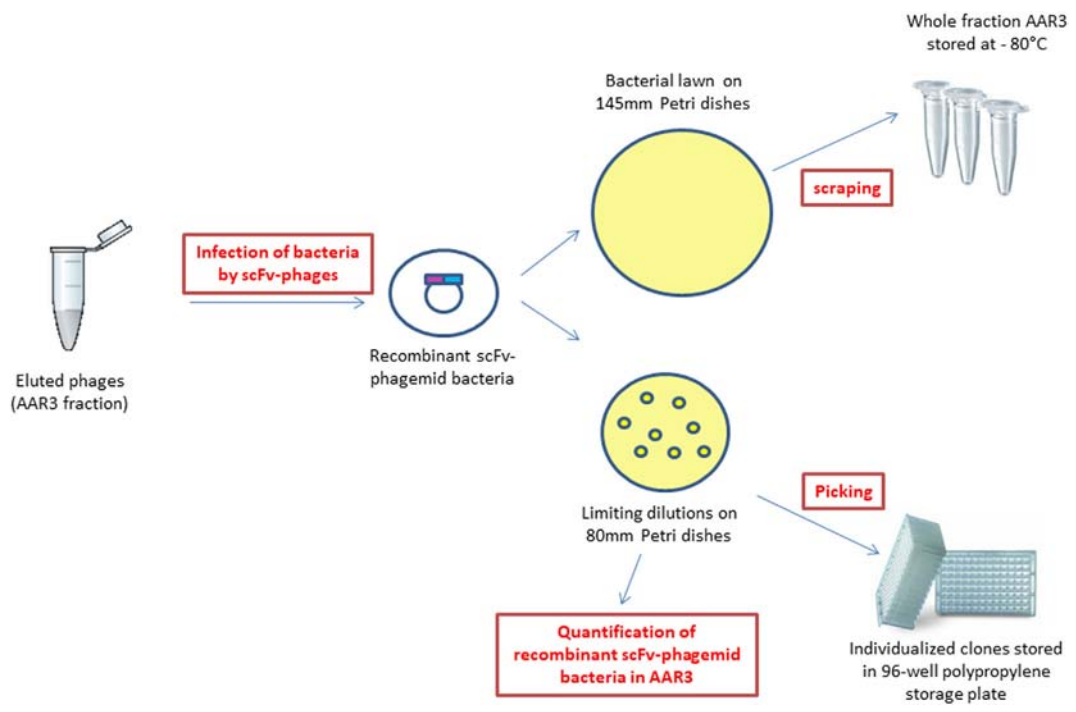


FIGURE 2 | Obtaining recombinant single chain fragment variable-phagemid bacteria from the AAR3 fraction, storage for next-generation sequencing, quantification, and picking of individual clones.

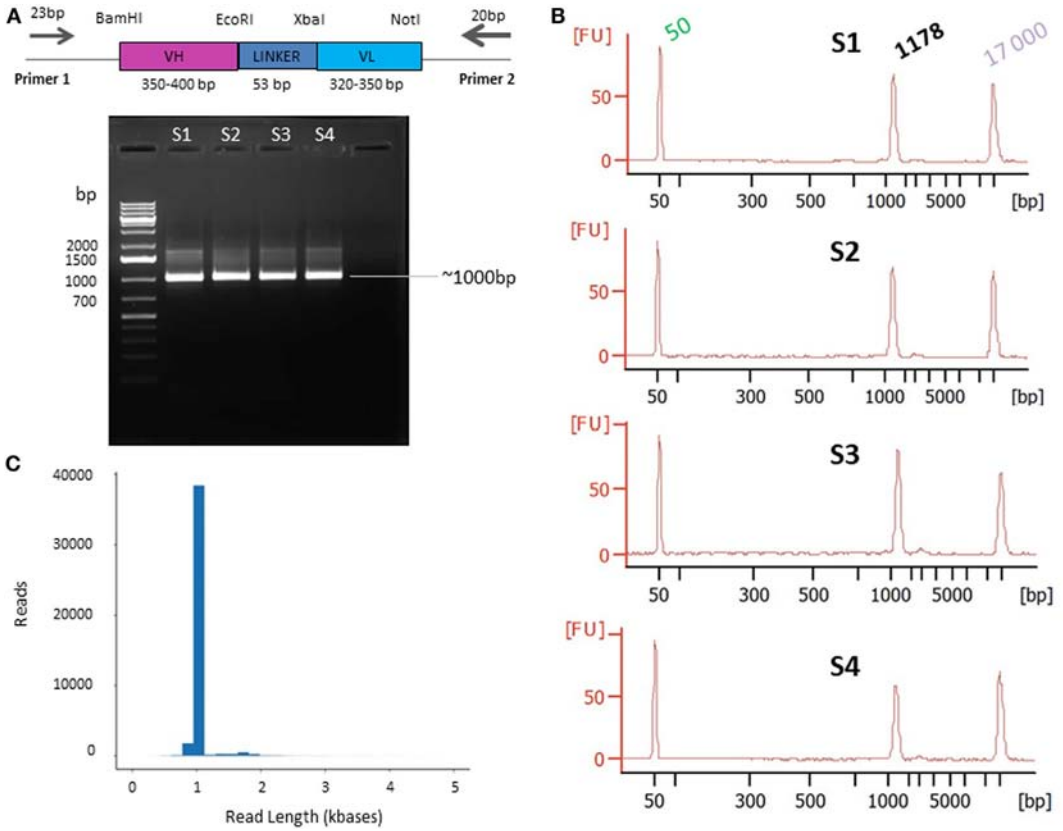


FIGURE 3 | Primer design **(A)** and quality control presequencing **(B)** and postsequencing **(C)**. **(A)** Primers designed on the phagemid vector and used for single chain fragment variable (scFv) PCR amplification. The scFv (VH-LINKER-VL) length range is between ~720 and ~800 bp [variable heavy (VH) between ~350 and ~400 bp and variable light (VL) between ~320 and ~350 bp]. The linker is 53 bp including the EcoRI and Xbal sites. The PCR products are expected to be ~1,000 bp on average, including the 5′ and 3′ region and the primers. **(B)** Agarose gel electrophoresis of PCR products. The DNA was amplified from the AAR3 fraction and PCR products were analyzed on 1.2% (w/v) agarose gel. The band at ~1,000 bp corresponds to the expected size for scFv amplicons. S1, S2, S3, and S4 correspond to the samples 1, 2, 3, and 4, respectively. The Bioanalyzer trace of the four samples shows the purity of amplicons with a high-quality single peak. **(C)** Pacific Biosciences RS II CCS2 read length distribution using P6-C4 chemistry for 1 SMRT cell (similar results were obtained for the 15 SMRT cells). Data are based on a 1-kb size-selected scFv library using a 6 h movie.

The PCR cycling protocol was chosen according to the manufacturer's instructions and consisted of 95°C for 3 min; 15 cycles of 98°C 20 s, 65°C 15 s, 72°C 30 s; one 72°C 1 min followed by one cycle at 4°C. The reduction of the number of PCR cycles minimizes the PCR bias, however when further decreases in the number of PCR cycles were attempted, it led to several contaminating, off-target bands on an agarose gel.

The required quantity in PacBio guidelines for insert sizes superior to 750 bp is 500 ng of DNA. Four PCR reactions were performed on the same AAR3 fraction for reproducibility purposes. Agarose gel electrophoresis was used to confirm amplification, correct fragment size, and to check for non-specific product contamination. A sizing marker was included to confirm size specificity (GeneRuler 1 kb Plus DNA Ladder, Thermo Fisher Scientific, France). Amplicons were cleaned using 1× ratio of AMPure PB Beads (Pacific Biosciences). DNA purity and quantification (sample volume, yields and size distributions) were evaluated and measured using the Agilent 2100 Bioanalyzer DNA12000 kit (Agilent technologies).

SMRTbell Template Preparation

SMRTbell templates (PacBio, CA, USA) were constructed following the standard Amplicon Sequencing Protocol.[4] The full procedure is explained in Figure S1 in Supplementary Material.

Briefly, PCR products were treated to repair DNA damage and then hairpin adapters were added *via* blunt end ligation to produce SMRTbell templates using the SMRTbell template prep kit 1.0. Exonucleases III and VII were used to remove failed ligation products and SMRTbell templates were purified with AMPure PB Beads. The ratio of sequencing primer and polymerase was determined by a PacBio calculator to correlate with SMRTbell concentrations and the 1,100-bp insert size. The sequencing primer was annealed to the single-stranded loop region of the SMRTbell template, and primer-annealed templates were then bound to DNA polymerase P6 using the DNA/polymerase binding kit P6v2. The DNA–polymerase complexes were loaded on

[4]http://www.pacb.com/wp-content/uploads/Procedure-Checklist-Amplicon-Template-Preparation-and-Sequencing.pdf.

15 SMRT cells using MagBeads onto the PacBio RS II system and sequenced using the C4 chemistry and 6-h movies.

Initial loading titrations were performed to identify the optimal loading concentration, identifying 0.03–0.035 nM as the best loading conditions for scFv PCR products. Each SMRT cell generates ~50,000 reads on average (considering that 50,000–75,000 ZMW can be optimally loaded with a single molecule of DNA). A total of 15 SMRT cells were loaded for the four different PCR products of the same AAR3 fraction (three SMRT for sample 1 and four SMRT for samples 2, 3, and 4) to cover, per PCR sample, the diversity of the AAR3 fraction (Figure S2 in Supplementary Material). The use of 15 SMRT cells was chosen to provide a thorough and sound proof-of-concept and read comparison from four different PCR amplicons generated from the same fraction (AAR3).

PacBio CCS Read Generation

The PacBio RSII instrument produces sequencing reads with an average read length of ~15 kb, which would theoretically pass over a 1,100 bp molecule more than 10 times, producing CCS2 sequences with an accuracy ~99.9%. Longer read lengths can be achieved by increasing the instrument run time, so these data were collected using 15 SMRT cells and 6 h movies, to maximize read length and, thus, number of passes. Following sequencing, the raw data were processed using the CCS2 pipeline (CCS code used available at https://github.com/PacificBiosciences/unanimity). All CCS2 reads that were 99.9% accurate or greater were exported for further analysis. The raw NGS data of the 85-related P3 sequences can be found in the NCBI Sequence Read Archive with the accession number SRP124616.

IMGT/HighV-QUEST Analysis of the PacBio CCS2 Reads

Characterizing IMGT/HighV-QUEST scFv Reads from AAR3

The FASTQ files of the PacBio CCS2 reads were imported and converted to FASTA sequences for submission to IMGT/HighV-QUEST[5] (29, 30), which implements IMGT/V-QUEST program version 3.4.2 (August 4, 2016) and IMGT/V-QUEST reference directory release 201631-4. The analysis was performed with the advanced functionality "Analysis of single chain Fragment variable (scFv)" (IMGT/HighV-QUEST Documentation, see text footnote 3).

Data filtering was applied with the following criteria to be fulfilled for each of the two V domains: (i) >85% of identity of the V-REGION of the V domain with the V-REGION of the closest germline IMGT gene and allele and (ii) in-frame V-(D)-J junction. Filtered sequences were then analyzed to identify the closest V, D (for VH) and J IMGT genes and alleles, to characterize the amino acid (AA) junction, to evaluate the mutations and to give a complete description of the scFv with IMGT labels (IMGT Index > scFv[6]).

[5]http://www.imgt.org/HighV-QUEST/login.action.

[6]http://www.imgt.org/IMGTindex/scFv.php.

Tracking and Analysis of Identical and Related PacBio Reads of the AAR3 scFv-Phage Clone P3

In order to evaluate the PacBio scFv read sequencing quality, reads identical or closely related to the sequence of a scFv-phage clone P3 (previously isolated from the same fraction AAR3 and Sanger sequenced) were tracked among the total reads generated in the 15 SMRT cells by two approaches. First, reads potentially related to P3 were searched for based on the expected VH-VL characteristics determined by IMGT/HighV-QUEST (same V and J genes and alleles, and same AA junctions). Second, a Blast search was performed to check whether P3-related reads could have been missed by the IMGT/HighV-QUEST filtering. The fasta headers of the sequences were modified to include the sequence set identifier and a blast database was built (formatdb 2.2.26) from the accumulated reads from the 15 SMRT cells. The database search was performed using the blastn program (2.2.26), with the combined following criteria: longest alignments on the P3 Sanger (767 bp) sequence, highest identity percentage, and maximum number of 20 mismatches or indels. Among the extracted reads only those fulfilling the P3 characteristics (in terms of V and J genes, allele names and AA junctions) were retained.

RESULTS

Generation and Analysis of PCR Products from the Whole AAR3 Fraction

The unbiased characterization of scFv from phage-display combinatorial libraries, in conjunction with sequencing on the PacBio RS II system, requires high-quality PCR products with undamaged, clean and non-chimeric amplicons. Creating PCR protocols to generate products that are truly representative of the starting cell population is a major challenge.

To achieve this aim: (1) the complete frozen AAR3 fraction was directly used, without amplification, as the source of recombinant bacteria; (2) efforts were made to limit the number of amplification cycles (the four samples subject to 15 cycles) in order to reduce quantitative distortions as well as error rates due to PCR artifact (data not shown); and (3) different high-fidelity Taq polymerases were tested so as to fit with our amplification system. The required quantity (500 ng) and the correct size of amplicons (~1,000 bp) were obtained with the KAPA HiFi™ HotStart polymerase (**Figure 3A**). The quality and quantity of amplicons was confirmed using an Agilent 2100 Bioanalyzer. After confirmation of the purity (**Figure 3B**), PCR samples were sequenced on the PacBio platform.

PacBio Sequencing and CCS2 Analysis

Sequencing was done on the PacBio RS II system using P6-C4 sequencing chemistry and SMRTbell libraries generated from DNA amplified from the AAR3 fraction (corresponding to 3.5×10^5 scFv-phagemids). For each of the four PCR samples, 3 or 4 SMRT cells were run to ensure adequate sampling of the scFv as described in Section "Materials and Methods" (Figure S2 in Supplementary Material). A total of 15 SMRT cells were used.

Circular consensus sequencing (CCS2) analysis of the 15 SMRT cells produced 450,558 reads. These reads were filtered to remove any double-loaded wells or other artifactual/lower

accuracy reads. The data obtained from filtered, post-CCS2 analysis represent the reads achieving 99.9% accuracy or above, derived only from wells loaded with a single molecule. Any reads that did not reach consensus coverage of QV30 (99.9% accurate) or above were filtered out. Those settings and filters are built into the CCS2 pipeline that is available *via* the PacBio web-based analysis software (SMRT Link). Thus, these 450,558 reads passed initial filtering with an average pass number of 24 and a quality score of minimum 99.94% accuracy (**Figure 3C**). Another contributing factor to data quality and throughput was the use of the longest movie lengths possible at the time (6 h) to ensure the longest read lengths for analysis.

IMGT/HighV-QUEST Analysis of the scFv PacBio Reads

The 450,558 FASTQ PacBio CCS2 reads were converted in FASTA sequences and analyzed using IMGT/HighV-QUEST, as described in Section "Materials and Methods." After analysis, a total of 391,655 scFv "candidates" (i.e., sequences with two V regions, IMGT label V-REGION) (86.93% of the submitted PacBio reads) were identified (**Table 1**). The scFv sequences were then filtered according to the criteria described in Section "Materials and Methods" [>85% of identity of the V-REGION of the V domains with the V-REGION of the closest germline IMGT genes and alleles and in-frame V-(D)-J junction, determined for both V domains]. The threshold of 85% of identity is the standard filter for classical IG repertoire analysis in IMGT/HighV-QUEST (29). Following this filtering, 348,659 full-length scFv representing 89.02% of the filtered sequences were identified (**Table 1**). These scFv reads include 346,934 VH-VL or VL-VH expected scFv sequences (Table S1 in Supplementary Material). The other 1,725 scFv reads comprise 171 VH-VH (5–22 found per SMRT cell) and 1,554 VL-VL (68–158 found per SMRT cell). These combinations most probably occurred during the construction of the original scFv-phage combinatorial library. Thus these results provide a useful and detailed overview of the content of the scFv combinatorial library.

The IMGT/HighV-QUEST analysis of the scFv reads included identification of the closest V, D (if VH), and J IMGT genes and alleles, characterization of the junction, evaluation of the mutations, and complete description of both V domains with IMGT labels (see text footnotes 2 and 5). These results demonstrate that both domains of the scFv reads sequenced by PacBio could be fully characterized with the functionality for scFv.

Tracking and Analysis of PacBio Reads Identical or Related to the P3 Clone across 15 SMRT Cells

Single chain fragment variable-phages issuing from the *in vivo* AAR3 selected fraction have been screened by a high-throughput flow cytometry method against atherosclerotic rabbit proteins. Some of the selected clones were then processed into scFv-Fc format in HEK293 cells and validated by immunohistochemistry on atheromatous aorta sections of two animal models of atherosclerosis (ApoE$^{-/-}$ mouse and New Zealand White rabbit submitted to hypercholesterolemic diet) and on human endarteriectomy

TABLE 1 | IMGT/HighV-QUEST analysis of the scFv Pacific Biosciences (PacBio) reads.

PCR sample no.	Number of reads with 99.9% predicted accuracy	Mean number of passes	Number of movies	SMRT cell no.	Number of PacBio CCS2 analyzed reads	Number of scFv candidates in analyzed reads	% of scFv candidates in analyzed reads, i.e., coverage	Number of filtered-in reads	% of scFv in filtered-in reads, i.e., coverage	% of scFv in analyzed reads, i.e., coverage
s1	91,828	24	3	1	29,224	25,419	86.98	22,906	90.11	78.38
				2	32,240	28,120	87.22	25,228	89.72	78.25
				3	30,364	26,496	87.26	23,799	89.82	78.38
s2	129,640	23	4	4	34,082	29,729	87.23	26,657	89.67	78.21
				5	33,510	29,213	87.18	26,407	90.39	78.80
				6	31,980	27,874	87.16	25,032	89.80	78.27
				7	30,068	26,289	87.43	23,695	90.13	78.80
s3	115,446	24	4	8	34,890	30,183	86.51	26,990	89.42	77.36
				9	29,373	25,314	86.18	22,468	88.76	76.49
				10	26,465	22,776	86.06	20,044	88.00	75.74
				11	24,718	21,358	86.41	18,741	87.75	75.82
s4	113,644	24	4	12	25,128	21,881	87.08	19,120	87.38	76.09
				13	23,693	20,515	86.59	17,756	86.55	74.94
				14	32,293	28,093	86.99	24,762	88.14	76.68
				15	32,530	28,395	87.29	25,054	88.23	77.02
Total	450,558				450,558	391,655	86.93	348,659	89.02	77.38

Each CCS read counts as "1× coverage" over the scFv molecules of interest. The coverage is given in percentage of the scFv of interest (analyzed reads, filtered-in reads and analyzed reads).

biopsies. A scFv-phage clone, named P3, selected for its high accuracy in biological assays in the two animal models and human biopsies was Sanger sequenced and patented (EP 17306337.1). The whole scFv P3 sequence (Figure S3 in Supplementary Material) was then tracked among the 346,934 VH-VL and VL-VH reads (Table S1 in Supplementary Material) from the 15 SMRT cells, by the two approaches described in Section "Materials and Methods," in order to evaluate the PacBio scFv read sequencing quality.

Analysis of PacBio Reads Identical to the P3 Clone

Sixty PacBio reads with 100% identity to the region aligned with the P3 Sanger sequence (767 bp) were obtained by both approaches. Interestingly, the P3 scFv was identified in the top 100 of the most abundant VH-VL associations in all the 15 data sets. These PacBio reads have a length of 977 bp, including the 5′ and 3′ regions and the primers. Fifty-three of the 60 reads have a 100% identity on their full length of 977 bp. As these PacBio reads were obtained from the 15 SMRT cells (**Table 2**), we can confidently say that no sequencing error was observed in the 51,781 bp of these 53 reads.

Seven reads showed 100% identity on a 939 bp length (**Table 2**). Interestingly their mutations are all localized at the ends of the primers and/or in the immediate vicinity of the 3′ primer (**Table 3**). A most probable explanation is that they occur during the sequencing polymerization (the priming step could be excluded as different mutations were observed in different SMRT cells and no degenerate bases were used in the primers). Ignoring the mutations in or next to the primers, no sequencing error was observed in 6,573 nucleotides of the 7 reads (100% identity on 939 bp). Combining these results with those of the 53 reads (100% identity on the full length of 977 bp), no sequencing error was detected on 58,354 nucleotides for the 60 scFv reads.

Analysis of PacBio CCS Reads Related to the P3 Clone

In addition to the 60 PacBio reads with a 100% identity (53 reads on 977 bp and seven reads on 939 bp) (**Table 2**), 25 "related" P3

TABLE 2 | Pacific Biosciences (PacBio) reads 100% identical to the aligned P3 Sanger sequence and 100% identical between them on 977 bp (53 reads) or 939 bp (7 reads).

PCR sample no.	Number of P3 PacBio reads per PCR sample	SMRT cell no.	Number of P3 PacBio reads per SMRT cell	100% on 977 bp (53 reads)	100% on 939 bp (7 reads)	GenBank/ENA/DDBJ accession number
s1	15	1	6	1, 2, 3, 4, 6, 8		MG272208
		2	5	10, 30	11, 12, 13	
		3	4	16, 17, 19	15	
s2	14	4	2	32, 33		
		5	2	36	35	
		6	3	37, 38, 39		
		7	7	20, 41, 44, 45, 46, 47, 48		
s3	15	8	6	49, 50, 51, 52, 53, 54		
		9	4	56, 57, 59, 60		
		10	4	63, 65, 66	64	
		11	1	68		
s4	16	12	4	70, 72, 73	71	
		13	2	75, 76		
		14	3	21, 80, 83		
		15	7	23, 24, 25, 26, 27, 28, 29		
Total	60		60	53	7	

TABLE 3 | Mutations observed at the 5′ and 3′ end of the seven Pacific Biosciences (PacBio) reads with 100% identity on 939 bp (positions 3–941).

PacBio read no. (assigned in the list 1–85)	PCR sample no.	SMRT cell no.	Mutation description[a]	Mutation localization	GenBank/ENA/DDBJ accession number
13	s1	2	One 1 nt-deletion (g2 > del)	5′ end of the 5′ primer	MG272209
11	s1	2	Two 1 nt-deletion (t975 > del, a977 > del)	3′ end of the 3′ primer	MG272210
15	s1	3	One 1 nt-substitution (c956 > t)	Vicinity of the 3′ primer	MG272211
12	s1	2	One 1 nt-deletion (a942 > del)[b]	Vicinity of the 3′ primer[b]	MG272212
35	s2	5			
71	s4	12			
64	s3	10	Two 1 nt-deletion (a942 > del),[b] (t975 > del)	Vicinity of the 3′ primer,[b] 3′ end of the 3′ primer	MG272213

Positions of the primers are 1–23 and 958–977.

[a]*Mutations are described according to the IMGT Scientific chart rules (http://www.imgt.org/IMGTScientificChart/Nomenclature/IMGTmutation.html) (40).*

[b]*The 1 nt-deletion (a942 > del) found in reads from the four samples most probably originates from the library.*

reads were identified (Figure S4 in Supplementary Material). These 25 P3-related reads showed 15 different mutation types in the insert and, for each type, a limited number of reads (1–6) (**Table 4**). Altogether, the 15 mutation types are described by 28 different individual mutations. This heterogeneity is in sharp contrast with the 60 reads with a 100% identical insert.

The observation of related reads was expected as the scFv are from a phagemid combinatorial library in which point mutations were initially introduced to mimic the IG somatic hypermutations. The analysis of the 25 reads was therefore performed in an attempt to distinguish among them related clones with potential biological interest (as reflecting the diversity of the original library) from reads with artifactual differences.

For 15 of these PacBio CCS reads, the origin of the mutations (mostly substitutions) was clearly from the scFv phagemid combinatorial library (and therefore of potential interest given their relatedness to P3). This was supported by the fact that identical mutated reads were obtained from different PCR and from different SMRT cells (category A in **Table 4**). These included 7 reads (highlighted in pink) with four identical individual mutations, 2 reads (highlighted in green) with three identical individual mutations different from those described above and one shared by both groups [c838 > g (VL)], 6 reads (highlighted in blue) with two identical individual mutations (and still different from those of the previous mutation types). All the substitution mutations are localized in the VL, in agreement with mutations of the VL and VH domains being targeted differently before the assembly into the scFv and confirming that the differences observed are intrinsic to the VL domain, and not due to PCR or sequencing errors. The IMGT/V-QUEST alignment of the 15 PacBio NGS sequences with the initial P3 Sanger sequence is provided as an additional PDF file in supplementary data (Figure S5 in Supplementary Material).

For the other 10 PacBio reads, present in one copy (category B in **Table 4**), the origin of the differences could not be determined with certainty: and any of the possible explanations: putative sequencing errors, PCR amplification errors or diversity of the combinatorial library could not be formally excluded at this stage. The 1 nt-deletion observed in VH and VL for reads 85 and 40, respectively, the large nt deletion at the 3′ end in read 78, the nt deletion at both 5′ and 3′ ends for read 81 and the one or two nt insertions in VH or VL for reads 84 and 9 could be considered as sequencing errors. The six reads with different single substitution and the one with 3 mutations could be attributed to PCR amplification errors or diversity of the combinatorial library. Moreover, and even if these differences are due to diversity of the combinatorial library, their single copy number suggests that they are from unselected (or poorly selected) scFv-phages from the library background and could be ignored in the library screening.

DISCUSSION

Antibody libraries are important resources to derive antibodies to be used for a wide range of diagnostic and therapeutic applications. *In vivo* or *ex vivo* phage-display selections have emerged as interesting ways to identify accurate antibodies in the context of the pathologic microenvironment (11). Although advancements in automation of biological assays have greatly improved screening strategies, high-throughput campaigns are still severely limited in the number of antibody fragments that can be interrogated, providing only a tiny fraction of the enriched phage library. Access to the genetic information encoded in antibody repertoires by NGS should allow more in depth analysis of the diversity of the selected library. However, the currently available NGS platforms that were capable of providing several million reads per run, generated only short reads of up to 300–700 bp (41). Therefore, only synthetic combinatorial scFv libraries, in which diversity was confined to CDR3 regions of the heavy and light chains, have benefited from NGS potential in the extensive *in silico* analysis of complex collections of selected antibodies (16).

While Mi-Seq, 454, or Ion-Torrent technologies will completely sequence heavy and light variable domains, they are currently insufficient to cover the full-length scFv, which are comprised of both VH and VL domains, connected by a peptide linker. These technologies often necessitate consensus building of multiple reads originating from the scFv fragment to obtain whole sequence information (42–44). Therefore, NGS sequencing of scFv fragments longer than 800 bp is still hampered by technical limitations in the length of reads. These limitations could be circumvented by the third generation PacBio sequencing platform. Capitalizing on PacBio SMRT DNA sequencing for high-resolution and high-throughput HLA typing (36, 37), we develop here a PacBio SMRT/CCS2 approach, combined with IMGT/HighV-QUEST analysis of full-length scFv reads to provide a straightforward protocol for characterization of the complete VH and VL domains of scFv from fully human combinatorial libraries.

Pacific Biosciences SMRT sequencing generated 450,558 reads of about 1,000 bp across 15 SMRT cells, following DNA amplification of the scFv insert from 3.5×10^5 *in vivo* selected scFv-phages. IMGT/HighV-QUEST with its scFv functionality allowed for further filtration and characterization of the 348,659 PacBio CCS reads as containing full-length scFv, which represent 89.02% of the overall filtered sequences from the 15 SMRT cells run. Among them, 346,934 identified expected VH-VL or VL-VH full-length scFv reads.

In order to evaluate the PacBio scFv read sequencing quality, a selected scFv-phage clone P3, previously identified from the AAR3 fraction by biological screenings and sequenced by the Sanger methodology, was tracked within the 346,934 VH-VL and VL-VH scFv reads characterized by IMGT/HighV-QUEST. Sixty PacBio CCS2 reads were identified from the 15 SMRT cells that demonstrated 100% identity on a length of 939 bp (including the complete scFv of 767 bp and the 5′ and 3′ regions) and for 53 of them on the full length of 977 bp (including the primers at both ends). Thus no sequencing error was observed on a total of 51,728 bp for these 53 reads, obtained from the 15 SMRT cells (and 58,354 bp if the seven reads with 100% on 939 bp are included), validating the capacity of the PacBio SMRT-CCS2 method to produce reads with an accuracy of >99.9% for complete sequences of inserts approaching 1,000 bp. It is of interest to note that these PacBio reads were obtained from the 15 SMRT cells.

In addition to the 60 PacBio reads with a 100% identity, 25 "related" P3 reads were identified and for 15 of them, mutation

TABLE 4 | Mutation heterogeneity observed in 25 P3-related PacBio reads, in contrast with the 60 P3 identical PacBio reads with 100% identity on the complete scFv.

Read categories	Pacific Biosciences (PacBio) read no. (assigned in the list 1–85)	PCR sample no.	SMRT cell No.	Number of reads/ mutation type	Mutation type	Mutation description[a]	GenBank/ENA/ DDBJ accession number
A (15 reads)	7, 18, 43, 67, 69, 79	1, 2, 3, 4	1, 3, 7, 11, 12, 14	6	Four 1 nt-substitution	a545 > g (VL), g686 > a (VL), a757 > g (VL), c838 > g (VL)	MG272218
	58	3	9	1	Four 1 nt-substitution with, in 3′, a large deletion	a545 > g (VL), g686 > a (VL), a757 > g (VL), c838 > g (VL), a886-a977 > del (92 nt)	MG272219
	61	3	9	2	Four 1 nt-substitution	c741 > t (VL), g837 > a (VL), c838 > g (VL), g843 > t (VL)	MG272220
	74	4	13		Four 1 nt-substitution *with, at the 3′ end of the 3′ primer, a 1 nt- deletion*	c741 > t (VL), g837 > a (VL), c838 > g (VL), g843 > t (VL), *a977 > del*	MG272221
	5, 14, 31, 34, 42, 82	1, 2, 4	1, 2, 4, 7, 14	6	Two 1 nt-substitution	c720 > t (VL), t744 > c (VL)	MG272223
B (10 reads)	85	2	4	2	One 1 nt-deletion	c242 > del (VH)	MG272216
	40	2	6		One 1 nt-deletion	g600 > del (VL)	MG272217
	77	4	13	6	One 1 nt-substitution	g495 > a (linker)	MG272227
	22	4	14		One 1 nt-substitution	t624 > c (VL)	MG272224
	55	3	9		One 1 nt-substitution	a627 > g (VL)	MG272225
	62	3	10		One 1 nt-substitution	g736 > a (VL)	MG272226
	78	4	13		One 1 nt-substitution *with, at the 3′ end of the 3′ primer, a 3 nt-deletion*	t599 > g (VL), *t975-a977 > del (3nt)*	MG272228
	81	4	14		One 1 nt-substitution *with, at the 5′ end of the 5′ primer, a 1 nt-deletion, and at the 3′ end, a 19-nt primer deletion*	*g2 > del*, a715 > g (VL), *c959-a977 > del (19 nt)*	MG272222
	84	2	5	1	Two 1 nt-insertion	209^210 > ins^a (VH), 762^763 > ins^t (VL)	MG272214
	9	1	2	1	One 1 nt-substitution in VH, one 2 nt-insertion + two 1 nt-substitution in VL	c322 > t (VH), 658^659 > ins^cc (VL), c659 > t (VL), t660 > a (VL)	MG272215
Total: 25				Total: 25			

Positions of the primers are 1–23 and 958–977. Category A: 15 P3-related reads of potential biological interest (mutations due to the VL diversity originating from the combinatorial library). Pink, green, and blue colors highlight groups of reads with in common identical substitution mutations. Category B: 10 P3-related reads with undefined origin of the mutations.

[a]*Mutations are described according to the IMGT Scientific chart rules (http://www.imgt.org/IMGTScientificChart/Nomenclature/IMGTmutation.html) (40). The mutations in the primers are shown in italics.*

analysis (category A in **Table 4**) clearly showed that they are related to P3. This was supported by the fact that identical mutated reads in the VL (in agreement with the diversity generated during the library construction) were obtained from different PCR samples and from different SMRT cells. P3 (60 reads) with its 15 related reads are among the top 100 most represented associated VH-VL domains in all the 15 data sets.

This study provides the first proof-of-concept that similar sequences could be tracked in phage-display selected scFv samples and their frequency determined by the number of reads. The favorite candidates chosen for their high frequency of enrichment could be rescued from these *in silico* data for implementation of downstream biological assays. This could be easily done by custom gene chemical synthesis, which offers the utmost flexibility and efficiency with high production yields (45).

Thus the large number of sequenced reads delivered, following the enrichment process, could be ideally suited for a more extensive evaluation of antibody candidates by biological assays. In that way, bringing full sequence data from NGS will accelerate search for identification of both the antibodies and their targeted biomarkers. We thus aimed to combine the sensitivity of the sequencing approach with the functional information provided by the immune assays. This *in silico* approach could be applied to any other pathology and phage-display screening methodology. Resolving the issue of complete scFv sequencing has the potential to profoundly impact the selection process of antibodies with desired properties after phage display biopanning with a special focus of *in vivo* selections. It is expected that this will contribute to therapeutic antibody discovery, selection and development.

Limitations

Scientists working on IGs are very concerned by the problems of sequencing errors versus mutations and of short read assembly. This study demonstrates that the PacBio SMRT-CCS2 method is able to produce reads with an accuracy of >99.9% for complete sequences of scFv inserts without the need of *in silico* VH and VL assembly as usually necessary using MiSeq or Ion Torrent technologies (43, 44). To assess the read quality of PacBio sequencing, we performed a pilot study using different PCR samples and 3 or 4 SMRT for each PCR, starting from the same AAR3 fraction. Using the biologically validated P3 clone as a reference, we demonstrated that P3 clone is among the top 100 *in vivo* selected clones with a representativity of 0.025%. Moreover, among the 85 P3 related sequences, 60 were 100% identical and 15 clearly originated from the scFv phagemid combinatorial library. From the IMGT/V-QUEST alignment of P3 pink, blue, and green groups of reads with the initial P3 Sanger sequence, we can confidently say that 88% of reads were free of sequencing errors. However, a doubt remains for 10 of them, present in just one copy. The nt-deletions or nt insertions observed in VH and VL for six of the 10 reads could be definitively considered as sequencing errors. For the single substitutions observed in seven reads, it is obviously impossible to determine their origin. To circumvent this limitation, other technologies could be considered, such as inverse PCR method based on VH CDR3 or VL CDR3 sequences (44) to synthesize these clones from the AAR3 fraction. Nonetheless, what is of upmost importance in our study is to identify over-represented clones (from the top 100 candidates) and to proceed to the rescue of highly enriched scFv and not isolated clones. Indeed, during phage-display selections, the reads of greatest interest will have the greatest depth of coverage, having expanded in the pool and thus receiving greater proportional read depth.

ETHICS STATEMENT

All animal experiments were performed in conformity with the Guide for the Care and Use of Laboratory Animals (NIH Publication No. 85–23, revised 1996) and were accredited by the local ethical committee (Animal Care and Use Committee of Bordeaux, France under the No. 50120192).

AUTHOR CONTRIBUTIONS

AH, FO, M-JJ-V, JL-T, MS, SK, M-PL, and GC-S conceived and designed the experiments. PM provided the library. MS sequenced the data. NH processed the raw data. VG designed the analysis algorithms and PD implemented the tool for NGS. CH, JKuijpers, AG, and JKorlach supervised the sequencing procedure. GC-S and M-PL supervised the project. AH, MS, GC-S and M-PL wrote the article. All the authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank Franck Salin for his helpful advice on DNA amplification and analysis using the Agilent 2100 Bioanalyzer. We are really grateful to Alexandre Fontayne for his help in the P3 patenting process. We thank Roberto Lleras for helpful information. We are grateful to Gérard Lefranc for helpful comments, Arthur Lavoie and Karthik Kalyan for IMGT/HighV-QUEST and to all IMGT team members for their constant motivation. IMGT® is Academic Institutional Member of the International Medical Informatics Association (IMIA) and of the Global Alliance for the Genomics and Health (GA4GH).

FUNDING

IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS); the Ministère de l'Enseignement Supérieur et de la Recherche (MESR); the Montpellier University, France; the Agence Nationale de la Recherche (ANR) Labex MabImprove [ANR-10-LABX-5301]; BioCampus Montpellier; Région Languedoc-Roussillon [Grand Plateau Technique pour la Recherche (GPTR)]. This work was granted access to the HPC@LR and to the High Performance Computing (HPC) resources of the Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation [036029] (2010–2017) made by GENCI (Grand Equipement National de Calcul Intensif). This study was achieved within the context of the Laboratory of

Excellence TRAIL ANR-10-LABX-57, referenced ANR-10-LABX-0057 and named TRAIL. A public grant from the French National Agency within the context of the Investments for the Future Program (ANR-13-BSV5-0018 SVSE5 Program), named ATHERANOS supported this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/fimmu.2017.01796/full#supplementary-material.

REFERENCES

1. McCafferty J, Griffiths AD, Winter G, Chiswell DJ. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* (1990) 348(6301):552–4. doi:10.10138/348552a0
2. Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J Mol Biol* (1991) 222(3):581–97. doi:10.1016/0022-2836(91)90498-U
3. Griffiths AD, Malmqvist M, Marks JD, Bye JM, Embleton MJ, McCafferty J, et al. Human anti-self antibodies with high specificity from phage display libraries. *EMBO J* (1993) 12(2):725–34.
4. Peneff C, Lefranc M-P, Dariavach P. Characterisation and specificity of two single-chain Fv antibodies directed to the protein tyrosine kinase Syk. *J Immunol Methods* (2000) 236:105–15. doi:10.1016/S0022-1759(99)00228-8
5. Pelat T, Hust M, Laffly E, Condemine F, Bottex C, Vidal D, et al. A high affinity, human-like antibody fragment (scFv) neutralising the lethal factor (LF) of *Bacillus anthracis* by inhibiting PA-LF complex formation. *Antimicrob Agents Chemother* (2007) 51:2758–64. doi:10.1128/AAC.01528-06
6. Thie H, Meyer T, Schirrmann T, Hust M, Dübel S. Phage display derived therapeutic antibodies. *Curr Pharm Biotechnol* (2008) 9:439–46. doi:10.2174/138920108786786349
7. Marcus WD, Wang H, Lindsay SM, Sierks MR. Characterization of an antibody scFv that recognizes fibrillar insulin and beta-amyloid using atomic force microscopy. *Nanomedicine* (2008) 4(1):1–7. doi:10.1016/j.nano.2007.11.003
8. Pelat T, Hust M, Hale M, Lefranc M-P, Dübel S, Thullier P. Isolation of a human-like antibody fragment (scFv) that neutralizes ricin biological activity. *BMC Biotechnol* (2009) 9:60. doi:10.1186/1472-6750-9-60
9. Tian H, Davidowitz E, Lopez P, He P, Schulz P, Moe J, et al. Isolation and characterization of antibody fragments selective for toxic oligomeric tau. *Neurobiol Aging* (2015) 36(3):1342–55. doi:10.1016/j.neurobiolaging.2014.12.002
10. Williams SM, Venkataraman L, Tian H, Khan G, Harris BT, Sierks MR. Novel atomic force microscopy based biopanning for isolation of morphology specific reagents against TDP-43 variants in amyotrophic lateral sclerosis. *J Vis Exp* (2015) 96:1–13. doi:10.3791/52584
11. Krag DN, Shukla GS, Shen GP, Pero S, Ashikaga T, Fuller S, et al. Selection of tumor-binding ligands in cancer patients with phage display libraries. *Cancer Res* (2006) 66:7724–33. doi:10.1158/0008-5472.CAN-05-4441
12. Deramchia K, Jacobin-Valat MJ, Laroche-Traineau J, Bonetto S, Sanchez S, Dos Santos P, et al. By-passing large screening experiments using sequencing as a tool to identify scFv fragments targeting atherosclerotic lesions in a novel *in vivo* phage display selection. *Int J Mol Sci* (2012) 13:6902–23. doi:10.3390/ijms13066902
13. Deramchia K, Jacobin-Valat MJ, Vallet A, Bazin H, Santarelli X, Sanchez S, et al. *In vivo* phage display to identify new human antibody fragments homing to atherosclerotic endothelial and subendothelial tissues. *Am J Pathol* (2012) 180:2576–89; Erratum in: *Am J Pathol* (2012) 181(5):1889. doi:10.1016/j.ajpath.2012.02.013
14. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, et al. By-passing in vitro screening – next generation sequencing technologies applied to antibody display and *in silico* candidate selection. *Nucleic Acids Res* (2010) 38(21):e193. doi:10.1093/nar/gkq789
15. Larman HB, Xu GJ, Pavlova NN, Elledge SJ. Construction of a rationally designed antibody platform for sequencing-assisted selection. *Proc Natl Acad Sci U S A* (2012) 109:18523–8. doi:10.1073/pnas.1215549109
16. Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods* (2013) 60(1):99–110. doi:10.1016/j.ymeth.2013.03.001
17. Lefranc M-P, Lefranc G. *The Immunoglobulin FactsBook*. London: Academic Press (2001). p. 1–458.
18. Lefranc M-P. Immunoglobulin and T cell receptor genes: IMGT(®) and the birth and rise of immunoinformatics. *Front Immunol* (2014) 5:22. doi:10.3389/fimmu.2014.0002223
19. Martinez O, Gangi E, Mordi D, Gupta S, Dorevitch S, Lefranc M-P, et al. Diversity in the complementarity determining region 3 (CDR3) of antibodies from mice with evolving anti-TSHR antibody responses. *Endocrinology* (2007) 148:752–61. doi:10.1210/en.2006-1096
20. Li L, Wang XH, Williams C, Volsky B, Steczko O, Seaman MS, et al. A broad range of mutations in HIV-1 neutralizing human monoclonal antibodies specific for V2, V3, and the CD4 binding site. *Mol Immunol* (2015) 66(2):364–74. doi:10.1016/j.molimm.2015.04.011
21. Marillet S, Lefranc M-P, Boudinot P, Cazals F. Novel structural parameters of Ig-Ag complexes yield a quantitative description of interaction specificity and binding affinity. *Front Immunol* (2017) 8:34. doi:10.3389/fimmu.2017.00034
22. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res* (2015) 43:D413–22. doi:10.1093/nar/gku1056
23. Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res* (2004) 32:W435–40. doi:10.1093/nar/gkh412
24. Brochet X, Lefranc M-P, Giudicelli V. 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* (2008) 36:W503–8. doi:10.1093/nar/gkn316
25. Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* (2011) 6:695–715. doi:10.1101/pdb.prot5633
26. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* (2004) 20:i379–85. doi:10.1093/bioinformatics/bth945
27. Giudicelli V, Lefranc M-P. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harb Protoc* (2011) 6:716–25. doi:10.1101/pdb.prot5634
28. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P. 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* (2012) 8(1):2. doi:10.4172/1745-7580.1000056
29. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
30. Li S, Lefranc M-P, Miles JJ, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) 4:2333. doi:10.1038/ncomms3333
31. Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, Kossida S. From IMGT-ONTOLOGY to IMGT/HighV-QUEST for NGS immunoglobulin (IG) and T cell receptor (TR) repertoires in autoimmune and infectious diseases. *Autoimmun Infec Dis* (2015) 1:1. doi:10.16966/aidoa.103
32. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc M-P. IMGT/HighV-QUEST statistical significance of IMGT clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles

of immunoglobulins and T cell receptors. *PLoS One* (2015) 10(11):e0142353. doi:10.1371/journal.pone.0142353
33. Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc M-P. IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. *Front Immunol* (2016) 7:339. doi:10.3389/fimmu.2016.00339
34. Giudicelli V, Duroux P, Kossida S, Lefranc M-P. IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. *BMC Immunol* (2017) 18(1):35. doi:10.1186/s12865-017-0218-8
35. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* (2015) 13:278–89. doi:10.1016/j.gpb.2015.08.002
36. Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, Midwinter W, et al. HLA typing for the next generation. *PLoS One* (2015) 10:e0127153. doi:10.1371/journal.pone.0127153
37. Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, Garcia K, et al. No assembly required: full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Hum Immunol* (2015) 76:891–6. doi:10.1016/j.humimm.2015.03.022
38. Carapito R, Radosavljevic M, Bahram S. Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. *Hum Immunol* (2016) 77:1016–23. doi:10.1016/j.humimm.2016.04.002
39. Hua TD, Souriau C, Marin M, Lefranc M-P, Weill M. Construction d'un répertoire de fragments d'anticorps scFv et son expression à la surface de phages filamenteux [French]. In: Lefranc and Lefranc, editor. *Ingénierie des anticorps. Banques combinatoires. Techniques en Immunologie*. (Chap. 5), Paris: Les Editions INSERM (1997). p. 39–54.
40. Lefranc M-P. IMGT locus on focus: a new section of experimental and clinical immunogenetics. *Exp Clin Immunogenet* (1998) 15:1–7. doi:10.1159/000019049
41. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, et al. Deep sequencing in library selection projects: what insight does it bring? *Curr Opin Struct Biol* (2015) 33:146–60. doi:10.1016/j.sbi.2015.09.001
42. Yang W, Yoon A, Lee S, Kim S, Han J, Chung J. Next-generation sequencing enables the discovery of more diverse positive clones from a phage-displayed antibody library. *Exp Mol Med* (2017) 49:e308. doi:10.1038/emm.2017.22
43. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2015) 21:86–91,1. doi:10.1038/nm.3743
44. D'Angelo S, Kumar S, Naranjo L, Ferrara F, Kiss C, Bradbury AR. From deep sequencing to actual clones. *Protein Eng Des Sel* (2014) 27:301–7. doi:10.1093/protein/gzu032
45. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* (2010) 28:965–71. doi:10.1038/nbt.1673

Conflict of Interest Statement: CH, NH, JKuijpers, and JKorlach are full-time employees at Pacific Biosciences, a company developing single-molecule sequencing technologies. MS was a full-time employee at Pacific Biosciences at the time of this work. PM is full-time employee at LFB, a pharmaceutical group specializing in biological medicinal products. The other authors declare that they have no conflict of interest.

Copyright © 2017 Hemadou, Giudicelli, Smith, Lefranc, Duroux, Kossida, Heiner, Hepler, Kuijpers, Groppi, Korlach, Mondon, Ottones, Jacobin-Valat, Laroche-Traineau and Clofent-Sanchez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

REVIEW
published: 01 February 2018
doi: 10.3389/fimmu.2018.00117

Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing

Arjun K. Mishra[1,2] *and Roy A. Mariuzza*[1,2]*

[1] *W. M. Keck Laboratory for Structural Biology, Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Rockville, MD, United States,* [2] *Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, College Park, MD, United States*

Affinity maturation is the process whereby the immune system generates antibodies of higher affinities during a response to antigen. It is unique in being the only evolutionary mechanism known to operate on a molecule in an organism's own body. Deciphering the structural mechanisms through which somatic mutations in antibody genes increase affinity is critical to understanding the evolution of immune repertoires. Next-generation sequencing (NGS) has allowed the reconstruction of antibody clonal lineages in response to viral pathogens, such as HIV-1, which was not possible in earlier studies of affinity maturation. Crystal structures of antibodies from these lineages bound to their target antigens have revealed, at the atomic level, how antibodies evolve to penetrate the glycan shield of envelope glycoproteins, and how viruses in turn evolve to escape neutralization. Collectively, structural studies of affinity maturation have shown that increased antibody affinity can arise from any one or any combination of multiple diverse mechanisms, including improved shape complementarity at the interface with antigen, increased buried surface area upon complex formation, additional interfacial polar or hydrophobic interactions, and preorganization or rigidification of the antigen-binding site.

Keywords: antibody, affinity maturation, somatic hypermutation, HIV-1, structural biology, next-generation sequencing

OPEN ACCESS

Edited by:
Prabakaran Ponraj, Intrexon, United States

Reviewed by:
Robyn Stanfield, The Scripps Research Institute, United States
Stanley Nithianantham, University of California, Davis, United States

***Correspondence:**
Roy A. Mariuzza
rmariuzz@umd.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *16 November 2017*
Accepted: *15 January 2018*
Published: *01 February 2018*

Citation:
Mishra AK and Mariuzza RA (2018) Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing. Front. Immunol. 9:117. doi: 10.3389/fimmu.2018.00117

INTRODUCTION

The ability of the humoral immune system to generate high-affinity binders for virtually any antigen is predicated on its capacity to produce large repertoires of antibodies encompassing a vast array of specificities and to then select members of this repertoire with high affinity for a particular immunogen (1–3). The extensive sequence diversity of antibody molecules derives from several sources: (i) combinatorial diversification, whereby two sets of light (L) chain gene segments, V_L and J_L, and three sets of heavy (H) chain gene segments, V_H, D, and J_H, rearrange to produce functional variable (V) regions; (ii) imprecise joining of these gene segments; and (iii) somatic hypermutation, by which point mutations, as well as insertions and deletions (indels), are introduced throughout the sequences encoding L and H chains (4). B cells expressing antibodies with improved affinity are better equipped to compete for antigen and thus receive signals that result in preferential expansion and further antibody sequence diversification *via* additional rounds of somatic hypermutation (1–3). Through this rapid evolutionary process of mutation and selection, antibody affinity typically improves 10- to 5,000-fold during the course of an immune response, bolstering host defense.

Recent advances in next-generation sequencing (NGS) have revolutionized the analysis of antibody repertoires by dramatically increasing sample depth compared to previous low-throughput

methods (5). In addition, new methods have been developed for single-cell sequencing, which allow large-scale determination of paired L and H chains. These advances, in conjunction with computational tools for reconstituting antibody clonal lineages, can provide a genetic record for the evolutionary processes of recombination and somatic hypermutation in immune responses to specific microbial pathogens. Crystal structures of affinity-matured and germ line antibodies from such lineages in complex with their target antigens have produced new insights into the molecular basis of affinity maturation.

Here, we first review our understanding of the basic biophysical principles underlying affinity maturation before the arrival of NGS. We then summarize what studies of bulk B cell populations by NGS have taught us about the general features of antibody repertoire selection. Finally, we discuss structural studies of reconstructed antibody clonal lineages with special emphasis on the immune response to HIV-1, which has so far benefited most from the application of NGS and single-cell analysis to better understanding affinity maturation (6) (**Table 1**).

STUDIES OF AFFINITY MATURATION PRIOR TO NGS

Before the advent of NGS, a number of structural studies were carried out comparing affinity-matured antibodies and their putative germ line precursors bound to the same antigen. In studies involving small molecules (haptens) such as phenyloxazolone and nitrophenyl phosphonate, rather than proteins, it was found that somatic mutations in complementarity-determining region (CDR) residues directly or indirectly implicated in binding hapten permit the formation of additional hydrogen bonds, electrostatic interactions, and van der Waals contacts (23–26). A particularly revealing case involved the matured 48G7 antibody, which binds nitrophenyl phosphonate ~3,000-fold more tightly than its germ line counterpart 48G7g (26). Large changes in the conformation of the antigen-binding site (paratope) of 48G7g were observed upon hapten engagement by this germ line antibody, whereas the free and hapten-bound forms of affinity-matured 48G7 showed few structural differences. Thus, affinity maturation in this case appeared to be driven largely by a mechanism of preorganizing the paratope into a conformation favorable for binding its hapten ligand (26). Such conformational preorganization was accompanied by a decrease in the flexibility of the paratope during the maturation process, which may increase specificity for the target antigen while reducing the possibility of cross-reactivity with other antigens, including self-antigens (27–29). The antibody maturation process appears to simultaneously select for both higher binding affinity and increased thermodynamic stability. In a study of matured antibody 93F3, which recognizes a small hapten, somatic mutations in the paratope that increased affinity were found to reduce the melting temperature of 93F3 compared to its germ line precursor. However, the destabilizing effects of these mutations were compensated by additional somatic mutations in the V_L/V_H interface, distal to the paratope (29).

The first structural study of the maturation of an antibody response to a protein antigen, instead of a hapten, involved a set of closely related antibodies specific for hen egg white lysozyme (HEL) (30). These antibodies represented different stages of affinity maturation, whereby the number of somatic mutations correlated with increasing affinity. Surprisingly, improved affinity could not be attributed to the formation of additional hydrogen bonds or salt bridges or to an increase in total buried surface area. Instead, affinity maturation resulted mainly from burial of increasing amounts of hydrophobic surface at the expense of polar surface, accompanied by improved shape complementarity at the V_H–HEL interface. The increase in hydrophobic interactions resulted from highly correlated structural rearrangements in antibody residues at the periphery of the interface with antigen, adjacent to the central energetic hot spot (30). Indeed, the periphery may offer more suitable sites for optimization because these regions are typically more flexible and tolerant to mutations than central sites (12), in agreement with the finding that somatic hypermutation spreads structural diversity generated by V(D)J recombination from central to peripheral regions of the antibody binding site (31).

Collectively, these structural studies showed that increased antibody affinity for small haptens or model protein antigens such as HEL can arise from any one or any combination of several variables, including additional interfacial hydrogen bonds or van der Waals contacts, conformational preorganization of the paratope, improved shape complementarity at the interface with antigen, or increased burial of total or hydrophobic surface area. These same basic strategies, as well as others, govern affinity maturation of antibody responses to biological antigens such as the envelope

TABLE 1 | Structural studies of antibody clonal lineages reconstructed using next-generation sequencing.

Antibody lineage	Specificity	PDB code	Reference
CH58	HIV-1 envelope glycoprotein (Env) gp120	4HPO	(7)
		4RIS, 4RIR	(8)
VRC01	HIV-1 Env gp120	4JPV, 4JPW, 4LSP, 4LSQ, 4LSR, 4LSS, 4LST, 4LSU, 4LSV	(9)
		5F7E, 5FA2, 5FEC, 5IGX, 5I90	(10)
		4JPK, 4JPI	(11)
VRC03	HIV-1 Env gp120	5JOF, 5JXA	(12)
CH103	HIV-1 Env gp120	4JAM, 4JAN	(13)
		4QHK, 4QHL, 4QHM, 4QHN	(14)
PGT121	HIV-1 Env gp120	4NCO	(15)
		4R26, 4R2G	(16)
		5CEX, 5CEY, 5CEZ	(17)
PCT64	HIV-1 Env gp120	5FEH	(18)
CAP256	HIV-1 Env gp120	4OCR, 4OCS, 4OCW, 4OD1, 4OD3, 4ODH, 4ORG	(19)
ANC195	HIV-1 Env gp120	5CJX	(20)
C65	Influenza A virus HA	4HK0, 4HK3, 4HKB, 4HKX	(21)
HV6-1 + HD3-3	Influenza A virus HA	5K9J, 5K9K, 5K9O, 5K9Q, 5KAN, 5KAQ	(22)

glycoproteins of HIV-1 and other viral pathogens, as discussed below. We first summarize what NGS of bulk B cell populations has taught us about antibody repertoire selection. We then discuss recent insights into affinity maturation gained from structural studies of antibody clonal lineages that were reconstructed using NGS (**Table 1**).

NGS ANALYSIS OF ANTIBODY REPERTOIRES IN BULK B CELL POPULATIONS

Next-generation sequencing of paired antibody L and H chains combined with computational modeling of antibody structures has been used to profile human antibody repertoire selection and maturation at the population level (5). In the most exhaustive study to date, a comparison of ~55,000 V_L/V_H pairs from naïve B cells with ~120,000 V_L/V_H pairs from antigen-experienced B cells, all isolated from human peripheral blood, showed that V_L and V_H genes pair in a purely combinatorial fashion with no detectable biases, but that certain V_L/V_H gene pairs are significantly depleted or enriched in the antigen-experienced repertoire compared to the naïve repertoire (32). Repertoire-wide computational structure prediction was carried out to characterize the physiochemical properties of the antibody paratopes. Whereas no appreciable differences in paratope hydrophobicity or solvent-accessible surface area were evident in antigen-experienced versus naïve antibodies, antigen-experienced V_LCDR3 and V_HCDR3 amino acid sequences displayed slightly increased positive charge compared to naïve sequences (32). Overall, however, the evolutionary processes of somatic hypermutation and affinity selection that occur in periphery blood did not leave a distinctive physiochemical imprint on the antigen-experienced antibody repertoire, even at the level of CDR3s, which are a major focus for somatic hypermutation. By contrast, bone marrow B cells expressing antibodies with positively charged CDR3 loops undergo preferential elimination at discrete developmental checkpoints before entering the periphery, possibly as a mechanism for reducing the risk of self-reactivity (33).

In a study to determine whether NGS could be used to detect antigen-specific sequences in bulk B cell populations, an analysis of identical CDR3 sequences that were shared by individuals previously vaccinated against *Haemophilus influenzae* type b identified a number of sequences known to be specific for this bacterium (34). Conserved CDR3 sequences were also observed in patients recovering from acute dengue infection (35), indicating convergent antibody evolution in different individuals exposed to the same antigens. In another study, NGS and single-cell sorting of peripheral blood plasmablasts were used to profile the acute antibody response to influenza A vaccination (36). Antibodies able to neutralize the virus were selected bioinformatically from clonal families. These vaccine-induced antibodies contained on average >30 somatic mutations overall. Notably, some antibodies exhibited higher affinities for hemagglutinins (HAs) from prior years' influenza strains than for the HA of the immunizing strain, suggesting recall of memory B cells expressing antibodies that had previously undergone affinity maturation (36).

STUDIES OF AFFINITY MATURATION AFTER NGS

Next-generation sequencing coupled with bioinformatics analysis has allowed, for the first time, the reconstruction of antibody clonal lineages and inference of germ line progenitor sequences, neither of which was possible in earlier studies of affinity maturation (37). However, an important caveat is that germ line sequences are predicted sequences that may differ from the true unmutated ancestor sequences. Whereas mutations in V_L and V_H gene segments can be identified with high confidence, the original V_LJ_L and (especially) V_HDJ_H junctional sequences of germ line antibodies are uncertain. In particular, it is impossible to know if insertions or deletions in these sequences took place during V(D)J recombination or were introduced during B cell affinity maturation. As a consequence, statistical methods must be used to infer the most likely unmutated common ancestor for an aligned set of sequences that are taken to be clonally related (37).

As measured by surface plasmon resonance (SPR), a putative germ line precursor of the anti-HIV-1 antibody 2F5, which recognizes the gp41 subunit of the HIV-1 envelope glycoprotein (Env), bound recombinant gp41 with ~500-fold lower affinity than the matured antibody (K_D = 0.7 μM versus 1.2 nM) (38). Micromolar K_Ds were also reported for germ line precursors of broadly neutralizing antibodies (bNAbs) CH01 and CH04, which recognize the V2/V3 quaternary epitope of the gp120 subunit of HIV-1 Env (39). Although seemingly low, such affinities are nevertheless sufficient to trigger affinity maturation of unmutated B cells *in vivo* (40, 41).

By contrast, the putative germ line ancestors of several bNAbs specific for the CD4 binding site of HIV-1 Env (b12, NIH45-46, and 3BNC60) failed to show detectable binding to recombinant Envs, raising the question of how B cell maturation leading to the eventual production of these bNAbs was initiated (42, 43). Moreover, this failure was observed even though the amino acid sequences of the V(D)J junctions of the affinity-matured antibodies were left unchanged in the reconstructed germ line versions. One possibility is that maturation of anti-CD4 binding site bNAb precursors was triggered by non-HIV antigens and that the resultant antibody intermediates serendipitously cross-reacted with Env. More likely, however, interactions between proteins in solution (3D affinity), as measured by SPR or related techniques, differ from those at contacts between two cells or between a cell and a virus (2D affinity) (44). For example, whereas SPR was unable to detect any binding between CD4 and MHC class II (45), the affinity of CD4 for MHC class II on B cells could be measured in 2D using CD4-functionalized supported lipid bilayers (46). Therefore, under physiological conditions, germ line precursors of anti-CD4 binding site bNAbs, expressed on the surface of B cells, might be engaged by membrane-anchored Env on the virion surface or on the surface of infected cells with sufficient affinity to trigger B cell maturation. In support of this idea, the germ line precursors of several anti-influenza HA bNAbs were found to bind to HA only when presented on membranes in the form of cell surface IgMs; as soluble IgGs, these precursors had no detectable affinity for HA (47). In the following sections, we have selected representative examples of affinity maturation

in order to illustrate the multiple structural strategies that the antibodies employ to increase potency and breadth of pathogen neutralization.

Preorganization, Rigidification, and Reorientation

Antibody CH58 was isolated from a participant in the RV144 HIV vaccine efficacy trial. Like most bNAbs elicited in response to HIV-1 infection, CH58 is highly mutated (7). The structure of affinity-matured CH58 in complex with an Env V2 peptide showed that this bNAb targets V2 residue Lys169, which is a site of vaccine-induced immune pressure. Structures have also been determined for the putative germ line precursor of CH58 in unliganded form and bound to the V2 peptide (8). A comparison of these structures revealed that affinity maturation of CH58 is driven by the formation of two new salt bridges linking V_LCDR1Asn31→Asp and V_HCDR1Ser28→Arg to Lys171 and Asp180 of V2, respectively (**Figure 1**) (8). In addition, V_LCDR3 in the unbound germ line precursor adopts a different conformation in the CH58–V2 complex, implying flexibility. By contrast, the conformation of V_LCDR3 in the CH58–V2 complex is nearly identical to that in unbound matured CH58. Such preorganization of the CH58 paratope into a configuration more suitable for binding V2, accompanied by rigidification of V_LCDR3 to lower the entropic cost of complex formation, further contributes to the 2,000-fold affinity increase during maturation. Paratope preorganization and rigidification have also been described for the CH65 lineage of anti-influenza virus HA antibodies (21), underscoring the general utility of these mechanisms for improving affinity (26–29).

A related study used hydrogen/deuterium exchange in combination with mass spectrometry (HDX/MS) to investigate how affinity maturation alters the dynamics of bNAbs against HIV-1 Env (12). Importantly, the high variability and constantly evolving nature of HIV-1 Env distinguish this conformationally dynamic glycoprotein from the static model antigens used in studies of affinity maturation prior to NGS (23–26, 30). HDX/MS directly measures local protein dynamics by monitoring backbone amide deuterium uptake when the protein is diluted into a solution of D_2O. Dynamic regions of a protein take up deuterium more rapidly than stable regions. HDX/MS was used to compare the local dynamics of the predicted germ line and affinity-matured forms of two bNAbs (VRC03 and VRC-PG04) specific for the CD4 binding site of HIV-1 Env (12). In both cases, the paratopes of the matured bNAbs were less dynamic overall that those of their germ line counterparts, in agreement with previous evidence that paratopes become more rigid during the maturation process (28, 30, 48). Surprisingly, however, the largest decreases in dynamics occurred at the periphery of the paratopes, at sites adjacent to Env glycans, rather than at primary Env-contacting sites. A similar pattern was observed for a bNAb (CAP256-VRC26) specific for

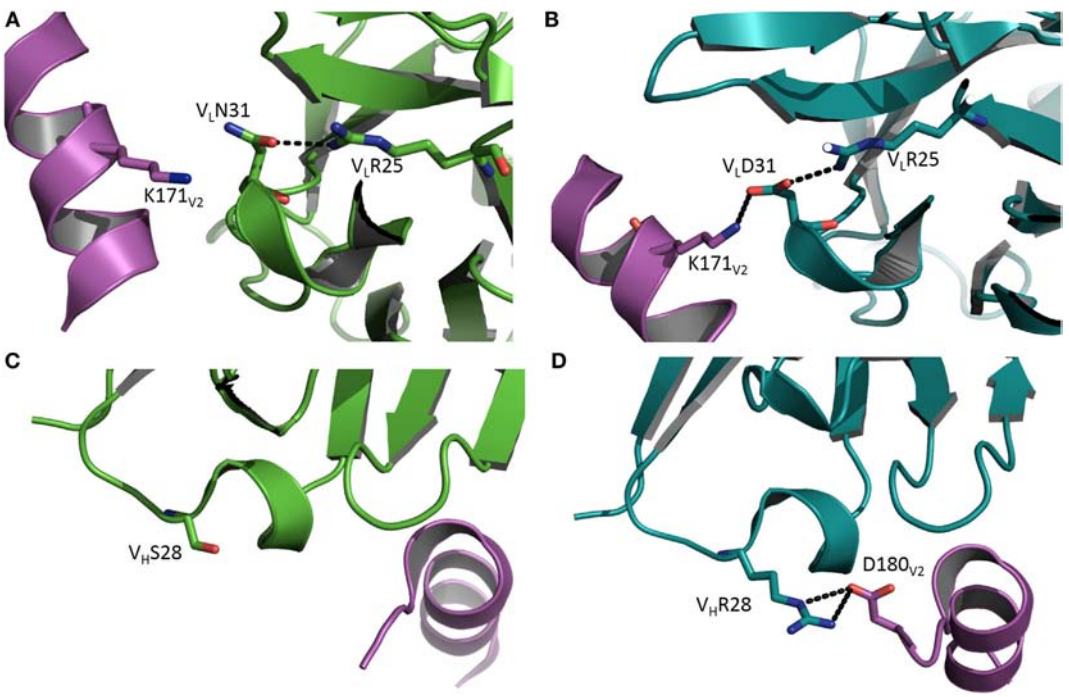


FIGURE 1 | Affinity maturation through formation of additional interactions with antigen. **(A)** Close-up view of the interface between the germ line precursor of antibody CH58 and the V2 peptide of HIV-1 envelope glycoprotein in the vicinity of V2 Lys171 (Protein Data Bank accession code 4RIS) (8). V_L is green; V2 is magenta. **(B)** In affinity-matured CH58, the somatic mutation V_LAsn31→Asp allows formation of a stabilizing salt bridge (dotted black line) to V2 Lys171. V_L is teal. **(C)** Close-up of the interface between the germ line precursor of CH58 and the V2 peptide in the vicinity of V_HSer28 (4HPO) (7). V_H is green; V2 is magenta. **(D)** In affinity-matured CH58, the mutation V_HSer28→Arg results in formation of a bidentate salt bridge with V2 Asp180. V_H is teal. This residue was not visible in **(C)** due to disorder in the C-terminus of the V2 peptide.

the V1/V2 quaternary epitope of HIV-1 Env. This stabilization of the paratope periphery may serve to minimize potential clashes with nearby Env glycans, while maintaining critical binding interactions mediated by the center of the paratope. It is probably not coincidental that a similar focus of affinity maturation on sites peripheral rather than central to the interface with antigen was also observed for anti-HEL antibodies (30).

Antibodies of the CH103 lineage block the CD4 binding site of HIV-1 *via* interactions dominated by V_HCDR3 (13). Affinity maturation of CH103 is associated with mutations in both contacting and non-contacting residues, including framework (FR) residues distant from the interface with Env. Structural analysis of the putative germ line precursor of CH103 and of two intermediates in the maturation pathway revealed a shift in the relative orientation of the V_L and V_H domains during evolution of the CH103 lineage, corresponding to a root mean squared deviation in α-carbon positions of 2.1 Å (**Figure 2A**) (14). This shift is mediated by several mutations at the V_L/V_H interface, including a leucine-to-valine substitution at FR position V_L46 that not only contributes to reorienting the V_L domain but also to reconfiguring V_HCDR3 (**Figure 2B**). Although not as important as the V_LLeu46→Val mutation, the neighboring V_LTyr49→Phe and V_HPhe100→Tyr mutations, which are also located in the V_L/V_H interface, may further contribute to V_L/V_H reorientation. Most likely, V_L/V_H reorientation occurred in response to insertions in the V5 loop of Env during infection, which had allowed the virus to escape neutralization by progenitors of CH103. Displacement of V_L away from V5 allowed accommodation of insertions in V5 without steric hindrance (14). In addition, the conformation of V_HCDR3 in the germ line precursor of CH103 is incompatible with gp120 binding, at least as observed in the CH103–gp120 complex (13), thereby necessitating rearrangement of this loop during the maturation process (14).

In sharp contrast to bNAbs against HIV-1, which are highly mutated, the potent human anti-Middle East respiratory syndrome coronavirus antibody m336 is almost germ line, with only one somatic mutation in the H chain (49). The structure of m336 in complex with the MERS-CoV receptor-binding domain showed that the IGV1-69-derived H chain contributes >85% of the binding surface. The subnanomolar affinity of m336 despite the near absence of somatic mutations results from direct interactions with germ line-encoded V_HCDR2 and V_H framework region 3 (V_HFR3) residues and with recombination-generated residues in the V_HDJ_H junction (49). Like m336, the anti-influenza HA bNAb CR6261 uses the IGV1-69 V_H gene segment (47). Virus neutralization depended solely on the H chain, and only seven somatic mutations in V_HCDR1 and V_HFR3 were required for maximum affinity.

To systematically dissect the contribution of mutations in FR residues to affinity maturation, deep mutational scanning was applied to the anti-vascular endothelial cell growth factor antibody G6.31 (50). A number of FR mutations at positions distal to the CDRs were found to improve the affinity and/or thermostability of G6.31. In particular, the FR mutation V_LPhe83→Ala, which is ~25 Å away from the antigen-binding site, increased both affinity and stability by altering the orientation of the constant domains (C_L and C_H1) relative to V_L and V_H, as well as the orientation of V_L to V_H (50). As measured by HDX/MS, the V_LPhe83→Ala mutation modulated the interdomain conformational dynamics of antibody G6.31. Furthermore, analysis of > 5,000 human V_L sequences showed that somatic mutations occur frequently at position 83, strongly suggesting its biological relevance in repertoire selection.

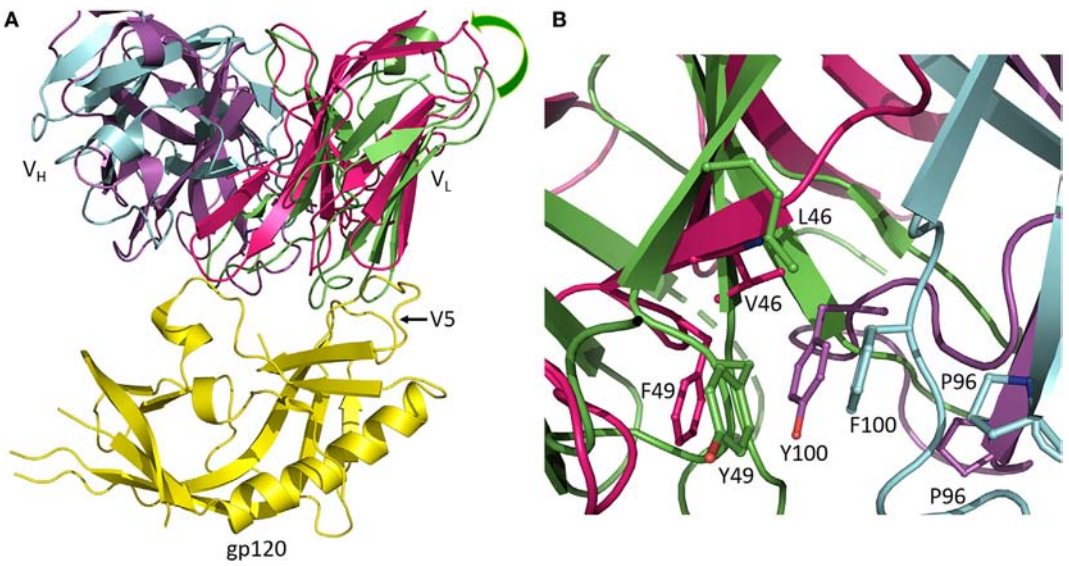


FIGURE 2 | Reorientation of V_L and V_H domains in response to viral escape mutations. **(A)** Superposition of the germ line precursor of CH103 (V_L and V_H domains are green and cyan, respectively) (4QHL) (14) onto matured CH103 (V_L and V_H domains are red and magenta, respectively) in complex with envelope glycoprotein gp120 (yellow) (4JAN) (13). During affinity maturation, a shift occurred in the orientation of V_L with respect to V_H. The shift is an adaptation to insertions in the gp120 V5 loop during infection. Movement of V_L away from gp120 enables accommodation of the V5 insertion. **(B)** Close-up view of the V_L/V_H interface in the vicinity of the V_LLeu46⟶Val mutation showing changes in interdomain contacts and rearrangement of V_HCDR3.

Affinity Maturation of Glycan-Binding Antibodies

Besides high sequence variability, another feature of HIV-1 Env that distinguishes it from model antigens such as HEL used in previous studies of affinity maturation is glycosylation. Indeed, extensive N-glycosylation masks much of the Env protein surface from antibody recognition. Nevertheless, a number of potent bNAbs have been discovered that penetrate this glycan shield and engage both carbohydrate and protein antigenic determinants (51, 52). In the most thoroughly studied example to date, the PGT121 family of bNAbs was shown to bind to N-glycans located in a high-mannose patch centered on the highly conserved Asn332 glycan and to protein elements at the base of the V3 loop (15, 16).

As revealed by NGS and X-ray crystallography, the putative germ line precursor of the PGT121 family splits into two evolutionary branches that differ considerably in how they interact with Env glycans (**Figures 3A,B**). One branch, exemplified by PGT124, contacts only the Asn332 glycan (16), whereas the other branch, exemplified by PGT122, also contacts the Asn137, Asn156, and Asn301 glycans (15). PGT124 and PGT122 employ a common set of CDR residues to contact the Asn332 glycan and V3 loop, nearly all of which are also found in the germ line precursor. This conservation suggests that a critical event in triggering the antibody response is simultaneous recognition of both carbohydrate and protein determinants.

Antibodies PGT124 and PGT122 share similar binding site architectures with long V_HCDR3 loops that pack against the entire length of the Asn332 glycan, thereby penetrating the glycan shield to reach the Env protein surface below (**Figure 3C**) (16). In both antibodies, as well as in the germ line precursor, a closed face on one side of the paratope engages the Asn332 glycan. A second, open face differs between PGT124 and PGT122. For PGT124, the open face enables avoidance of neighboring glycans through a shift of V_HCDR3 away from its equivalent position in PGT122 (**Figure 3C**). For PGT122, by contrast, mutations during affinity maturation result in productive interactions between V_HCDR2 and the Asn137 glycan. In agreement with the structural data, deletion of the Asn332 glycan abolished PGT124 neutralization of nearly all HIV-1 isolates, whereas deletion of neighboring glycans had little or no effect (16). Conversely, PGT122 is less reliant on the Asn332 glycan than PGT124 for virus neutralization because of its ability to utilize alternative glycans at Asn137 and Asn301 to achieve high-affinity binding.

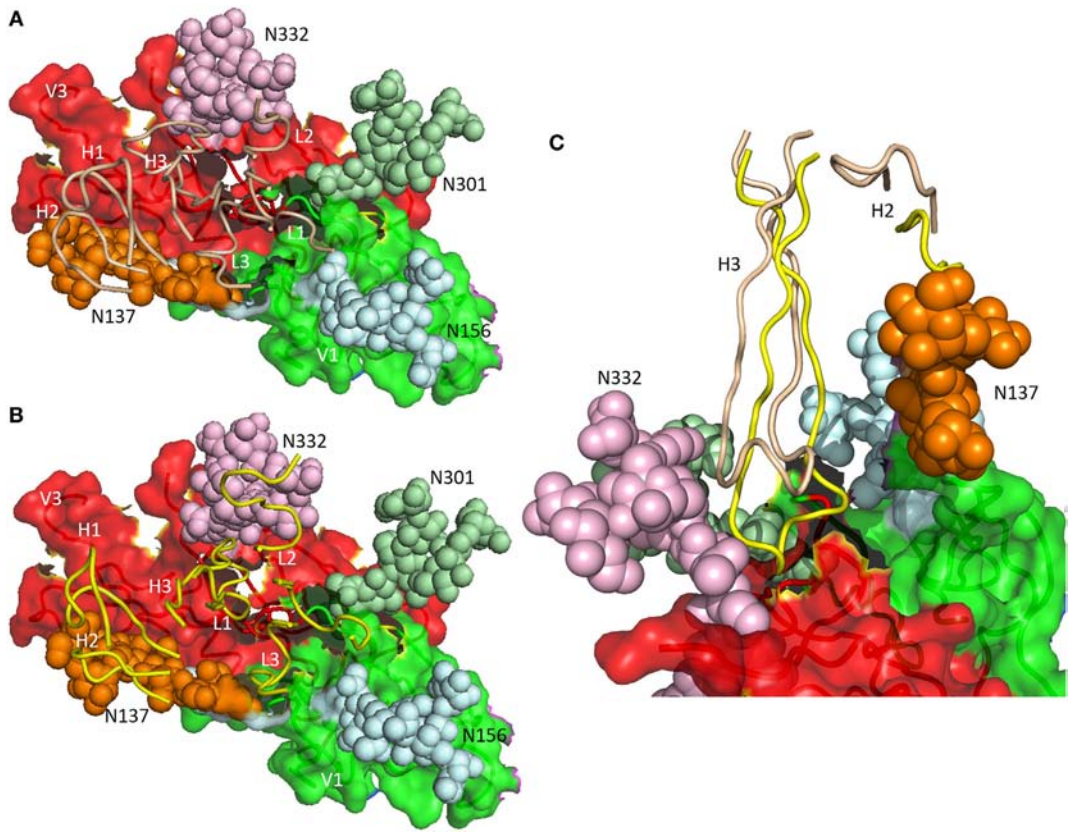


FIGURE 3 | Differential glycan recognition by affinity-matured antibodies. **(A)** Positions of complementarity-determining region (CDR) loops (wheat) of matured antibody PGT124 on envelope glycoprotein (Env) gp120 (4R2G) (16). V_L CDR loops are labeled L1–L3; V_H CDR loops are labeled H1–H3. The V1 (green) and V3 (red) regions of gp120 are depicted as molecular surfaces. Glycans are drawn as spheres and labeled N137, N156, N301, and N332. The PGT124 epitope is composed of V1, V3, and the N332 glycan. **(B)** Positions of CDR loops (yellow) of matured antibody PGT122 on Env gp120 (4NCO) (15). The PGT122 epitope is composed of V1 and V3, as well as all four glycans at N137, N156, N301, and N332. **(C)** V_HCDR3 of PGT122 (yellow) contacts both the N137 and N332 glycans. V_HCDR3 of PGT124 (wheat) contacts only the N332 glycan due to a shift in position relative to V_HCDR3 of PGT122.

Recently, NGS of HIV-1 longitudinal cohorts has been performed to investigate coevolution of HIV-1 Env and antibodies targeting this glycoprotein and to identify viral variants that initiate maturation of bNAbs (18, 53, 54). These studies have shown that viral escape generates a pool of diverse epitope variants and that somatic hypermutation, acting in parallel, creates antibodies with differential ability to neutralize these variants. Thus, antibodies from the CAP256 and PCT64 lineages, which were isolated independently from two different HIV-1 infected African patients, both target the V2 apex epitope of the Env trimer (18, 53, 54). The crystal structure of bNAb PCT64-35B, isolated at 35 months postinfection, showed that the somatically mutated 25-residue V_HCDR3 loop adopts a β-hairpin conformation that projects above the other CDRs (18). The extended conformation and anionic character of this V_HCDR3, which contains at least one sulfotyrosine, probably enable this loop to penetrate between the glycans that shield the V2 apex epitope to contact the positively charged V1/V2 protein surface of the Env trimer. Moreover, in the PCT64 donor, Env glycoform heterogeneity may have played a role in activating B cell precursors of the PCT64 lineage by allowing germ line antibody binding to early Env trimer forms lacking complex or hybrid glycans at key positions (18).

According to the polyreactivity hypothesis, B cells initially produce germ line antibodies with conformationally flexible combining sites that are able to recognize diverse antigens with low affinity (12, 28, 48). In this way, the immune system can respond to an enormous variety of potential antigens, whose numbers dwarf theoretical estimates of the clonal diversity of germ line antibodies. As described earlier, studies of immune responses to haptens, model proteins, and HIV-1 Env have shown that affinity maturation generates antibodies with higher affinity and specificity than germ line antibodies, at least in part through rigidification of the paratope. However, a recent study of glycan-specific antibodies has suggested an alternative pathway for antibody evolution that is distinct from the polyreactive germ line pathway (55).

A combination of glycan microarrays and molecular dynamics (MD) simulations was used to investigate the affinity maturation of two antibodies (3F8 and ch14.18) specific for the ganglioside GD2, a tumor-associated carbohydrate antigen overexpressed in various cancers, notably neuroblastoma and melanoma (55). Surprisingly, the putative germ line antibodies, although of lower affinity than their matured counterparts, were just as highly selective for CD2 as the matured antibodies when screened against hundreds of glycans and thousands of proteins. Possible reasons for this lack of observable polyreactivity were investigated by MD simulations of germ line and affinity-matured 3F8 and ch14.18 (55). These simulations revealed that, rather than becoming more rigid, the binding sites of the matured anti-G2 antibodies showed an increase in flexibility relative to the binding sites of the germ line antibodies. Most likely, affinity maturation of 3F8 and ch14.18 was enthalpically driven by an increase in direct or water-mediated hydrogen bonds to G2, rather than entropically driven by a decrease in binding site flexibility. The high selectivity of germ line antibodies to the ganglioside G2, compared to the polyreactivity of antibodies to other antigens (12, 28, 48), may reduce the risk of developing autoimmune diseases such as Guillain–Barré syndrome, which is associated with antibodies to the structurally related ganglioside GM2 (56).

Indels in Antibody Affinity Maturation

Multibase in-frame indels are introduced during somatic hypermutation in germinal center B cells along with point mutations (57). As revealed by NGS, the frequency of indels among total somatic mutations in normal human B cell repertoires is low (~2%) (58). In sharp contrast, ~40% of bNAbs against HIV-1 Env contain indels, ranging in size from 1 to 11 amino acids (59). Indels are the primary means by which antibodies can effect large changes in steric volume comparable to those associated with addition or removal of glycans. Indeed, the role of indels in bNAb maturation is to accommodate glycans, penetrate the glycan shield, and/or increase the number of interfacial contacts.

A distinguishing feature of VRC01-class bNAbs that target the CD4 binding site of Env is a deletion of two to six amino acids in V_LCDR1 (9, 10). The function of this deletion is to avoid a steric clash with the Asn276 glycan of loop D of gp120, such that reversion of the deletion to germ line markedly diminished binding. Notably, the V_LCDR1 deletion appears across multiple lineages of VRC01-class bNAbs from different individuals, suggesting that there are few other viable solutions to fitting an antibody into the CD4 binding site of Env.

The structure of antibody 8ANC195 in complex with trimeric Env showed that this bNAb recognizes an epitope that spans the gp41 and gp120 Env subunits (20). An insertion of five amino acids in V_HFR3 extends between the Asn234 and Asn276 glycans of gp120, establishing productive interactions with both glycans and enabling 8ANC195 to penetrate the glycan shield to contact the protein surface of Env (**Figure 4**). Like 8ANC195, bNAb 35O22 binds trimeric Env at the gp41–gp120 interface (60). An

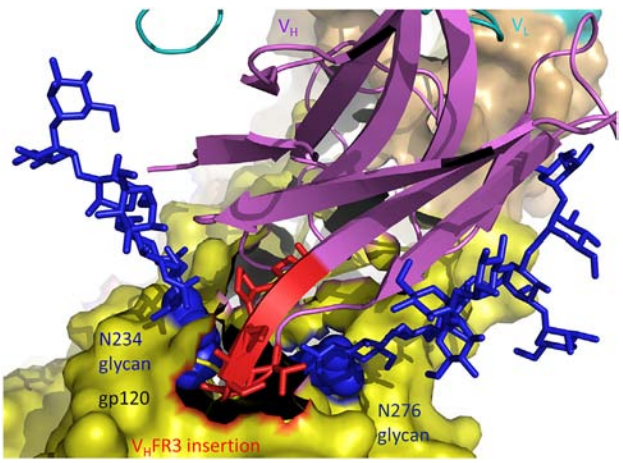


FIGURE 4 | Role of insertions in antibody affinity maturation. An insertion of five amino acids in V_H framework region 3 (V_HFR3) of matured antibody 8ANC195 enables V_HFR3 to extend between the N234 and N276 glycans of envelope glycoprotein (Env) gp120 and contact protein beneath the glycan shield (5CJX) (20). V_H and V_L are magenta and cyan, respectively. The insertion in V_HFR3 is red. Env gp120 (yellow) is drawn as a molecular surface. Glycans are represented as blue sticks.

insertion of eight amino acids in V_HFR3 projects into the cleft between the Env subunits, resulting in additional stabilizing contacts with antigen that increase the neutralization potency of the matured antibody compared to its germ line precursor.

CONCLUSION

Structural studies of antibody affinity maturation spanning nearly 30 years have identified a diversity of biophysical mechanisms underlying this prototypical example of molecular evolution. These include improved shape complementarity at the interface with antigen, increased buried surface area upon complex formation, additional interfacial polar or hydrophobic interactions, preorganization or rigidification of the antigen-binding site, and V_L/V_H reorientation. Over the last 5 years, NGS has allowed the reconstruction of antibody clonal lineages in immune responses to viral pathogens, mainly HIV-1, which was not possible previously. These remarkable studies have revealed how antibodies evolve to penetrate the glycan shield of HIV-1 Env and how the virus in turn evolves to escape neutralization. New insights into the coevolution of viruses and antiviral antibodies will come from NGS of both Env variants and antibodies targeting these variants in well-documented longitudinal cohorts of African HIV-1 patients, as demonstrated recently (18, 53, 54). Although HIV-1 has so far been the primary focus of studies using NGS and single-cell analysis to identify germ line progenitors and intermediates along antibody maturation pathways, the future application of these methods to other pathogens, such as dengue virus and Ebola virus, will undoubtedly uncover new structural strategies for generating high-affinity binders to bolster host immune defense.

It is becoming increasingly evident that the contribution of somatic hypermutation to bNAb development must be considered in designing vaccine immunogens for different viral pathogens. At one extreme, bNAbs against some viruses, such as MERS-CoV (49) and hepatitis C virus (61), appear to exist naturally with relatively few somatic mutations. Importantly, somatic mutations were found not to be required for binding of germ line ancestors of these bNAbs to their viral targets, which may facilitate elicitation of effective vaccine responses. At the other extreme, bNAbs against viruses such as HIV-1 are highly somatically mutated and generally require years to develop in infected individuals. Furthermore, germ line precursors of these bNAbs often exhibit little or no detectable affinity for HIV-1 Env, making elicitation of such antibodies by vaccination a formidable challenge (62). Current strategies to overcome this roadblock involve initial activation of naïve mature B cells expressing germ line B cell receptors with engineered germ line-binding immunogens, followed by sequential vaccinations with immunogens designed to bind intermediate antibodies in order to guide the immune system through complex maturation pathways that ultimately lead to antibodies with specific somatic mutations conferring high affinity and neutralization potency (63–66).

AUTHOR CONTRIBUTIONS

AM and RM conceived and wrote the manuscript.

FUNDING

This study was supported by National Institutes of Health grant AI132213 to RM.

REFERENCES

1. Rajewsky K. Clonal selection and learning in the antibody system. *Nature* (1996) 381:751–8. doi:10.1038/381751a0
2. Neuberger MS. Antibodies: a paradigm for the evolution of molecular recognition. *Biochem Soc Trans* (2002) 30:341–50. doi:10.1042/bst030a047
3. Sundberg EJ, Mariuzza RA. Molecular recognition in antigen-antibody complexes. *Adv Protein Chem* (2002) 61:119–60. doi:10.1016/S0065-3233(02)61004-6
4. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302:575–81. doi:10.1038/302575a0
5. Wardemann H, Busse CE. Novel approaches to analyze immunoglobulin repertoires. *Trends Immunol* (2017) 38:471–82. doi:10.1016/j.it.2017.05.003
6. Kepler TB, Wiehe K. Genetic and structural analyses of affinity maturation in the humoral response to HIV-1. *Immunol Rev* (2017) 275:129–44. doi:10.1111/imr.12513
7. Liao HX, Bonsignori M, Alam SM, McLellan JS, Tomaras GD, Moody MA, et al. Vaccine induction of antibodies against a structurally heterogeneous site of immune pressure within HIV-1 envelope protein variable regions 1 and 2. *Immunity* (2013) 38:176–86. doi:10.1016/j.immuni.2012.11.011
8. Nicely NI, Wiehe K, Kepler TB, Jaeger FH, Dennison SM, Rerks-Ngarm S, et al. Structural analysis of the unmutated ancestor of the HIV-1 envelope V2 region antibody CH58 isolated from an RV144 vaccine efficacy trial vaccinee. *EBioMedicine* (2015) 2:713–22. doi:10.1016/j.ebiom.2015.06.016
9. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* (2013) 39:245–58. doi:10.1016/j.immuni.2013.04.012
10. Scharf L, West AP, Sievers SA, Chen C, Jiang S, Gao H, et al. Structural basis for germline antibody recognition of HIV-1 immunogens. *Elife* (2016) 5:1–24. doi:10.7554/eLife.13783
11. Jardine J, Julien JP, Menis S, Ota T, Kalyuzhniy O, McGuire A, et al. Rational HIV immunogen design to target specific germline B cell receptors. *Science* (2013) 340:711–6. doi:10.1126/science.1234150
12. Davenport TM, Gorman J, Joyce MG, Zhou T, Soto C, Guttman M, et al. Somatic hypermutation-induced changes in the structure and dynamics of HIV-1 broadly neutralizing antibodies. *Structure* (2016) 24:1346–57. doi:10.1016/j.str.2016.06.012
13. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* (2013) 496:469–76. doi:10.1038/nature12053
14. Fera D, Schmidt AG, Haynes BF, Gao F, Liao HX, Kepler TB, et al. Affinity maturation in a broadly neutralizing B-cell lineage through reorientation of variable domains. *Proc Natl Acad Sci U S A* (2014) 111:10275–80. doi:10.1073/pnas.1409954111
15. Julien JP, Cupo A, Sok D, Stanfield RL, Lyumkis D, Deller MC, et al. Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* (2013) 342:1477–83. doi:10.1126/science.1245625
16. Garces F, Sok D, Kong L, McBride R, Kim HJ, Saye-Francisco KF, et al. Structural evolution of glycan recognition by a family of potent HIV antibodies. *Cell* (2014) 159:69–79. doi:10.1016/j.cell.2014.09.009
17. Garces F, Lee JH, de Val N, de la Pena AT, Kong L, Puchades C, et al. Affinity maturation of a potent family of HIV antibodies is primarily focused

on accommodating or avoiding glycans. *Immunity* (2015) 43:1053–63. doi:10.1016/j.immuni.2015.11.007
18. Landais E, Murrell B, Briney B, Murrell S, Rantalainen K, Berndsen ZT, et al. HIV envelope glycoform heterogeneity and localized diversity govern the initiation and maturation of a V2 apex broadly neutralizing antibody lineage. *Immunity* (2017) 47:990–1003. doi:10.1016/j.immuni.2017.11.002
19. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* (2014) 509:55–62. doi:10.1038/nature13036
20. Scharf L, Wang H, Gao H, Chen S, McDowall AW, Bjorkman PJ. Broadly neutralizing antibody 8ANC195 recognizes closed and open states of HIV-1 Env. *Cell* (2015) 162:1379–90. doi:10.1016/j.cell.2015.08.035
21. Schmidt AG, Xu H, Khan AR, O'Donnell T, Khurana S, King LR, et al. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proc Natl Acad Sci U S A* (2013) 110:264–9. doi:10.1073/pnas.1218256109
22. Joyce MG, Wheatley AK, Thomas PV, Chuang GY, Soto C, Bailer RT, et al. Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. *Cell* (2016) 166:609–23. doi:10.1016/j.cell.2016.06.043
23. Alzari PM, Spinelli S, Mariuzza RA, Boulot G, Poljak RJ, Jarvis JM, et al. Three-dimensional structure determination of an anti-2-phenyloxazolone antibody: the role of somatic mutation and heavy/light chain pairing in the maturation of an immune response. *EMBO J* (1990) 9:3807–14.
24. Mizutani R, Miura K, Nakayama T, Shimada I, Arata Y, Satow Y. Three-dimensional structures of the Fab fragment of murine N1G9 antibody from the primary immune response and of its complex with (4-hydroxy-3-nitrophenyl)acetate. *J Mol Biol* (1995) 254:208–22. doi:10.1006/jmbi.1995.0612
25. Yuhasz SC, Parry C, Strand M, Amzel LM. Structural analysis of affinity maturation: the three-dimensional structures of complexes of an anti-nitrophenol antibody. *Mol Immunol* (1995) 32:1143–55. doi:10.1016/0161-5890(95)00063-1
26. Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC. Structural insights into the evolution of an antibody combining site. *Science* (1997) 276:1665–9. doi:10.1126/science.276.5319.1665
27. Furukawa K, Akasako-Furukawa A, Shirai H, Nakamura H, Azuma T. Junctional amino acids determine the maturation pathway of an antibody. *Immunity* (1999) 11:329–38. doi:10.1016/S1074-7613(00)80108-9
28. Manivel V, Sahoo NC, Salunke DM, Rao KV. Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity* (2000) 13:611–20. doi:10.1016/S1074-7613(00)00061-3
29. Wang F, Sen S, Zhang Y, Ahmad I, Zhu X, Wilson IA, et al. Somatic hypermutation maintains antibody thermodynamic stability during affinity maturation. *Proc Natl Acad Sci U S A* (2013) 110:4261–6. doi:10.1073/pnas.1301810110
30. Li Y, Li H, Yang F, Smith-Gill SJ, Mariuzza RA. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Biol* (2003) 10:482–8. doi:10.1038/nsb930
31. Tomlinson IM, Walter G, Jones PT, Dear PH, Sonnhammer EL, Winter G. The imprint of somatic hypermutation on the repertoire of human germline V genes. *J Mol Biol* (1996) 256:813–7. doi:10.1006/jmbi.1996.0127
32. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113:E2636–45. doi:10.1073/pnas.1525510113
33. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science* (2003) 301:1374–7. doi:10.1126/science.1086907
34. Trück J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J Immunol* (2015) 194:252–61. doi:10.4049/jimmunol.1401405
35. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) 13:691–700. doi:10.1016/j.chom.2013.05.008
36. Tan YC, Blum LK, Kongpachith S, Ju CH, Cai X, Lindstrom TM, et al. High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin Immunol* (2014) 151:55–65. doi:10.1016/j.clim.2013.12.008
37. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res* (2013) 2:103. doi:10.12688/f1000research.2-103.v1
38. Alam SM, Liao HX, Dennison SM, Jaeger F, Parks R, Anasti K, et al. Differential reactivity of germ line allelic variants of a broadly neutralizing HIV-1 antibody to a gp41 fusion intermediate conformation. *J Virol* (2011) 85:11725–31. doi:10.1128/JVI.05680-11
39. Bonsignori M, Hwang KK, Chen X, Tsao CY, Morris L, Gray E, et al. Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *J Virol* (2011) 85:9998–10009. doi:10.1128/JVI.05045-11
40. Dal Porto JM, Haberman AM, Kelsoe G, Shlomchik MJ. Very low affinity B cells form germinal centers, become memory B cells, and participate in secondary immune responses when higher affinity competition is reduced. *J Exp Med* (2002) 195:1215–21. doi:10.1084/jem.20011550
41. Shih TA, Meffre E, Roederer M, Nussenzweig MC. Role of BCR affinity in T cell dependent antibody responses in vivo. *Nat Immunol* (2002) 3:570–5. doi:10.1038/ni803
42. Hoot S, McGuire AT, Cohen KW, Strong RK, Hangartner L, Klein F, et al. Recombinant HIV envelope proteins fail to engage germline versions of anti-CD4bs bNAbs. *PLoS Pathog* (2013) 9:e1003106. doi:10.1371/journal.ppat.1003106
43. Xiao X, Chen W, Feng Y, Zhu Z, Prabakaran P, Wang Y, et al. 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* (2009) 390:404–9. doi:10.1016/j.bbrc.2009.09.029
44. Huang J, Zarnitsyna VI, Liu B, Edwards LJ, Jiang N, Evavold BD, et al. The kinetics of two-dimensional TCR and pMHC interactions determine T-cell responsiveness. *Nature* (2010) 464:932–6. doi:10.1038/nature08944
45. Wang XX, Li Y, Yin Y, Mo M, Wang Q, Gao W, et al. Affinity maturation of human CD4 by yeast surface display and crystal structure of a CD4-HLA-DR1 complex. *Proc Natl Acad Sci U S A* (2011) 108:15960–5. doi:10.1073/pnas.1109438108
46. Jönsson P, Southcombe JH, Santos AM, Huo J, Fernandes RA, McColl J, et al. Remarkably low affinity of CD4/peptide-major histocompatibility complex class II protein interactions. *Proc Natl Acad Sci U S A* (2016) 113:5682–7. doi:10.1073/pnas.1513918113
47. Lingwood D, McTamney PM, Yassine HM, Whittle JR, Guo X, Boyington JC, et al. Structural and genetic basis for development of broadly neutralizing influenza antibodies. *Nature* (2012) 489:566–70. doi:10.1038/nature11371
48. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. *Science* (2003) 299:1362–7. doi:10.1126/science.1079731
49. Ying T, Prabakaran P, Du L, Shi W, Feng Y, Wang Y, et al. Junctional and allele-specific residues are critical for MERS-CoV neutralization by an exceptionally potent germline-like antibody. *Nat Commun* (2015) 6:8223. doi:10.1038/ncomms9223
50. Koenig P, Lee CV, Walters BT, Janakiraman V, Stinson J, Patapoff TW, et al. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proc Natl Acad Sci U S A* (2017) 114:E486–95. doi:10.1073/pnas.1613231114
51. Ward AB, Wilson IA. The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunol Rev* (2017) 275:21–32. doi:10.1111/imr.12507
52. MacLeod DT, Choi NM, Briney B, Garces F, Ver LS, Landais E, et al. Early antibody lineage diversification and independent limb maturation lead to broad HIV-1 neutralization targeting the Env high-mannose patch. *Immunity* (2016) 44:1215–26. doi:10.1016/j.immuni.2016.04.016
53. Bhiman JN, Anthony C, Doria-Rose NA, Karimanzira O, Schramm CA, Khoza T, et al. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nat Med* (2015) 21:1332–6. doi:10.1038/nm.3963
54. Andrabi R, Su CY, Liang CH, Shivatare SS, Briney B, Voss JE, et al. Glycans function as anchors for antibodies and help drive HIV broadly neutralizing antibody development. *Immunity* (2017) 47:524–37. doi:10.1016/j.immuni.2017.08.006
55. Sterner E, Peach ML, Nicklaus MC, Gildersleeve JC. Therapeutic antibodies to ganglioside GD2 evolved from highly selective germline antibodies. *Cell Rep* (2017) 20:1681–91. doi:10.1016/j.celrep.2017.07.050
56. Ang CW, Jacobs BC, Brandenburg AH, Laman JD, van der Meché FG, Osterhaus AD, et al. Cross-reactive antibodies against GM2 and CMV-infected fibroblasts in Guillain-Barré syndrome. *Neurology* (2000) 54:1453–8. doi:10.1212/WNL.54.7.1453

57. Fukita Y, Jacobs H, Rajewsky K. Somatic hypermutation in the heavy chain locus correlates with transcription. *Immunity* (1998) 9:105–14. doi:10.1016/S1074-7613(00)80592-0
58. Briney BS, Willis JR, Crowe JE Jr. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun* (2012) 13:523–9. doi:10.1038/gene.2012.28
59. Kepler TB, Liao HX, Alam SM, Bhaskarabhatla R, Zhang R, Yandava C, et al. Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. *Cell Host Microbe* (2014) 16:304–13. doi:10.1016/j.chom.2014.08.006
60. Huang J, Kang BH, Pancera M, Lee JH, Tong T, Feng Y, et al. Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-gp120 interface. *Nature* (2014) 515:138–43. doi:10.1038/nature13601
61. Bailey JR, Flyak AI, Cohen VJ, Li H, Wasilewski LN, Snider AE, et al. Broadly neutralizing antibodies with few somatic mutations and hepatitis C virus clearance. *JCI Insight* (2017) 2:92872. doi:10.1172/jci.insight.92872
62. Haynes BF, Kelsoe G, Harrison SC, Kepler TB. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat Biotechnol* (2012) 30:423–33. doi:10.1038/nbt.2197
63. Briney B, Sok D, Jardine JG, Kulp DW, Skog P, Menis S, et al. Tailored immunogens direct affinity maturation toward HIV neutralizing antibodies. *Cell* (2016) 166(6):1459–70.e11. doi:10.1016/j.cell.2016.08.005
64. Escolano A, Steichen JM, Dosenovic P, Kulp DW, Golijanin J, Sok D, et al. Sequential immunization elicits broadly neutralizing anti-HIV-1 antibodies in Ig knockin mice. *Cell* (2016) 166:1445–8. doi:10.1016/j.cell.2016.07.030
65. Jardine JG, Kulp DW, Havenar-Daughton C, Sarkar A, Briney B, Sok D, et al. HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science* (2016) 351:1458–63. doi:10.1126/science.aad9195
66. Steichen JM, Kulp DW, Tokatlian T, Escolano A, Dosenovic P, Stanfield RL, et al. HIV vaccine design to target germline precursors of glycan-dependent broadly neutralizing antibodies. *Immunity* (2016) 45:483–96. doi:10.1016/j.immuni.2016.08.016

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.

Copyright © 2018 Mishra and Mariuzza. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

MINI REVIEW
published: 02 February 2018
doi: 10.3389/fimmu.2018.00118

Next-Generation Sequencing of Antibody Display Repertoires

Romain Rouet[1], Katherine J. L. Jackson[1], David B. Langley[1] and Daniel Christ[1,2]*

[1] *Garvan Institute of Medical Research, Sydney, NSW, Australia,* [2] *Faculty of Medicine, St Vincent's Clinical School, The University of New South Wales, Sydney, NSW, Australia*

In vitro selection technology has transformed the development of therapeutic monoclonal antibodies. Using methods such as phage, ribosome, and yeast display, high affinity binders can be selected from diverse repertoires. Here, we review strategies for the next-generation sequencing (NGS) of phage- and other antibody-display libraries, as well as NGS platforms and analysis tools. Moreover, we discuss recent examples relating to the use of NGS to assess library diversity, clonal enrichment, and affinity maturation.

Keywords: antibody display technology, next-generation sequencing, phage display, antibody libraries, *in vitro* selection, antibody therapeutics

OPEN ACCESS

Edited by:
Prabakaran Ponraj, Intrexon, United States

Reviewed by:
Yang Feng, National Cancer Institute at Frederick, United States
Masaki Hikida, Kyoto University, Japan

****Correspondence:***
Romain Rouet
r.rouet@garvan.org.au

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *27 November 2017*
Accepted: *15 January 2018*
Published: *02 February 2018*

Citation:
Rouet R, Jackson KJL, Langley DB and Christ D (2018) Next-Generation Sequencing of Antibody Display Repertoires. Front. Immunol. 9:118. doi: 10.3389/fimmu.2018.00118

INTRODUCTION

The development of antibody display technology such as phage (1), ribosome (2), yeast (3), and mammalian display (4) has enabled the rapid selection of binders from diverse libraries. These technologies bypass the use of animals and allow for the enrichment of binders within days to weeks. The power of *in vitro* selection technologies relies on a direct physical link between phenotype (displayed antibody construct) and genotype (antibody variable domain genes), allowing for the identification of binders through sequencing of their encoding genes. Multiple rounds of selection are generally required to identify antigen-specific binders, either by binding to a solid support or through cellular sorting (5). In many cases, later rounds of selections tend to be dominated by a handful of clones, which are then further characterized for affinity. Such clonal dominance can reflect genuine selection for high antigen affinity but might also reflect other properties such as superior expression or display. Consequently, clones with superior affinity may be present at low frequency and may not be readily detectable using traditional screening methods such as ELISA (6).

Recent advances in DNA sequencing technologies and computing power over the last decade has led to a dramatic reduction in the cost of sequencing and has simplified data analyses (7). Although initially developed for genomics applications, such as whole-genome sequencing, transcriptome sequencing, and epigenetics, next-generation sequencing (NGS) technology is now increasingly being applied other fields, including to basic and applied immunology. This includes the sequencing of the paired human heavy and light chain repertoire from isolated naïve (8, 9) and antigen-specific B-cells (10, 11), as well as T-cell receptor (12) and antibody display repertoires (13). While most NGS platforms were originally designed for short reads, technology is evolving rapidly, extending both read length and depth. Here, we review recent advances in NGS technology and key applications to phage display and other *in vitro* selection technologies.

Strategies for NGS of Antibody Repertoires

Traditionally, antibody display libraries are analyzed by isolation of 10^2–10^3 clones in combination with Sanger sequencing (5). Although this approach is sufficient to identify dominant clones after selection, or to broadly validate design objectives, the data obtained represent only a limited snapshot of actual library diversity. By contrast, NGS approaches allow for far-greater insights into library diversity by providing up to 10^7 sequences (approximately 10,000-fold more sequences than Sanger sequencing).

One of the main challenges in the use of NGS for the analysis of antibody selection systems relates to the size of the encoded genes: the smallest antibody fragments (variable domains) range between 300 and 400 bp in length, while the commonly used scFv and Fab antibody fragment formats range from 700 to 800 bp to over 1,500 bp, respectively. While NGS technologies are particularly well suited for high throughput sequencing of short reads (less than 100 bp), many platforms can nevertheless sequence up to 300–400 bp with reasonable throughput. In particular, Illumina Miseq and Hiseq, 454 GS FLX (instrument discontinued), and Ion Torrent PMG are suited for this task (8, 14, 15); in addition, PacBio sequencing generates particularly long reads at the cost of reduced read numbers (**Table 1**) (16). Long sequences can also be generated by using paired-end reads: this method is particularly useful for scFv formats, enabling the sequencing of multiple CDR regions of V_H and V_L domains. In addition to analysis of longer antibody fragment sequences, some studies have focused on sequencing the relatively short V_H CDR3 repertoire only (23) [which forms the center of the antigen binding site and is a major determinant of antigen binding (24)].

The use of NGS requires particular attention be paid to sequencing errors (25). DNA amplification inevitably results in polymerase errors, which can be context dependent. Although the error rates of polymerases are generally low (10^{-5}–10^{-6} per base), errors will inevitably be present in large NGS datasets that encompass billions of bases. In addition, the NGS technologies themselves can be susceptible to the introduction of errors, such as cluster misamplification and base misincorporation, with frequencies ranging from 10^{-2} (PacBio, Ion Torrent) to 10^{-3} (Illumina). To help identify PCR and sequencing errors, unique molecular identifiers (UMIs; stretches of 8–10 degenerate DNA bases) can be added to primers during the first two cycles of PCR amplification. Reads that share the same UMI have a high probability of being derived from the same original template. Such reads can then grouped after sequencing and used for error correction (26).

Bioinformatic Tools to Analyze NGS Data

While the analysis of the limited number of clones obtained by Sanger sequencing can be carried out manually, the larger sample size of NGS approaches necessitates the use bioinformatics tools. Following confirmation of the quality of the NGS read data by a tool such as FastQC (27), the data are further processed to clean up reads before analysis of antibody or antibody fragment sequences. The steps undertaken will be highly dependent on the NGS platform utilized and the format of the amplicons but generally will focus on the: removal of adapter sequences [e.g., PRINSEQ (28)], de-multiplexing (if barcodes were used), UMI identification and consensus building, and error correction (26), read quality trimming and filtering [e.g., Trimmomatic (29)] and, if paired-end sequencing was performed, the merging of the read pairs with a program such as PEAR (30). Separate analysis of heavy and light chains may be required for antibody formats such as scFv, where the presence of synthetic linkers can complicate analyses.

Programs such as IMPre (31), IgBLAST (32), IMGT/High V-QUEST (33, 34), and ImmundiveRsity (35), which were originally developed for the analysis of B and T cell receptor repertoires, identify V_H and V_L germlines as well as V_H and V_L CDRs. The selection of a tool will depend on the number of NGS reads being analyzed and the computational skill level of the researcher. IgBLAST and IMGT/High V-QUEST are both available as web-based submission systems, with IMGT/High V-QUEST permitting a larger number of reads to be analysis per submission. IMGT/High V-QUEST returns an output format compatible with programs such as Microsoft Excel or OpenOffice, whereas IgBLAST output is text based. The tools use different alignment algorithms, BLAST (IgBLAST) and modified Smith–Waterman (V-QUEST), but both restrict the germline gene repertoires to those defined by the tool's creators. A stand-alone version of IgBLAST is available, and it has no restriction on the number of input reads, permits the user-defined germline gene databases, provides additional output formats, and can be parallelized on a cluster for processing of large datasets; however, its use does require some command line basics.

Postprocessing of the output of tools such as IgBLAST and IMGT/High V-QUEST is required to generate information about the clone structure within the dataset, and to pair V_H and V_L domain sequences. Clone structures can be inferred by applying sequence clustering tools, such as CD-hit (36) or UCLUST (37) to CDR3s alone, at either the amino acid or nucleotide sequence level, or to the full-length sequence, to group closely related sequences into "clonal" groups. The choice of parameters will depend on the diversity of the library. Finally, scripts can be used to analyze and summarize the diversity and other compositional characteristics of the library.

A custom pipeline as described earlier requires a level of informatics skills not always available to researchers, therefore, specialized pipelines for the analysis of recombinant antibody libraries, either naïve or *in vitro* selected against particular antigens, have been developed. The AbMining Toolbox is particularly suited for identifying V_H CDR3, which is determined by using a hidden Markov model (HMM) that captures the conserved sequences upstream and downstream of the CDR3 (38). N^2GSAb can rapidly identify germline and V_H CDR3 and provides a tool

TABLE 1 | Next-generation sequencing platforms for the analysis of display libraries.

Platform	Read length	Max. depth	Error type (percentage)	Reference
Illumina Miseq	300 bp PE	40×10^6 reads	Substitutions (~0.1)	(15, 17)
Illumina Hiseq 2500	250 bp PE	600×10^6 reads	Substitutions (~0.1)	(18, 19)
Ion Torrent PMG	400 bp	5.5×10^6 reads	Indels (~1)	(14, 20)
454 GS FLX	Up to 1 kb	1×10^6 reads	Indels (~1)	(13, 21)
PacBio	250 bp–40 kb	0.4×10^6 reads	Indels (~1)	(22)

for clustering unique sequences (39). VDJFasta uses an HMM to accurately predict all V_H and V_L CDRs, as well as the GS linker sequence for scFv fragments, and can generate library diversity plots (13). The ImmuneDB package aligns sequences based on a query sequence, such as a framework region, to delineate CDR regions (40). ImmuneDB also performs mutational and statistical analysis on the sequence library and can construct lineage trees to aid in the interpretation of antigen-selected libraries. More recently, DEAL was developed to better predict library diversity by identifying and correcting sequencing errors (17). In the published example, the library was not generated by PCR but rather by ligation of adapters to avoid any amplification bias and focus on sequencing errors. Reads are clustered using seed sequences of 10–20 bp and analyzed by binary comparison. The clusters are then compared with each read, taking into account the Phred quality score for each base and the error rate of the Phi-X control to identify sequencing errors. A list of software is outlined in **Table 2**.

It is also possible to outsource the sequencing and/or analysis of antibody libraries to commercial suppliers. Examples include (but are not limited to) CD Genomics and Molecular Cloning Laboratories. Such companies offer a range of options from basic consulting on designing primers for multiplexing and sequencing, to complete analysis from purified DNA or phages.

Application to Design Validation and the Analyses of Naïve Antibody Libraries

When generating antibody display repertoires, either synthetic or derived from immunized animals, it is important to assess the clonal diversity of the library before selection. Several studies have demonstrated the use of NGS to measure diversity to validate the design of displayed libraries. In an early example, Novimmune designed scFv libraries with both synthetic diversity, using degenerated oligonucleotides, and semi-synthetic diversity, from human or rabbit donors (39). Sequencing of V_H CDR3 using the Illumina platform revealed that the synthetic libraries had many more unique clones compared with donor-derived libraries, with between 1–16 and 31–69% clonal redundancy, respectively (39). Intriguingly, the extent of clonal redundancy in the donor-derived libraries suggested an upper limit of human V_H diversity of around $2–3 \times 10^6$ unique clones. This figure correlates with other NGS studies aimed at determining human B-cell diversity ($3–9 \times 10^6$) (41). In addition, the Novimmune sequencing results also validated the V_H CDR3 length distribution in human antibodies, which closely matched that of the IMGT repertoire (39).

A second example, relates to the Ylanthia synthetic antibody library developed by MorphoSys (21). The library was designed to encode a range of V_H CDR3 lengths to closely match the natural human antibody repertoire and was analyzed by the Roche 454 sequencing platform (21). The library was found to be composed of about 95% unique clones, and there was no indication of amplification biases during antibody library construction. In addition, the authors used NGS data to validate V_H CDR3 diversity and length, as well as V_H and V_L germline frequencies.

High throughput sequencing approaches are not limited to human sequences, with a recent study assessing the diversity of rabbit (V_H and V_L) Fab libraries by NGS (Ion PGM) (20). Surprisingly, and unlike human libraries derived from donors, these studies detected very low levels of redundancy within the rabbit libraries, with over 98% of V_H clones being of a unique nature (~3×10^9 sequence reads were analyzed).

Next-generation sequencing has also been used to accurately determine library size. A recent study generated a donor-derived V_H library for this purpose, which was then sequenced using Illumina adapter ligation (circumventing the need for PCR amplification) (17). Sequencing depth for the V_H library exceeded the library size by three-fold suggesting that the diversity was well represented in the NGS output. The authors estimated the minimal functional diversity to be 1.2×10^6 individual unique clones representing just one-fifth of the original number of bacterial clones.

TABLE 2 | Software for next-generation sequencing analysis.

Software package	Strengths	Reference
IMGT/High V-Quest	Fast germline identification, CDR determination, and batch submission	(34)
ImmunediveRsity	Quality filtering and noise correction	(35)
VDJFasta	Hidden Markov model to determine all CDRs and frequency analysis, very rapid analysis	(13)
N^2GSAb	Rapid germline and V_H CDR3 determination and sequence clustering	(39)
ImmuneDB	Alignment based on sequence query to determine CDRs and frequency	(40)
DEAL	Sequencing error correction before analysis	(17)

Application to Affinity Maturation and Epitope Mapping

Next-generation sequencing can also be used to guide selection toward high affinity clones. For example, one seminal study employed NGS to guide maturation of an scFv fragment directed against ErbB2 to a final affinity of 25 pM (resulting in a 158-fold improvement over wild type) (18). Guided by structure-based design, individual CDR regions (excluding V_L CDR2) were randomized, selected against ErbB2 antigen, and analyzed by NGS before and after panning. This revealed enrichment of novel sequence motifs at diversified CDR positions, with the exception of V_H CDR3, which was enriched toward the wild-type motif (suggesting an already optimal sequence). Next, the most frequent CDR substitutions were combined to generate a secondary library (V_H CDR3 being reverted to wild type), which was selected against the target. This resulted in improved affinities of between 300 and 25 pM, compared with the wild-type affinity of 4 nM, highlighting the power of this stepwise approach for affinity maturation.

In a further study, deep mutational scanning analysis using NGS was performed on a humanized version of the anti-EGFR monoclonal cetuximab (42). More specifically, independent V_H and V_L libraries (encoding over 1,000 single amino acid substitutions at 59 different positions—32 in V_H and 27 in V_L) were selected by mammalian cell display and flow cytometry.

Gated populations were analyzed by NGS to identify permissive mutations and to generate a heat map of the antigen binding site. Overall, this strategy identified 67 substitutions that increased affinity, including one mutation with a five-fold K_D improvement. Similar strategies can also be used to map epitope surfaces, as exemplified by the interaction of *S. aureus* toxin with neutralizing antibodies (43).

CONCLUSION

Next-generation sequencing holds great promise for the development of therapeutic monoclonal antibodies, by allowing unprecedented insights into library diversity and clonal enrichment. Although current NGS platforms were not designed with antibody libraries in mind, the technologies are now at a stage where unique sequence insights into all stages of the selection process can be obtained. Moreover, with ongoing advances in sequencing technology, depth and read length is improving continuously: for instance, the PacBio Sequel system generates approximately seven times more sequences than the previous RS II system but maintains its long-read capability (Pacific Biosciences), while nanopore systems such as the MinIOn (Oxford Nanopore) offer the promise of real-time DNA sequencing in combination with ultra-long reads. We conclude that, with NGS technology evolving at a rapid pace, its importance in the sequence analyses of phage- and other antibody-display libraries is likely to continue to increase.

AUTHOR CONTRIBUTIONS

RR wrote the manuscript. KJ, DL, and DC edited the manuscript.

FUNDING

This work was supported by the Australian National Health and Medical Research Council (APP1090875, APP1148051, APP11 13904, APP1113790) and the Australian Research Council (DP16010491).

REFERENCES

1. Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* (1985) 228:1315–7. doi:10.1126/science.4001944
2. Hanes J, Pluckthun A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* (1997) 94:4937–42. doi:10.1073/pnas.94.10.4937
3. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* (1997) 15:553–7. doi:10.1038/nbt0697-553
4. Ho M, Nagata S, Pastan I. Isolation of anti-CD22 Fv with high affinity by Fv display on human cells. *Proc Natl Acad Sci U S A* (2006) 103:9637–42. doi:10.1073/pnas.0603653103
5. Lee CM, Iorno N, Sierro F, Christ D. Selection of human antibody fragments by phage display. *Nat Protoc* (2007) 2:3001–8. doi:10.1038/nprot.2007.448
6. Rouet R, Lowe D, Dudgeon K, Roome B, Schofield P, Langley D, et al. Expression of high-affinity human antibody fragments in bacteria. *Nat Protoc* (2012) 7:364–73. doi:10.1038/nprot.2011.448
7. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* (2016) 17:53. doi:10.1186/s13059-016-0917-0
8. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31:166–9. doi:10.1038/nbt.2492
9. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2015) 21:86–91. doi:10.1038/nm.3743
10. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* (2011) 333:1593–602. doi:10.1126/science.1207532
11. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* (2014) 509:55–62. doi:10.1038/nature13036
12. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* (2009) 114:4099–107. doi:10.1182/blood-2009-04-217604
13. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106:20216–21. doi:10.1073/pnas.0909775106
14. Moutel S, Bery N, Bernard V, Keller L, Lemesre E, de Marco A, et al. NaLi-H1: A universal synthetic library of humanized nanobodies providing highly functional antibodies and intrabodies. *Elife* (2016) 5:e16228. doi:10.7554/eLife.16228
15. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, et al. By-passing in vitro screening – next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* (2010) 38:e193. doi:10.1093/nar/gkq789
16. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* (2016) 17:333–51. doi:10.1038/nrg.2016.49
17. Fantini M, Pandolfini L, Lisi S, Chirichella M, Arisi I, Terrigno M, et al. Assessment of antibody library diversity through next generation sequencing and technical error compensation. *PLoS One* (2017) 12:e0177574. doi:10.1371/journal.pone.0177574
18. Hu D, Hu S, Wan W, Xu M, Du R, Zhao W, et al. Effective optimization of antibody affinity by phage display integrated with high-throughput DNA synthesis and sequencing technologies. *PLoS One* (2015) 10:e0129125. doi:10.1371/journal.pone.0129125
19. Larman HB, Xu GJ, Pavlova NN, Elledge SJ. Construction of a rationally designed antibody platform for sequencing-assisted selection. *Proc Natl Acad Sci U S A* (2012) 109:18523–8. doi:10.1073/pnas.1215549109
20. Peng H, Nerreter T, Chang J, Qi J, Li X, Karunadharma P, et al. Mining naive rabbit antibody repertoires by phage display for monoclonal antibodies of therapeutic utility. *J Mol Biol* (2017) 429(19):2954–73. doi:10.1016/j.jmb.2017.08.003
21. Tiller T, Schuster I, Deppe D, Siegers K, Strohner R, Herrmann T, et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *MAbs* (2013) 5:445–70. doi:10.4161/mabs.24218
22. Vollmers C, Penland L, Kanbar JN, Quake SR. Novel exons and splice variants in the human antibody heavy chain identified by single cell and single molecule sequencing. *PLoS One* (2015) 10:e0117050. doi:10.1371/journal.pone.0117050
23. Weinstein JA, Jiang N, White RA III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science* (2009) 324:807–10. doi:10.1126/science.1170020
24. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* (2000) 13:37–45. doi:10.1016/S1074-7613(00)00006-6
25. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of next generation sequencing platforms. *Next Gener Seq Appl* (2014) 1. doi:10.4172/jngsa.1000106
26. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* (2017) 27:491–9. doi:10.1101/gr.209601.116

27. S. Andrews. FastQC: a quality control tool for high throughput sequencing data (2010). Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
28. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* (2011) 27:863–4. doi:10.1093/bioinformatics/btr026
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) 30:2114–20. doi:10.1093/bioinformatics/btu170
30. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate illumina paired-end reAd mergeR. *Bioinformatics* (2014) 30:614–20. doi:10.1093/bioinformatics/btt593
31. Zhang W, Wang IM, Wang C, Lin L, Chai X, Wu J, et al. IMPre: an accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol* (2016) 7:457. doi:10.3389/fimmu.2016.00457
32. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41:W34–40. doi:10.1093/nar/gkt382
33. Brochet X, Lefranc MP, Giudicelli V. 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* (2008) 36:W503–8. doi:10.1093/nar/gkn316
34. Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) 4:2333. doi:10.1038/ncomms3333
35. Cortina-Ceballos B, Godoy-Lozano EE, Samano-Sanchez H, Aguilar-Salgado A, Velasco-Herrera Mdel C, Vargas-Chavez C, et al. Reconstructing and mining the B cell repertoire with immunediversity. *MAbs* (2015) 7:516–24. doi:10.1080/19420862.2015.1026502
36. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* (2006) 22:1658–9. doi:10.1093/bioinformatics/btl158
37. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* (2010) 26:2460–1. doi:10.1093/bioinformatics/btq461
38. D'Angelo S, Glanville J, Ferrara F, Naranjo L, Gleasner CD, Shen X, et al. The antibody mining toolbox: an open source tool for the rapid analysis of antibody repertoires. *MAbs* (2014) 6:160–72. doi:10.4161/mabs.27105
39. Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods* (2013) 60:99–110. doi:10.1016/j.ymeth.2013.03.001
40. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. ImmuneDB: a system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* (2017) 33:292–3. doi:10.1093/bioinformatics/btw593
41. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* (2011) 6:e22365. doi:10.1371/journal.pone.0022365
42. Forsyth CM, Juan V, Akamatsu Y, DuBridge RB, Doan M, Ivanov AV, et al. Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs* (2013) 5:523–32. doi:10.4161/mabs.24979
43. Van Blarcom T, Rossi A, Foletti D, Sundar P, Pitts S, Bee C, et al. Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing. *J Mol Biol* (2015) 427:1513–34. doi:10.1016/j.jmb.2014.09.020

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.

Copyright © 2018 Rouet, Jackson, Langley and Christ. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 05 February 2018
doi: 10.3389/fimmu.2018.00128

In-Depth Analysis of Human Neonatal and Adult IgM Antibody Repertoires

*Binbin Hong[1,2], Yanling Wu[1], Wei Li[3], Xun Wang[4], Yumei Wen[1], Shibo Jiang[1], Dimiter S. Dimitrov[3] and Tianlei Ying[1]**

[1] Key Laboratory of Medical Molecular Virology of Ministries of Education and Health, School of Basic Medical Sciences, Fudan University, Shanghai, China, [2] Central Laboratory, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, China, [3] Protein Interactions Section, Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD, United States, [4] Shanghai Blood Center, WHO Collaborating Center for Blood Transfusion Services, Shanghai, China

Although high-throughput sequencing and associated bioinformatics technologies have enabled the in-depth, sequence-based characterization of human immune repertoires, only a few studies on a relatively small number of sequences explored the characteristics of antibody repertoires in neonates, with contradictory conclusions. To gain a more comprehensive understanding of the human IgM antibody repertoire, we performed Illumina sequencing and IMGT/HighV-QUEST analysis of IgM heavy chain repertoire of the B lymphocytes from the cord blood (CB) of neonates, as well as the repertoire from peripheral blood of healthy human adults (HH). The comparative study revealed unexpectedly high levels of similarity between the neonatal and adult repertoires. In both repertoires, the VDJ gene usage showed no significant difference, and the most frequently used VDJ gene was IGHV4-59, IGHD3-10, and IGHJ3. The average amino acid (aa) length of CDR1 (CB: 8.5, HH: 8.4) and CDR2 (CB: 7.6, HH: 7.5), as well as the aa composition and the average hydrophobicity of the CDR3 demonstrated no significant difference between the two repertories. However, the average aa length of CDR3 was longer in the HH repertoire than the CB repertoire (CB: 14.5, HH: 15.5). Besides, the frequencies of aa mutations in CDR1 (CB: 19.33%, HH: 25.84%) and CDR2 (CB: 9.26%, HH: 17.82%) were higher in the HH repertoire compared to the CB repertoire. Interestingly, the most prominent difference between the two repertoires was the occurrence of N2 addition (CB: 64.87%, HH: 85.69%), a process that occurs during V-D-J recombination for introducing random nucleotide additions between D- and J-gene segments. The antibody repertoire of healthy adults was more diverse than that of neonates largely due to the higher occurrence of N2 addition. These findings may lead to a better understanding of antibody development and evolution pathways and may have potential practical value for facilitating the generation of more effective antibody therapeutics and vaccines.

Keywords: high-throughput sequencing, antibody repertoire, cord blood, VDJ rearrangement, junctional modification, N2 addition

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Brandon DeKosky, University of Kansas, United States
Yong-Sung Kim, Ajou University, South Korea

****Correspondence:***
Tianlei Ying
tlying@fudan.edu.cn

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *21 September 2017*
Accepted: *16 January 2018*
Published: *05 February 2018*

Citation:
Hong B, Wu Y, Li W, Wang X, Wen Y, Jiang S, Dimitrov DS and Ying T (2018) In-Depth Analysis of Human Neonatal and Adult IgM Antibody Repertoires. Front. Immunol. 9:128. doi: 10.3389/fimmu.2018.00128

INTRODUCTION

High-throughput sequencing of antibody repertoire and related bioinformatics analysis are becoming increasingly important tools that allow unprecedented insight into the in-depth, sequence-based composition of human immune repertoires (1–3). Such information may lead to a better understanding of antibody development and evolution pathways and facilitate the generation of more effective antibody therapeutics and vaccines (4, 5). However, despite the extensive efforts over the past decade, our understanding of human antibody repertoire remains limited due to its two fundamental characteristics. First, the antibody repertoire of an individual is highly dynamic, which varies greatly not only in response to environmental (for example, infection) but also to intrinsic (for example, aging) factors. Furthermore, a thorough analysis of antibody repertoire has been hindered by its enormous diversity. There are three primary mechanisms contributing to the antibody repertoire diversity: the combinatorial diversity created by rearrangements of the variable (V), diversity (D), and joining (J) gene segments; the junctional diversity resulted from exonuclease trimmings and the random addition of nucleotides; and the somatic hypermutations that occur during the immunoglobulin synthesis. By these mechanisms, a virtually unlimited number of different antibodies could be achieved using a limited number of germline immunoglobulin genes (6, 7). Therefore, it could be technically challenging to analyze the highly dynamic and diverse human antibody repertoires.

Notably, the antibody repertoire of the fetus or umbilical cord blood (CB) represents a source of un-mutated or minimally mutated antibodies, thus providing a unique opportunity to gain a general understanding of the human antibody repertoire. Compared with adults, human neonates are believed to have a limited ability to generate effective antibodies because they have not been exposed to exogenous antigens and do not develop an effective immunological memory response to the antigens (8, 9). Accordingly, the fetal repertoire is more restricted than the adult repertoire (10). Several earlier studies have revealed the characteristics of the fetal and adult immune repertoires, including the preferential VDJ gene usages, somatic mutations, and the length of the CDR3, which vary in different periods of human fetal life and adulthood (11–14). In contrast, another study compared the repertoires of human CB and adult sources reconstituted by NOD/SCID/$\beta2m^{-/-}$ mice with human B-cell progenitors and found nearly identical IGHV and IGHJ gene segment usage and only modest differences in CDR3 of the antibody heavy chain (15). Such inconsistency may partly result from the relatively small size of the examined samples or the sequenced repertoires. Indeed, we estimate that there are at least 10^7 B lymphocytes in a typical CB sample (100–200 mL), whereas the 454 pyrosequencing technology, as used in most previous studies, was only able to generate roughly 10^5 reads per sample (16), and thus underpowered to evaluate the full scale of antibody repertoires.

Recently, the Illumina-based sequencing is becoming the dominant high-throughput sequencing strategy, enabling the acquisition of millions to billions of sequences in a single experiment. The greater sequencing depth allows comprehensive investigation of the human antibody repertoires with high diversity (17). In this study, we described the characterization of the IgM heavy chain (IgH) repertoires from the B lymphocytes of the CB of neonates and the peripheral blood of adults using the Illumina sequencing platform. Over 10^7 unique antibody clones were identified, but less than 1% of these unique clones were shared by both neonates and adults, indicating the extremely large diversity of human antibody repertoires. Interestingly, despite the difference in sequences, we found unexpectedly high levels of similarity between the neonatal and adult repertoires regarding the VDJ gene usage, the characteristics of CDRs, and the occurrence of certain junctional modifications. The IgH repertoire of healthy adults was more diverse than that of neonates, largely due to the higher occurrence of N2 addition, a process that occurs during V-D-J recombination for introducing random nucleotide additions between D- and J-gene segments. These findings suggest a critical and previously unrecognized role for antibody junctional modifications, especially N2 addition, in the development and evolution of antibody repertoires in healthy individuals.

MATERIALS AND METHODS

Samples

The CB samples from 10 newborn babies (4 boys and 6 girls) were provided by National Disease Research Interchange (NDRI, Philadelphia, PA, USA) with approval of institutional research board and donor consent. Care was taken not to contaminate the samples with maternal blood. The blood samples from healthy adults were collected from 33 healthy adults (16 females and 17 males; age range, 27–62 years; average age, 44.1 years), who underwent a routine health check with no history of known major diseases, with approval of institutional research board and donor consent. The basic characteristics of the study population were summarized in **Table 1**, and the detailed inclusion criteria were summarized in the Supplementary Materials.

Establishment of IGH Repertoires for Deep Sequencing

As the source for amplification of antibody sequences, cDNA was reverse transcripted from the total RNA extracted from lymphocytes and was prepared according to the reported protocols (18). PCR amplifications were applied to establish the IGH repertoire libraries. Primers used in PCR amplifications were highly specific to the N-terminal and C-terminal regions of the IgM-derived heavy chains as described previously (19). Briefly, PCR amplifications were performed with a mixture of primers in which the 3′-ends ligated to the first seven codons of IGHV1 to IGHV7 gene families, and PCR amplifications of the

TABLE 1 | Characteristics of the study population.

	Neonates	Adults
Gender (F/M)	6/4	16/17
Age	0 day	F: 43.8 ± 9.9 (years) M: 45.0 ± 9.4 (years)
Weight	3,379.6 ± 561.8 (g)	N/A

F, female; M, male; N/A, not available.

constant domains were performed by a sense primer specific for CH1 domain of IGHM spanning first eight codons (3′–5′ strand) according to the ImMunoGeneTics database (www.imgt.org). The PCR amplifications were performed again to produce shorter IgM fragments for Illumina sequencing. Multiplexed PCR was employed to amply rearranged IGH sequences using forward primers matching the first frame regions in IGHV gene segments and reverse primers aligning the fourth frame regions in IGHJ gene segments, which covered the antibody variable domains consisting of the three CDRs. The primers used in our study were listed in the Supplementary Materials. PCR amplification were performed in a volume of 50 μL, using 25 μL Pfu mastermix (CWbio, China), 1 μL template, and 1 μL (50 nM) each primer mixture. The PCR conditions were as follows: initial denaturation at 94°C for 5 min, 35 cycles of denaturation at 94°C for 30 s, annealing at 56°C for 1 min, extension at 72°C for 1 min, and final extension at 72°C for 10 min. The PCR amplicons were purified using the QIAquick Gel Extraction Kit (Qiagen, Germany), then underwent high-throughput sequencing based on Illumina Hiseq platform according to the manufacturer's protocol.

Sequences Processing

A series of stringent quality control criteria were applied to exclude biologically implausible sequences. First, raw reads were filtered for Phred quality score of 20 over 80% of nucleotides to gain clean data to exclude sequences with PCR errors and sequencing artifacts. The sequences were classified to productive and unproductive groups according to the analysis of IMGT/HighV-QUEST. The unproductive VDJ rearrangements were eliminated from the dataset, and the productive sequences were excluded when containing insertions and deletions (indels) or stop codons in V- and J-gene segments. These indels or stop codons would break the reading frames in VDJ segments. It is believed that the B cells need a functional antigen receptor to survive (20), and therefore, when such breaks appeared, sequences might contain either PCR errors or sequencing artifacts (21). Furthermore, sequences carrying substitutions or mutations in the conserved amino acids at specified positions were removed to avoid the substitution errors in Illumina platform. The possibility of misclassification of VDJ gene segments in the algorithms of the IMGT tool for VDJ region searching mainly depended on these special amino acids (22). Additionally, the redundant sequences were eliminated to avoid the accumulation of one single sequence due to PCR amplification. The unique clones were defined by sequences containing unique VDJ, including unique alleles and CDR3 sequences. The number of sequences after each step of sorting is listed in **Table 2**. The sequences have been deposited in the NCBI SRA database (SUB3220644). IMGT/High V-QUEST (version 1.5.1) was used for sequence annotation to determine the V(D)J genes, CDRs, and junctional modification and to identify indels errors (23). Results from IMGT/High V-QUEST analysis were imported into PostgreSQL database, and Structured Query Language (SQL) was used to retrieve the data for statistical analysis.

Statistical Analysis

Data analyses were performed using GraphPad Prism, Perl, and R programs. Student's *t*-test, Pearson's chi-test, and logistic regression analyses were used in the statistical significance analyses when required. Because statistically significant differences are more likely to occur with large sample sizes, effect sizes are necessary to understand if the differences are meaningful. The effect size of Student's *t*-test is Cohen's *d* value, which is used to measure the standardized difference between two means, as initially suggested by Cohen (24): when $d = 0.20$, the ES or the difference is considered to be small; when $d = 0.50$, the ES is medium; and when $d = 0.80$, the ES is large. For chi-square analyses and logistic regression, the odds ratio (OR) was used as the effect size (25). Generally, OR values that ranged from 0.9 to 1.1 was considered to be not significantly different; when $1.2 < OR < 1.4$ or $0.7 < OR < 0.8$, the difference was slight; when $1.5 < OR < 2.9$ or $0.4 < OR < 0.6$, the difference was medium; and when $OR > 3.0$ or $OR < 0.3$, the difference was large or the association was strong.

TABLE 2 | The number of input cells and sequencing data.

	Neonates	Adults
Theoretical number of lymphocytes[a]	$1.1–2.1 \times 10^9$	$6–9.6 \times 10^9$
Input cells (10^8/100 mL)	1.0×10^9	6.6×10^9
Raw sequences (clean data)	10,122,711 (1% of input cells)	15,978,350 (0.24% of input cells)
Unique sequences (nt)	8,475,193	15,057,048
Productive sequences	6,532,659 (77.0% of unique sequences)	11,820,648 (78.5% of unique sequences)
Unproductive sequences	428,411 (5.1% of unique sequences)	629,333 (4.2% of unique sequences)
Error sequences	1,514,123 (17.9% of unique sequences)	2,607,067 (17.3% of unique sequences)
Unique clones (productive)	3,209,817	7,303,188

[a]*The theoretical number of lymphocytes was estimated by the estimated number of lymphocytes reported previously ($0.5–0.9 \times 10^9$ cells/L in neonates and $0.16–0.68 \times 10^9$ cells/L in adults, Table S1 in Supplementary Materials).*

Ethics Statement

The CB samples were provided by NDRI (Philadelphia, PA, USA) with approval of the institutional research board and the donor consent. Procedures followed in this study were in accordance with the ethical standards of concerned institutional policies and the Research Donor Program of National Cancer Institute.

RESULTS

The Repertoire Diversity

By high-throughput sequencing, we obtained two immune repertoires of IgHs, one from the B cells in the CB of healthy neonates (CB), and the other from the B cells in peripheral blood of healthy human adults (HH). Initially, 10,122,711 raw sequences were collected from CB, and 15,978,350 sequences were obtained from HH. Next, we performed a series of stringent data filtering and cleaning procedures to exclude unproductive or biologically implausible sequences, as described under the Section "Materials and Methods." The sequences that had unique VDJ gene rearrangements, including those contained unique CDR3 amino acid (aa) sequences, or had identical CDR3 but distinct

VDJ rearrangements were defined as "unique clones." A total of 3,209,817 unique clones (31.7% of raw sequences) were identified in the CB repertoire, and 7,303,188 unique clones (45.7% of raw sequences) were found in the HH repertoire (**Figure 1A**). To exclude the bias caused by the number of input cells or the sequences, we randomly selected sequences from each datasets using the randomized table generated by R program (repeated three times), which represents the computational simulation to sample the same amount of input cells or sequences. Then, we calculated the proportion of the unique clones out of the randomly selected sequences. The results showed that, when the sample size was small, the proportion of the unique clones did not differ greatly between CB and HH. In contrast, when the sample size increase to 1,000,000, the proportion of the unique clones began to show difference between CB and HH (CB: 66%, HH 77%), indicating that the HH repertoire was more diverse than CB (Figure S1 in Supplementary Material).

Interestingly, we found that the HH and CB repertoires only shared 21,753 unique clones, constituting 0.7% of the CB and 0.3% of the HH unique clones, respectively (**Figure 1A**). Among these unique clones, 1,496,278 unique CDR3 aa sequences (46.6% of unique clones) were identified in the CB repertoire, and 3,428,850 unique CDR3 (46.7% of unique clones) were found in the HH repertoire (**Figure 1A**). Similarly, only 47,640 CDR3 sequences were shared by both repertoires, constituting 3.2% of the CB and 1.4% of the HH unique CDR3 sequences.

Next, we analyzed the VDJ rearrangement patterns using the IMGT/High V-QUEST tool (version 1.5.1). We included the gene allele information in the calculation of VDJ gene patterns to estimate the antibody repertoire diversity, because the allele information represents the genetic polymorphism that also results in repertoire diversity. There were 30,309 and 34,688 unique VDJ patterns in CB and HH repertoires, respectively, rearranged by 178 germline V-, 27 D-, and 13 J-gene segments (**Figures 1A,B**). The two repertoires shared 25,704 identical VDJ rearrangement patterns, which accounted for 84.8% of patterns in CB repertoire and 74.1% in HH repertoire. Taken together, these results highlight the overwhelmingly high diversity of human IgH repertoires as little antibody sequences were shared by two different repertoires, although recombined from similar VDJ genes and rearrangement patterns.

VDJ Gene Usage

To find the preferentially utilized VDJ gene in the two repertoires, the usages of the IGHV, IGHD, and IGHJ gene segments were calculated and shown in **Figure 2**. In the VDJ gene usage analyses, the gene alleles were not included to pack the data and reduce the data groups. There are 51 IGHV genes belonged to 7 gene families (**Figures 2A,B**). The top three preferred IGHV genes were IGHV4-59 > IGHV4-34 > IGHV2-5 in the CB repertoire (**Figures 2A,D**), and were IGHV4-59 > IGHV1-69 > IGHV4-34 in the HH repertoire (**Figures 2B,D**). Of all seven IGHV gene families, IGHV1, IGHV2, IGHV3, and IGHV4 gene families were mainly used, and together accounted for 94.5% in CB repertoire and 99.9% in HH repertoire (**Figure 2C**). On the other hand, a dramatic decreased use of IGHV5, IGHV6, and IGHV7 gene families was found in the HH repertoire as compared to that in the CB (0.1% vs. 5.5%). The usage of IGHV4 gene family (43.5%) was much higher than the other gene families in the HH repertoire, while both IGHV3 and IGHV4 gene families were frequently observed in the CB repertoire, with a rate about 30% (**Figure 2C**).

In both repertoires, the most populated group in the IGHD sets was IGHD3, with a rate of 28.7% in CB and 34.1% in HH (**Figure 2C**). IGHD7 or IGHD7-27, the only member in IGHD7 family, was rarely observed in the HH repertoire (1.3%) but accounted for about 10% in CB. The detailed classifications of IGHD gene groups revealed the top three frequently used IGHD genes: IGHD3-10 > IGHD6-13 > IGHD7-27 in CB and IGHD3-10 > IGHD3-22 > IGHD1-26 in the HH repertoire (**Figure 2D**).

For IGHJ genes, the usage of IGHJ2, IGHJ3, IGHJ4, and IGHJ6 accounted for a large proportion (more than 90%), while IGHJ1 and IGHJ5 were comparatively used less in both repertoires, with

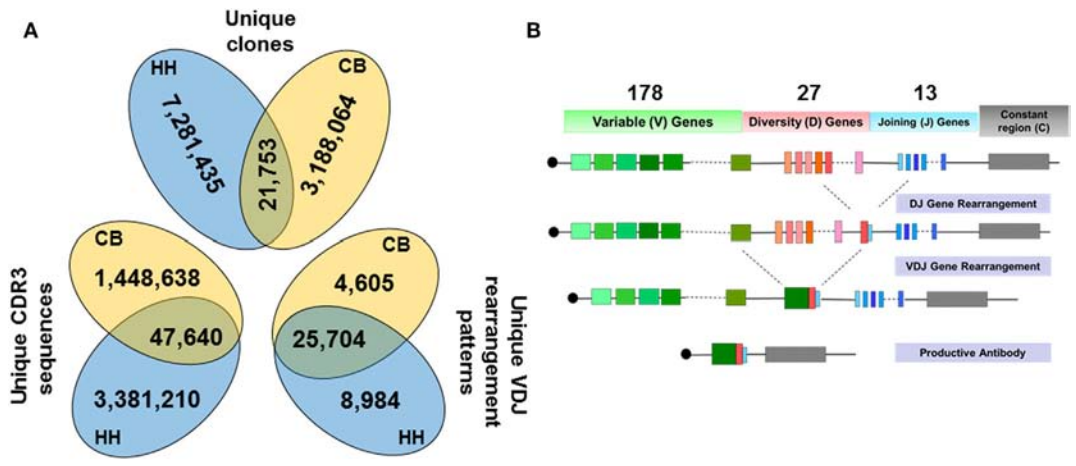


FIGURE 1 | **(A)** Data summary of the unique clones, the unique CDR3 sequences, and the unique VDJ gene rearrangement patterns from the repertoires of the neonates and adults. **(B)** Schematic of VDJ gene rearrangement, showing the number of V, D, and J gene (including alleles) observed in our dataset.

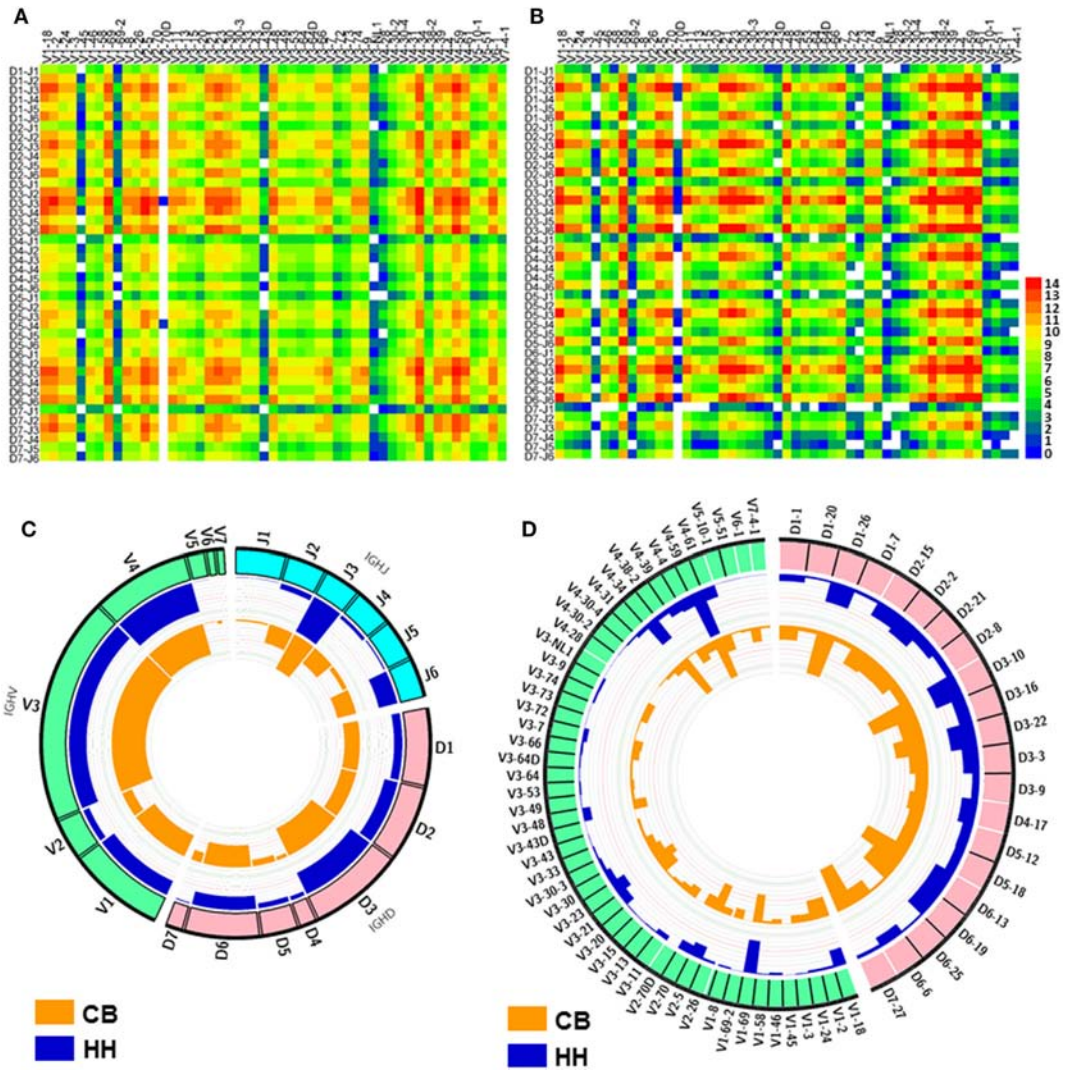


FIGURE 2 | The VDJ gene rearrangements and VDJ gene usage observed in the repertoires of the neonates and adults. **(A,B)** The heatmap of VDJ gene rearrangements observed in the IgH repertoire of the neonates **(A)** and the adults **(B)**, The base-2 logarithm of the count was used to present the frequency of each type of rearrangements. **(C)** The usage of VDJ gene families. The outsider arcs represent the V (green arc), D (pink arc), and J (cyan arc) gene families, and the histograms inside the circle represent the usage of each gene in the repertoire of the neonates (orange) and adults (blue). **(D)** The usage of V (green arc) and D (pink arc) gene subgroups in the repertoire of the neonates and adults. The histograms inside the circle represent the usage of each gene subgroups in the repertoire of the neonates (orange) and adults (blue).

a rate of 6.9% in CB and 2.0% in the HH repertoire. IGHJ3 was the most frequently used group in both repertoires, with a rate of 37.4% in CB and 57.3% in HH.

Although the comparison of VDJ gene usage between the two repertoires showed that the difference was statistically significant ($p < 0.05$, **Table 3**), the OR values were close to 1.0, indicating that the effect was small, and the low p value was mainly due to the large sample size. Indeed, as shown in **Figure 2**, the difference was only slight between the two repertoires. In both repertoires, the most frequently used VDJ gene was IGHV4-59, IGHD3-10, and IGHJ3. Taken together, these results suggest that the VDJ gene usage was similar in the IgH repertoire of neonates and adults.

The Characteristics of CDRs

The CDRs play critical roles in the binding of antibodies to antigens. In both repertoires, the length of CDR1 ranged from 8 to 10 aa (**Figure 3A**) and CDR2 ranged from 7 to 10 aa (**Figure 3B**). The CDR1 length of 8 aa was the most common observed (CB: 75.38%, HH: 77.63%), and the CDR2 length of 7 aa and 8 aa together accounted for the majority of the repertoires (CB: 94.91%, HH: 99.56%). As shown in **Table 4**, there was no apparent difference in the average length of CDR1 (CB: 8.5, HH: 8.4) or CDR2 (CB: 7.6, HH: 7.5) between the CB and the HH repertoires.

Interestingly, we found that the CDR3 length in the HH repertoire was evidently longer than the CB repertoire. Compared to that of CDR1 and CDR2, the length of CDR3 was much more

TABLE 3 | The VDJ gene usage distribution in the repertoires of neonates and adults.

Gene group	Repertoire		p^b	OR[b]	95% CI[b]
	CB[a] %	HH[a] %			
IGHV1	22.08	22.05	2.27E−58	1.009	(1.008, 1.010)
IGHV2	10.39	9.19			
IGHV3	32.53	25.16			
IGHV4	30.42	43.47			
IGHV5	0.86	0.10			
IGHV6	3.27	0.02			
IGHV7	0.45	0.02			
IGHD1	14.78	12.07	<2.2E−16	0.894	(0.893, 0.894)
IGHD2	14.93	19.94			
IGHD3	28.72	34.10			
IGHD4	3.86	5.89			
IGHD5	6.47	7.40			
IGHD6	21.23	19.33			
IGHD7	10.01	1.29			
IGHJ1	1.98	0.38	<2.2E−16	1.113	(1.112, 1.114)
IGHJ2	18.36	8.01			
IGHJ3	37.44	57.25			
IGHJ4	18.83	4.97			
IGHJ5	4.93	1.64			
IGHJ6	18.48	27.74			

[a]*CB: the repertoire of the neonates; HH: the repertoire of the adults.*
[b]*Calculated by the logistic regression.*
OR, odds ratio; 95% CI, 95% confidence interval.

variable, ranged from 3 to 42 aa in the CB repertoire and from 3 to 38 aa in the HH repertoire. As shown in **Figure 3C**, the CDR3 length of 14 aa was the most frequently observed in the CB repertoire, while the 15 aa CDR3 accounted for the largest proportion in the HH repertoire. Furthermore, for CDR3 length of 14 aa or smaller, the CB repertoire exhibited significantly higher frequencies, but just the opposite for CDR3 length of 15 aa or larger.

The aa changes in CDR1 and CDR2 as compared to germline sequences were also calculated in our analysis. The proportion of sequences with aa changes in CDR1 region was 19.33% in the CB repertoire and 25.84% in the HH repertoire (**Figure 3D**). Logistic regression showed that the rate of aa changes in CDR1 region of the HH repertoire was about 1.5 times higher than the CB [OR = 1.454, 95% CI: (1.449, 1.459)]. Similarly, the proportion of sequences with aa changes in CDR2 region was 9.26% in the CB repertoire and 17.82% in HH (**Figure 3E**), and the rate of aa changes in CDR2 region of HH was about twice as much as that of the CB repertoire as defined by logistic regression [OR = 2.124, 95% CI: (2.115, 2.133)]. As expected, these results indicate that the extent of somatic hypermutation occurred in the CDR1 and CDR2 regions was higher in the IgH repertoire of adults than that of the neonates.

The aa changes in CDR3 region cannot be calculated due to the extremely high flexibility of this region. Therefore, the aa usage of CDR3 region was analyzed instead, as shown in **Figure 3F**. Tyrosine, alanine, glycine, and aspartic acid were the most frequently occurring amino acids in CDR3. The aa composition of the CDR3 demonstrated no significant difference between the two repertories [OR = 0.993, 95% CI: (0.993, 0.993)]. The hydrophobicity value of the amino acids was determined by Kyte-Doolittle scale. The average hydrophobicity value of all the CDR3 sequences was -0.43 ± 2.70 for the CB repertoire and -0.28 ± 2.82 for HH and showed no significant difference [t-value = −296.29, Cohen's $d = 0.05$, $p < 2.2E{-}16$, 95% CI: (−0.144, −0.141)]. Taken together, these results suggest that the CDRs in the IgH repertoire of adults have characteristics similar to that of neonates, except for the slightly higher level of somatic hypermutation and significantly longer CDR3 regions.

V-D-J Junction Analysis

In addition to recombinational diversity, the diversity of IgH repertoire also came from the V-D-J junctions including the palindromic nucleotides (P) addition and the non-template randomized nucleotides (N) addition, as well as the deletion of nucleotides caused by exonuclease trimming (T). The N additions happened at the region between the 3′-end of V gene and the 5′-end of D gene (N1) and the region between the 3′-end of D gene and the 5′-end of J gene (N2). The P additions and the exonuclease trimming were observed at 3′-end of V regions (3VP and 3VT), 5′-end and 3′-end of D genes (5DP and 5DT, 3DP and 3DT) and 5′-end of J genes (5JP and 5JT). The occurrence of the P/N addition and exonuclease trimming is shown in **Figure 4**. Notably, the occurrence of N2 addition were significantly higher in the HH repertoire than CB, and the occurrence of 3DP and 3DT, 5JP, and 5JT showed slight difference between the two repertoires, while other types of modification showed no significant difference (**Table 5**). The average length of N2 addition was also greater in the HH repertoire (6.03 nt) than CB (5.08 nt), and there was no evident difference in the length of N1 addition (CB, 6.38 nt; HH, 6.41 nt) between the two repertoires (**Table 6**). The diversity of N2 addition in the HH repertoire is 3.5-fold higher than that in the CB repertoire, representing the most prominent difference among all the junctional modifications (**Table 7**).

Next, we analyzed the association between VDJ genes and the occurrence of N/P addition, along with exonuclease trimming in junctions (**Figure 5**). When combined the CB and HH repertoires together, most of the IGHV, IGHD, or IGHJ gene showed no or only slight statistical differences among the different subgroups, except for IGHD7, which displayed higher occurrence of 3DP and lower occurrence of 3DT than the other IGHD subgroups (**Figure 5C**; **Table 8**). Intriguingly, we found that such statistical difference was solely resulted from the CB repertoire. As shown in **Figure 5C**, the occurrence of the 3DP addition related to IGHD7 was evidently higher in CB as compared to the HH repertoire, while the 3DT trimming was lower. Furthermore, the N2 addition related to IGHD7 was significantly higher than that of any other IGHD subgroup in the CB repertoire, but lower than any other IGHD subgroup in the HH repertoire (**Figure 5G**). Except for IGHD7, all other gene subgroups had a considerably higher occurrence of N2 addition in the HH repertoire than that in the CB repertoire, and the occurrence of other types of addition or trimming (3VP, 3VT, 5DP, 5DT, 3DP, 3DT, 5JP, 5JT, and N1) did not show a significant difference between the two repertoires.

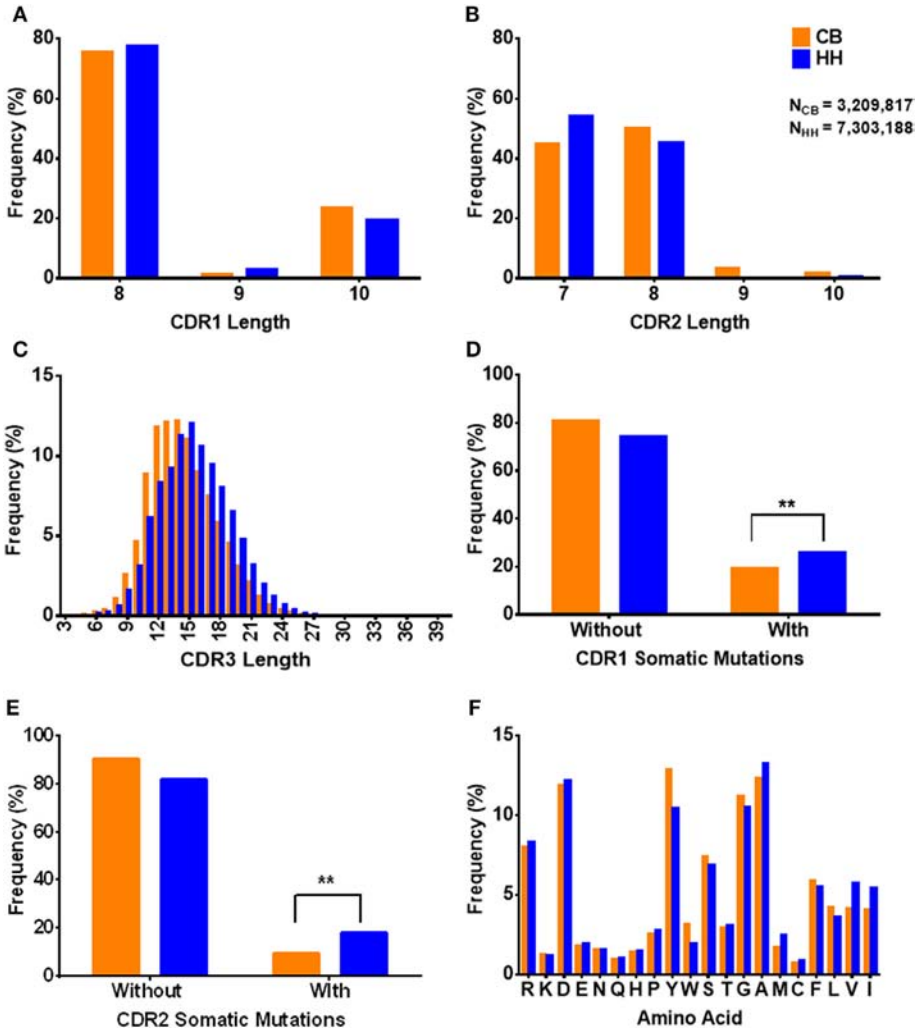


FIGURE 3 | The characteristics of CDRs in the IgH repertoire of the neonates and the adults. **(A)** The CDR1 length distribution. **(B)** The CDR2 length distribution. **(C)** The CDR3 length distribution. **(D)** The occurrence of amino acids changes in CDR1 regions. "Without" represented no amino acids changes, and "With" represented at less one amino acid change happened when compared to the germline IgH gene. **(E)** The occurrence of amino acids changes in CDR2 regions. **(F)** The amino acids usage of CDR3 sequences.

TABLE 4 | The average length (aa) of CDRs in the repertoires of neonates and adults.

CDR	Repertoire		T^a	*d*	*p*	95% CI
	CB	HH				
CDR1	8.48 ± 0.85	8.42 ± 0.8	108.443	0.07	<2.2E−16	(0.059, 0.062)
CDR2	7.62 ± 0.64	7.47 ± 0.52	381.801	0.28	<2.2E−16	(0.154, 0.156)
CDR3	14.51 ± 3.33	15.48 ± 3.43	−432.565	0.29	<2.2E−16	(−0.978, −0.969)

CDR, the complementarity-determining region.
[a]*Calculated by Student's t-test.*

DISCUSSION

In this study, we adapted Illumina-based high-throughput sequencing to analyze characteristics of the IgH repertoires of the CB samples from neonates and peripheral blood samples from healthy adults. A total of 26,101,061 antibody sequences were initially obtained from 43 individuals, and a series of strict data cleaning procedures were employed to remove unproductive or biologically implausible sequences. Furthermore, we introduced a strict definition of "unique" antibody clone, which only refers to the unique antibody sequence containing a unique VDJ gene rearrangement or a unique CDR3 sequence. Although the unique

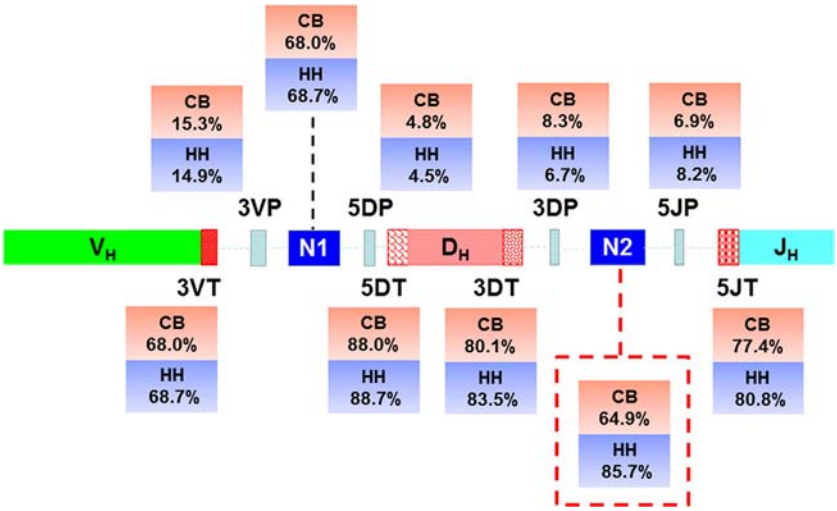


FIGURE 4 | Schematic of the junctional modifications during VDJ rearrangement, showing the locations and the occurrence of different types of junctional modifications. 3VP and 3VT: the palindromic nucleotides (P) additions and exonuclease trimmings observed at 3′-end of V regions, respectively; 5DP and 5DT: the P additions and exonuclease trimmings observed at 5′-end of D genes, respectively; 3DP and 3DT: the P additions and exonuclease trimmings observed at 3′-end of D genes, respectively; 5JP and 5JT: the P additions and exonuclease trimmings observed at 5′-end of J genes, respectively; N1: the non-template randomized nucleotides (N) additions happened at the region between the 3′-end of V gene and the 5′-end of D gene; N2: N additions happened at the region between the 3′-end of D gene and the 5′-end of J gene.

TABLE 5 | The occurrence of junctional modifications in the repertoires of neonates and adults.

Junctional modification	Repertoire		p^a	OR	95% CI
	CB %	HH %			
3VP	15.33	14.87	<2.2E−16	0.964	(0.961, 0.968)
N1	87.9	87.35	<2.2E−16	0.951	(0.947, 0.954)
5DP	4.79	4.46	2.45E−123	0.927	(0.921, 0.933)
3DP	8.28	6.74	<2.2E−16	0.80	(0.796, 0.804)
N2	64.87	85.69	<2.2E−16	3.243	(3.232, 3.253)
5JP	6.93	8.17	<2.2E−16	1.195	(1.189, 1.201)
3VT	68.07	68.66	8.52E−79	1.028	(1.025, 1.031)
5DT	88.01	88.67	<2.2E−16	1.067	(1.062, 1.071)
3DT	80.11	83.45	<2.2E−16	1.252	(1.248, 1.256)
5JT	77.35	80.84	<2.2E−16	1.236	(1.232, 1.240)

[a]*Calculated by the logistic regression.*

VDJ gene rearrangements can represent the genetic background of antibody clones, the junctional modification occurred in the CDR3 regions serves as one of the critical mechanisms for the generation of antibody diversity. Therefore, we include the sequences containing the identical VDJ gene rearrangement but unique CDR3 sequences for the representation of junctional diversity. For the sequences with the same VDJ gene rearrangement and the identical CDR3, only one sequence can be preserved in our dataset. Under this circumstance, we may lose sequences with nucleotide polymorphisms in the IGHV, but the representativeness of our data was not affected, since more than 98% of our original sequences share at least 90% identity with the sequences from IGMT database (data not shown). Finally, a total of 10,513,005 unique antibody clones (40.3% of raw sequences) were identified. By using these procedures, we can condense the large data size and eliminate the noise of dataset, thereby facilitating the subsequent data statistics and analysis.

A couple of interesting findings were made from the comparative study of the neonatal and adult IgH repertoires. Our study confirmed the extremely large diversity of human IgH repertoires, as less than 1% of unique clones were shared by the two repertoires. Despite the difference in sequences, we found unexpectedly high levels of similarity between the two repertoires regarding the VDJ gene usage, the characteristics of CDRs, and the occurrence of certain junctional modifications. Surprisingly, the most significant difference came from the occurrence frequency of N2 addition, a process that occurs during V-D-J recombination for introducing random nucleotide additions between D- and J-gene segments. Our study also demonstrated that the IgH repertoire of healthy adults was more diverse than that of neonates, which added the evidence that the fetal repertoires were relatively limited compared to the adult repertoires. The higher occurrence of hypermutations and the N2 addition might be the reason why the repertoire of adults had higher diversity than the neonates.

The VDJ gene usage has been a topic of considerable interest because it is possible that the immune repertoires could be skewed toward a single VDJ gene family or a single VDJ gene. For the IGHV usage in our data, we found that the most frequently used IGHV gene family was IGHV3 in the neonatal repertoire and IGHV4 in the adult repertoire, and the most preferentially utilized IGHV subgroup was IGHV4-59 in both repertoires. Previous studies using 454 sequencing showed that IGHV1 group was the most predominant IGHV group in the CB IgM repertoire (26), whereas the IGHV3 group was the most populated group in the IgM repertoire of adult populations (7). For IGHJ gene usage, the IGHJ3 group was the mostly used one in both repertoires. However, previous studies showed that IGHJ4 was mostly found

TABLE 6 | The average length (nt) of junctional modifications in the repertoires of neonates and adults.

Junctional modification	Repertoire		t^a	*d*	*p*	95% CI
	CB	HH				
3VP	1.56 ± 0.75	1.56 ± 0.75	5.117	0.01	3.11125E−07	(0.004, 0.009)
N1	6.38 ± 4.43	6.41 ± 4.57	−10.679	−0.01	1.28016E−26	(−0.041, −0.028)
5DP	1.64 ± 0.8	1.6 ± 0.78	16.360	0.05	3.92882E−60	(0.036, 0.045)
3DP	1.51 ± 0.78	1.38 ± 0.7	74.297	0.18	<2.2E−16	(0.132, 0.139)
N2	5.08 ± 4.25	6.03 ± 4.68	−270.882	−0.21	<2.2E−16	(−0.959, −0.945)
5JP	1.43 ± 0.73	1.37 ± 0.72	31.284	0.08	1.4071E−214	(0.053, 0.061)
3VT	2.84 ± 1.99	2.75 ± 1.89	53.757	0.04	<2.2E−16	(0.083, 0.089)
5DT	6.91 ± 5.01	7.21 ± 5.12	−81.316	−0.06	<2.2E−16	(−0.302, −0.288)
3DT	5.93 ± 4.15	6.55 ± 4.46	−196.632	−0.15	<2.2E−16	(−0.632, −0.620)
5JT	4.97 ± 4.37	5.78 ± 5.02	−649.265	−0.17	<2.2E−16	(−2.251, −2.238)

[a]*Calculated by Student's t-test.*

TABLE 7 | The number of different types of P/N additions in the repertoires of neonates and adults.

Library	P/N additions					
	3VP	5DP	3DP	5JP	N1	N2
CB	67	62	65	33	466,692	213,226
HH	72	74	101	38	935,620	751,780

in CB samples and in adults' repertoire (7, 26). For IGHD gene utilization, IGHD3 and IGHD6 groups formed almost half of the total IGHD gene usages in our data, which was also observed in the in the CB IgM repertoire (26). The preferential usage of IGHD7-27 (DQ52) in fetal samples was reported in some previous studies (27). In our data, IGHD7-27 was also frequently observed, accounting for about 10% of the neonatal repertoire; however, IGHD 7-27 only accounted for about 1% in the repertoire of adults. The reason why our results were not exactly consistent with the previous 454 sequencing-based studies of IgM repertoire in neonates or adults might be the different sequencing depth, the vast variety of the Ig repertoires, and the difference in individuals' genetic background.

In our study, we found that the VDJ gene usage were not significantly different between the neonates and adults. We calculated the VDJ gene usage by including the information of IgH gene alleles. In total, 178 V-, 27 D-, and 13 J-gene segments were found in our study, and only two IGHV alleles were found not shared by the two repertoires. Theoretically, the frequency of each VDJ gene allele could be 0.56% in V-, 3.7% in D-, and 7.69% in J-gene if each VDJ genes were used randomly in the VDJ rearrangements. We used these theoretical values as the threshold and divided the VDJ genes into two groups that were the frequently used (FUD) genes and the rarely used (RUD) genes (Figure S3 in Supplementary Material). For V genes, there were only 47 out of 178V genes in CB as well as 36 out of 178V genes in HH whose usage were more than 0.56% that can be defined as the FUD V genes, and most of these genes (29 FUD V genes) were shared by the two repertoires. Similarly, most of the FUD D genes (7 out of 8 FUD D genes in CB and 11 in HH) and J genes (3 out of 4 FUD J genes in CB and 3 in HH) were also shared by the two repertoires. Therefore, the reason why the VDJ gene usage did not show significant difference could be that the two repertoires shared the majority of these FUD VDJ genes. However, our data also showed that the preferred VDJ genes were not exactly same between the two repertoires. This may partly due to the effects of age (antigen exposure), but we cannot exclude the influence of the individual difference. A longitudinal investigation on the same individual(s) could be more ideal to clarify this point.

The IgH CDR3 region is the most diverse component of the antibody and typically plays a critical role in defining the specificity of antibodies (28–30). In our data, the CDR3 in the repertoire of adults were much more diverse than the neonates, but the aa usage was similar between the two repertoires. Interestingly, we found that the major difference stems from the length of CDR3 regions. The length of CDR1 or CDR2 was similar in both repertoires, since the length of CDR1 and CDR2 was mainly determined by the IGHV genes whose length diversity was restricted. In contrast, the adult repertoire displayed higher frequencies of CDR3 with 15 aa or longer, and lower frequencies of 14 aa or shorter (**Figure 3C**), resulting in a longer CDR3 in average in the adult repertoire (15.5 aa) as compared to the neonatal repertoire (14.5 aa), which were also observed when the dataset was divided into un-mutated and mutated sequences (Figure S2 and Table S3 in Supplementary Material). Despite this, we found that the majority of CDR3 length ranged from 10 to 20 aa in both repertoires (CB, 90.72%; HH, 86.43%), and antibodies with CDR3 longer than 20 aa only accounted for a small proportion in the two repertoires (CB, 4.92%; HH, 7.84%). Some previous studies suggested that the length of the HCDR3 sequences from the fetal repertoire were considerably shorter because of the preferred utilization of the shortest D gene, IGHD7-27 (10, 13, 27, 29, 31). Indeed, we also found that the IgH repertoire of neonates had higher usage of IGHD7-27 gene than the adults, but such effects could be compromised by the fact that the neonatal repertoire also exhibited increased occurrence of N/P additions and the smaller degree of exonuclease trimming in IGHD7-27 gene. The underlying mechanism for this phenomenon requires further investigation.

The long HCDR3 loops have previously shown to be associated with antibody auto-reactivity and poly-reactivity that can be removed from the human repertoire during B-cell development (32–34). Indeed, our data suggested that the most antibodies in

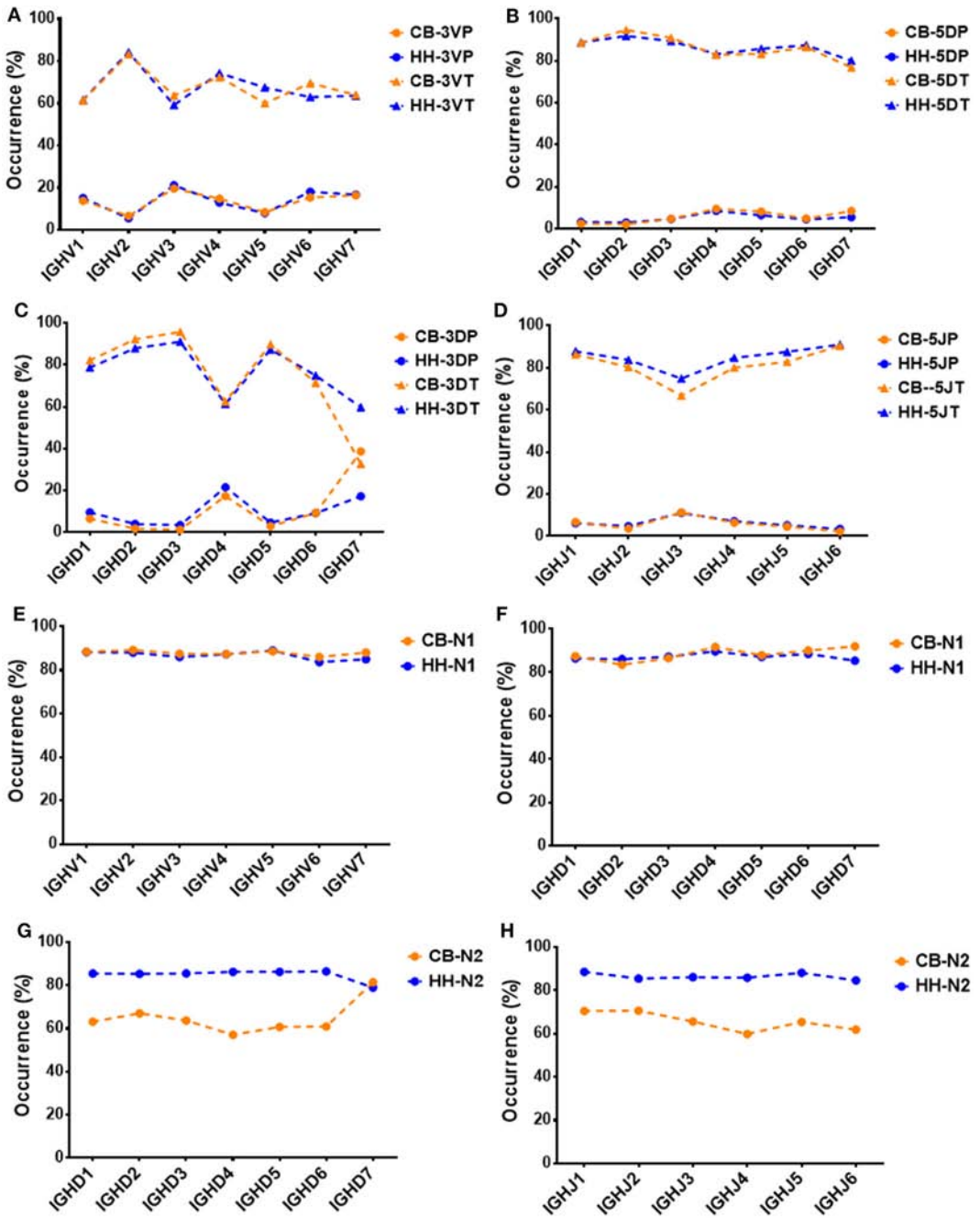


FIGURE 5 | The association between VDJ gene groups and the occurrence of P, N additions or exonuclease trimmings in VDJ junctions. **(A)** The association between the occurrence of 3VP, 3VT, and IGHV gene groups. **(B)** The association between the occurrence of 5DP, 5DT, and IGHD gene groups. **(C)** The association between the occurrence of 3DP, 3DT, and IGHD gene groups. **(D)** The association between the occurrence of 5JP, 5JT, and IGHJ gene groups. **(E)** The association between the occurrence of N1 and IGHV gene groups. **(F)** The association between the occurrence of N1 and IGHD gene groups. **(G)** The association between the occurrence of N2 and IGHD gene groups. **(H)** The association between the occurrence of N2 and IGHJ gene groups.

human IgH repertoires had the proper length of CDR3 loops but also retained a small proportion containing long CDR3 loops. A proper CDR3 length would be necessary to the survival and mature of B cells, including proper and efficient protein folding, proper pairing with the surrogate light chain to generate a functional antibody, and finally the ability to overcome negative selection of auto-reactive receptors. However, the retention of some longer CDR3 loops would be expected to increase the repertoire diversity and facilitate binding to recessed epitopes in pathogens or the active sites of enzymes (35–38).

The junctional diversification plays an important role in expanding the diversity of CDR3. Some previous studies also

TABLE 8 | The association between VDJ genes and the occurrence of N/P addition along with exonuclease trimming in junctions.

	CB			HH		
	p^a	OR	95% CI	*p*	OR	95% CI
IGHV-3VP	<2.2E−16	1.062	(1.060, 1.065)	1.93E−15	0.993	(0.991, 0.995)
IGHV-3VT	<2.2E−16	1.074	(1.072, 1.076)	<2.2E−16	1.144	(1.143, 1.146)
IGHV-N1	4.90E−187	0.962	(0.960, 0.965)	7.90E−241	0.968	(0.967, 0.97)
IGHD-N1	<2.2E−16	1.102	(1.10, 1.104)	<2.2E−16	1.039	(1.038, 1.04)
IGHD-5DP	<2.2E−16	1.192	(1.189, 1.195)	<2.2E−16	1.099	(1.097, 1.101)
IGHD-5DT	<2.2E−16	0.839	(0.837, 0.84)	<2.2E−16	0.915	(0.913, 0.916)
IGHD-3DP	<2.2E−16	1.659	(1.655, 1.663)	<2.2E−16	1.138	(1.136, 1.14)
IGHD-3DT	<2.2E−16	0.642	(0.641, 0.643)	<2.2E−16	0.848	(0.847, 0.849)
IGHD-N2	<2.2E−16	1.039	(1.038, 1.041)	5.93E−56	1.01	(1.009,1.011)
IGHJ-N2	<2.2E−16	0.917	(0.915, 0.918)	<2.2E−16	0.969	(0.968, 0.971)
IGHJ-5JP	<2.2E−16	0.823	(0.820, 0.825)	<2.2E−16	0.749	(0.747, 0.751)
IGHJ-5JT	<2.2E−16	1.264	(1.262, 1.267)	<2.2E−16	1.367	(1.365, 1.369)

[a]*Calculated by the logistic regression.*

described the characteristics of the junctional modifications in CDR3 regions. For instance, by analyzing hundreds of productive and nonproductive VDJ rearrangements, Souto-Carneiro et al. found that the average length and occurrence of N2 insertions of fetal, preterm, and full-term neonates were significantly less than that of the adult rearrangements in the productive B-cell repertoires. The mean length of N1, 3DT, and 5JT was also less in the neonatal productive repertoires than that of adults (39). Our study showed the similar characteristics of N2 addition, and we found that the N2 addition related to IGHD7 was significantly higher than that of any other IGHD subgroups in the neonatal repertoire, but lower than any other IGHD subgroups in the adult repertoire. In another study, the mean length of N addition and 5JT trimming was also observed to be longer in human adults than fetus (15.2 ± 0.8 vs. 8.8 ± 0.6, 7.4 ± 1.3 vs. 3.9 ± 0.9, respectively), but the nucleotides loss due to 5DT was greater in human fetus than in human adults (10.2 ± 1.1 vs. 6.2 ± 1.3) (40). By performing high-throughput 454 sequencing and IMGT/HighV-QUEST analysis of 28,169 antibody heavy chain sequences from two babies, Prabakaran et al. found that N addition (93%) and exonuclease trimming (97%) had very high occurrence rates as compared to that of P additions (26%) (26). In addition, using immunodeficient mice reconstituted with human B-cell progenitors, Kolar et al. found that the N addition of $sIgM^-$ cells was longer in adult chimeric mice than the fetal and CB chimeras, and the fetal chimeras had less N2 addition in comparison with adult chimeras. The N addition of $sIgM^+$ cells was slightly longer in CB chimeras than the fetal and adults (15). To compare with the previous findings, we re-calculated the occurrence and mean length of the total N addition by adding the N1 additions to the N2 of each sequences. The data suggested that the occurrence and mean length of total N addition had no significant difference between the neonates and the adults (occurrence: 88.51% in CB, 87.91% in HH; mean length: 6.48 ± 4.46 in CB, 6.48 ± 4.59 in HH), which was due to the abundant N1 additions in both repertoires that may cover the difference in N2 insertions.

Besides, a number of studies analyzed the repertoires of B-cell subpopulations and described the characteristics of VDJ usage, CDR3 length, junctional modification and somatic mutation in different B-cell subsets. The IGHV3 and IGHJ4 families were often found to be the most commonly observed gene families in the previous studies, although different grouping standards were used to divide the B-cell populations, and the usage of the VDJ gene usually showed some difference among the B-cells subpopulations (41–45). However, in our adult repertoire, IGHV4, IGHD3, and IGHJ3 were the most observe genes. It was surprisingly to find that the CDR3 lengths of IgD^+CD27^+ memory B cells were shorter than that of IgD^+CD27^- naïve B cells, and a slight reduction in CDRH3 length was also observed in antigen-experienced repertoires compared with naive repertoires (41, 43, 45–47). Besides, lower occurrence and shorter length of N addition, as well as higher occurrence and longer length of exonuclease trimming was observed in memory B cell population (41, 44). The higher affinity antigen-experienced B cells were considered to harbor the shorter CDR3 (44, 48–50). It is noteworthy to point out that we did not discriminate different cell populations, but rather pooled B cells and extracted all the IgM antibody gene by using the specific IGHM constant region primers. Therefore, our study represents a large sample surveying of the IgM repertoires of the neonates and the adults.

In this study, we were able to achieve much deeper sequencing depth with Illumina sequencing than that of 454 pyrosequencing. Importantly, we found that most of the characteristics in the repertoires of neonates and adults were similar, but the adults possess much higher occurrence of N2 addition, which may play important role in the age-related antibody repertoire changes. Taken together, the results suggested that the major source of diversity arose from the CDR3 region, and that the junctional modulations could be one of the major determinants for the increased diversity in the healthy adults, highlighting the importance of VDJ junctional modifications, especially the N2 addition.

In-depth analyses of the IgM repertoires could not only lead to a better understanding of the components in the human humoral immune system, but also have potential practical value for the development of antibody therapeutics and vaccines. For example, previous studies suggested that bioinformatics analysis can be used to identify potentially effective antibodies similar to a targeted functional antibody by analyzing the sequenced antibody

repertoire adapting a Phylogeny-based method (1, 4). Therefore, our sequence data could serve as a large database to search for potentially effective antibodies. Indeed, panels of potent human monoclonal antibodies against various disease targets have been identified recently that had no or very few somatic mutations (51–56). Additionally, with the awareness of the importance of N2 junctional motif in the antibody heavy chain, it is possible to achieve more effective antibody affinity maturation by diversifying N2 junctions inside CDR3, instead of introducing extensive somatic mutations throughout the entire antibody heavy chain. Moreover, the structural analysis of antigen-antibody complexes in repertoire-scale could be facilitated by bioinformatics methods such as Pyrosetta or Rosetta Antibody (46). These information may guide the design of vaccine candidates able to induce antibodies encoded by the most frequently used VDJ rearrangements in an individual, paving the way to the development of personalized vaccination.

ETHICS STATEMENT

The cord blood samples were provided by NDRI (Philadelphia, PA, USA) with approval of institutional research board and donor consent. Procedures followed in this study were in accordance with the ethical standards of concerned institutional policies and the Research Donor Program of National Cancer Institute.

AUTHOR CONTRIBUTIONS

TY, DD, and YWE conceived and designed the project. BH, YWU, WL, and XW carried out the experiments. BH analyzed the data. TY, BH, and SJ wrote the paper with input from all co-authors.

FUNDING

This work was supported by the National Natural Science Foundation of China (31570936, 81630090, 81561128006), the 1000 Young Talents Program of China, and the Intramural Research Program, National Cancer Institute, National Institutes of Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/fimmu.2018.00128/full#supplementary-material.

REFERENCES

1. Kwong PD, Chuang GY, Dekosky BJ, Gindin T, Georgiev IS, Lemmin T, et al. Antibodyomics: bioinformatics technologies for understanding B-cell immunity to HIV-1. *Immunol Rev* (2017) 275(1):108. doi:10.1111/imr.12480
2. Hou D, Chen C, John SE, Chen S, Song Y. High-throughput sequencing-based immune repertoire study during infectious disease. *Front Immunol* (2016) 7(7):336. doi:10.3389/fimmu.2016.00336
3. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, et al. Deep sequencing in library selection projects: what insight does it bring? *Curr Opin Struct Biol* (2015) 33(23):146–60. doi:10.1016/j.sbi.2015.09.001
4. Zhu J, Wu X, Zhang B, Mckee K, O'Dell S, Soto C, et al. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc Natl Acad Sci U S A* (2013) 110(43):4088–97. doi:10.1073/pnas.1306262110
5. Bonsignori M, Zhou T, Sheng Z, Chen L, Gao F, Joyce MG, et al. Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. *Cell* (2016) 165(2):449. doi:10.1016/j.cell.2016.02.022
6. Furukawa K, Akasakofurukawa A, Shirai H, Nakamura H, Azuma T. Junctional amino acids determine the maturation pathway of an antibody. *Immunity* (1999) 11(3):329. doi:10.1016/S1074-7613(00)80108-9
7. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106(48):20216–21. doi:10.1073/pnas.0909775106
8. Bauer K, Zemlin M, Hummel M, Pfeiffer S, Karstaedt J, Steinhauser G, et al. Diversification of Ig heavy chain genes in human preterm neonates prematurely exposed to environmental antigens. *J Immunol* (2002) 169(3):1349. doi:10.4049/jimmunol.169.3.1349
9. Schallert N, Pihlgren M, Kovarik J, Roduit C, Tougne C, Bozzotti P, et al. Generation of adult-like antibody avidity profiles after early-life immunization with protein vaccines. *Eur J Immunol* (2002) 32(3):752–60. doi:10.1002/1521-4141(200203)32:3<752::AID-IMMU752>3.0.CO;2-5
10. Schroeder HW, Hillson JL, Perlmutter RM. Early restriction of the human antibody repertoire. *Science* (1987) 238(4828):791–3. doi:10.1126/science.3118465
11. Wang X, Stollar BD. Immunoglobulin VH gene expression in human aging. *Clin Immunol* (1999) 93(2):132–42. doi:10.1006/clim.1999.4781
12. Schroeder HW, Wang JY. Preferential utilization of conserved immunoglobulin heavy chain variable gene segments during human fetal life. *Proc Natl Acad Sci U S A* (1990) 87(16):6146. doi:10.1073/pnas.87.16.6146
13. Raaphorst FM, Timmers E, Kenter MJH, Tol MJDV, Vossen JM, Schuurman RKB. Restricted utilization of germ-line VH3 genes and short diverse third complementarity-determining regions (CDR3) in human fetal B lymphocyte immunoglobulin heavy chain rearrangements. *Eur J Immunol* (1992) 22(1):247–51. doi:10.1002/eji.1830220136
14. Shiokawa S, Mortari F, Lima JO, Nuñez C, Rd BF, Kirkham PM, et al. IgM heavy chain complementarity-determining region 3 diversity is constrained by genetic and somatic mechanisms until two months after birth. *J Immunol* (1999) 162(10):6060.
15. Kolar GR, Yokota T, Rossi MI, Nath SK, Capra JD. Human fetal, cord blood, and adult lymphocyte progenitors have similar potential for generating B cells with a diverse immunoglobulin repertoire. *Blood* (2004) 104(9):2981. doi:10.1182/blood-2003-11-3961
16. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005) 437(7057):376. doi:10.1038/nature03959
17. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* (2008) 456(7218):53–9. doi:10.1038/nature07517
18. Chen W, Zhu Z, Xiao X, Dimitrov DS. Construction of a human antibody domain (VH) library. *Methods Mol Biol* (2009) 525:81. doi:10.1007/978-1-59745-554-1_4
19. Zhu Z, Dimitrov DS. Construction of a large naïve human phage-displayed fab library through one-step cloning. *Methods Mol Biol* (2009) 525(525):129. doi:10.1007/978-1-59745-554-1_6
20. Kraj P, Rao SP, Glas AM, Hardy RR, Milner EC, Silberstein LE. The human heavy chain Ig V region gene repertoire is biased at all stages of B cell ontogeny, including early pre-B cells. *J Immunol* (1997) 158(12):5824.
21. Hansen TØ, Lange AB, Barington T. Sterile DJH rearrangements reveal that distance between gene segments on the human Ig H chain locus influences their ability to rearrange. *J Immunol* (2015) 194(3):973–82. doi:10.4049/jimmunol.1401443
22. Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* (2011) 2011(6):633. doi:10.1101/pdb.ip86

23. Ehrenmann F, Giudicelli V, Duroux P, Lefranc MP. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb Protoc* (2011) 2011(6):726. doi:10.1101/pdb.prot5635
24. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates (1988). 20 p.
25. Muth JED. *Basic Statistics and Pharmaceutical Statistical Applications*. Boca Raton: BChapman & Hall/CRC Press (2006). 477 p.
26. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, et al. 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* (2012) 64(5):337. doi:10.1007/s00251-011-0595-8
27. Pascual V, Verkruyse L, Casey ML, Capra JD. Analysis of Ig H chain gene segment utilization in human fetal liver. Revisiting the "proximal utilization hypothesis". *J Immunol* (1993) 151(8):4164–72.
28. Larimore K, Mccormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol* (2012) 189(6):3221–30. doi:10.4049/jimmunol.1201303
29. Xu JL, Davis MM. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* (2000) 13(1):37. doi:10.1016/S1074-7613(00)00006-6
30. Wu TT, Johnson G, Kabat EA. Length distribution of CDRH3 in antibodies. *Proteins* (1993) 16(1):1–7. doi:10.1002/prot.340160102
31. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* (2011) 333(6049):1593–602. doi:10.1126/science.1207532
32. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science* (2003) 301(5638):1374. doi:10.1126/science.1086907
33. Aguilera I, Melero J, Nuñez-Roldan A, Sanchez B. Molecular structure of eight human autoreactive monoclonal antibodies. *Immunology* (2001) 102(3):273–80. doi:10.1046/j.1365-2567.2001.01159.x
34. Ichiyoshi Y, Casali P. Analysis of the structural correlates for antibody polyreactivity by multiple reassortments of chimeric human immunoglobulin heavy and light chain V segments. *J Exp Med* (1994) 180(3):885–95. doi:10.1084/jem.180.3.885
35. Zwick MB, Komori HK, Stanfield RL, Church S, Wang M, Parren PW, et al. The long third complementarity-determining region of the heavy chain is important in the activity of the broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2F5. *J Virol* (2004) 78(6):3155–61. doi:10.1128/JVI.78.6.3155-3161.2004
36. Genst ED, Silence K, Decanniere K, Conrath K, Loris R, Kinne JR, et al. Molecular basis for the preferential cleft recognition by dromedary heavy-chain antibodies. *Proc Natl Acad Sci U S A* (2006) 103(12):4586–91. doi:10.1073/pnas.0505379103
37. Bond CJ, Marsters JC, Sidhu SS. Contributions of CDR3 to V H H domain stability and the design of monobody scaffolds for naive antibody libraries. *J Mol Biol* (2003) 332(3):643–55. doi:10.1016/S0022-2836(03)00967-7
38. Desmyter A, Transue TR, Ghahroudi MA, Thi MH, Poortmans F, Hamers R, et al. Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. *Nat Struct Biol* (1996) 3(9):803. doi:10.1038/nsb0996-803
39. Souto-Carneiro MM, Sims GP, Girschik H, Lee J, Lipsky PE. Developmental changes in the human heavy chain CDR3. *J Immunol* (2005) 175(11):7425. doi:10.4049/jimmunol.175.11.7425
40. Link JM, Larson JE, Schroeder HW. Despite extensive similarity in germline DH and JH sequence, the adult rhesus macaque CDR-H3 repertoire differs from human. *Mol Immunol* (2005) 42(8):943. doi:10.1016/j.molimm.2004.09.027
41. Tian C, Luskin GK, Dischert KM, Higginbotham JN, Shepherd BE, Crowe JE Jr. Evidence for preferential Ig gene usage and differential TdT and exonuclease activities in human naïve and memory B cells. *Mol Immunol* (2007) 44(9):2173–83. doi:10.1016/j.molimm.2006.11.020
42. Briney BS, Willis JR, Mckinney BA, Crowe JE Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naïve and memory repertoires that extends across individuals. *Genes Immun* (2012) 13(6):469–73. doi:10.1038/gene.2012.20
43. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010) 116(7):1070–8. doi:10.1182/blood-2010-03-275859
44. Rosner K, Winter DB, Tarone RE, Skovgaard GL, Bohr VA, Gearhart PJ. Third complementarity-determining region of mutated VH immunoglobulin genes contains shorter V, D, J, P, and N components than non-mutated genes. *Immunology* (2001) 103(2):179. doi:10.1046/j.1365-2567.2001.01220.x
45. Wu YCB, David K, Dunn-Walters DK. The relationship between CD27 negative and positive B cell populations in human peripheral blood. *Front Immunol* (2011) 2(21):81. doi:10.3389/fimmu.2011.00081
46. Dekosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113(19):E2636. doi:10.1073/pnas.1525510113
47. Brezinschek HP, Foster SJ, Dörner T, Brezinschek RI, Lipsky PE. Pairing of variable heavy and variable kappa chains in individual naive and memory B cells. *J Immunol* (1998) 160(10):4762–7.
48. Pini A, Viti F, Santucci A, Carnemolla B, Zardi L, Neri P, et al. Design and use of a phage display library. Human antibodies with subnanomolar affinity against a marker of angiogenesis eluted from a two-dimensional gel. *J Biol Chem* (1998) 273(34):21769–76. doi:10.1074/jbc.273.34.21769
49. Padlan EA, Silverton EW, Sheriff S, Cohen GH, Smithgill SJ, Davies DR. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc Natl Acad Sci U S A* (1989) 86(15):5938–42. doi:10.1073/pnas.86.15.5938
50. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J Immunol* (1991) 147(5):1709.
51. Yeung YA, Foletti D, Deng X, Abdiche Y, Strop P, Glanville J, et al. Germline-encoded neutralization of a *Staphylococcus aureus* virulence factor by the human antibody repertoire. *Nat Commun* (2016) 7:13376. doi:10.1038/ncomms13376
52. Lingwood D, Mctamney PM, Yassine HM, Whittle JR, Guo X, Boyington JC, et al. Structural and genetic basis for development of broadly neutralizing influenza antibodies. *Nature* (2012) 489(7417):566. doi:10.1038/nature11371
53. Magnani DM, Cgt S, Rosen BC, Ricciardi MJ, Pedreñolopez N, Gutman MJ, et al. A human inferred germline antibody binds to an immunodominant epitope and neutralizes Zika virus. *PLoS Negl Trop Dis* (2017) 11(6):e0005655. doi:10.1371/journal.pntd.0005655
54. Bailey JR, Flyak AI, Cohen VJ, Li H, Wasilewski LN, Snider AE, et al. Broadly neutralizing antibodies with few somatic mutations and hepatitis C virus clearance. *JCI Insight* (2017) 2(9):92872. doi:10.1172/jci.insight.92872
55. Wen X, Mousa JJ, Bates JT, Lamb RA, Crowe JE Jr, Jardetzky TS. Structural basis for antibody cross-neutralization of respiratory syncytial virus and human metapneumovirus. *Nat Microbiol* (2017) 2:16272. doi:10.1038/nmicrobiol.2016.272
56. Fu Y, Zhang Z, Sheehan J, Avnir Y, Ridenour C, Sachnik T, et al. A broadly neutralizing anti-influenza antibody reveals ongoing capacity of haemagglutinin-specific memory B cells to evolve. *Nat Commun* (2016) 7:12780. doi:10.1038/ncomms12780

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.

Copyright © 2018 Hong, Wu, Li, Wang, Wen, Jiang, Dimitrov and Ying. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

REVIEW
published: 21 February 2018
doi: 10.3389/fimmu.2018.00224

Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires

Enkelejda Miho[1,2], Alexander Yermanos[1], Cédric R. Weber[1], Christoph T. Berger[3,4], Sai T. Reddy[1] and Victor Greiff[1,5]**

[1] Department for Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland, [2] aiNET GmbH, ETH Zürich, Basel, Switzerland, [3] Department of Biomedicine, University Hospital Basel, Basel, Switzerland, [4] Department of Internal Medicine, Clinical Immunology, University Hospital Basel, Basel, Switzerland, [5] Department of Immunology, University of Oslo, Oslo, Norway

OPEN ACCESS

Edited by:
Jacob Glanville,
Distributed Bio, United States

Reviewed by:
Benny Chain,
University College London,
United Kingdom
Claude-Agnes Reynaud,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

***Correspondence:**
Sai T. Reddy
sai.reddy@ethz.ch;
Victor Greiff
victor.greiff@medisin.uio.no

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: 22 November 2017
Accepted: 26 January 2018
Published: 21 February 2018

Citation:
Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST and Greiff V (2018) Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. Front. Immunol. 9:224. doi: 10.3389/fimmu.2018.00224

The adaptive immune system recognizes antigens *via* an immense array of antigen-binding antibodies and T-cell receptors, the immune repertoire. The interrogation of immune repertoires is of high relevance for understanding the adaptive immune response in disease and infection (e.g., autoimmunity, cancer, HIV). Adaptive immune receptor repertoire sequencing (AIRR-seq) has driven the quantitative and molecular-level profiling of immune repertoires, thereby revealing the high-dimensional complexity of the immune receptor sequence landscape. Several methods for the computational and statistical analysis of large-scale AIRR-seq data have been developed to resolve immune repertoire complexity and to understand the dynamics of adaptive immunity. Here, we review the current research on (i) diversity, (ii) clustering and network, (iii) phylogenetic, and (iv) machine learning methods applied to dissect, quantify, and compare the architecture, evolution, and specificity of immune repertoires. We summarize outstanding questions in computational immunology and propose future directions for systems immunology toward coupling AIRR-seq with the computational discovery of immunotherapeutics, vaccines, and immunodiagnostics.

Keywords: systems immunology, B-cell receptor, T-cell receptor, phylogenetics, networks, artificial intelligence, immunogenomics, antibody discovery

INTRODUCTION

The adaptive immune system is responsible for the specific recognition and elimination of antigens originating from infection and disease. Molecular recognition of antigens is achieved through the vast diversity of antibody (B-cell receptor) and T-cell receptors (TCRs). The genetic diversity of these adaptive immune receptors is generated through a somatic recombination process that acts on their constituent V, D, and J segments (1, 2). During the gene rearrangement process, additional sequence diversity is created by nucleotide deletion and addition, resulting in a potential diversity of $>10^{13}$ unique B- and T-cell immune receptor sequences (3–6). The adaptive immune repertoire often refers to the collection of all antibody and T-cell immune receptors within an individual and

represents both the ongoing and the past immune status of an individual. Current threats, for example of pathogenic nature, are countered by B- and T-cell clonal expansion and selection (7), whereas past ones are archived in immunological memory compartments (8). Immune repertoires are highly dynamic. They are constantly evolving within the repertoire sequence space, which is defined as the set of all biologically achievable immune receptor sequences. Repertoire dynamics and evolution span several orders of magnitude in size (germline gene to clonal diversity), physical components (molecular to cellular dynamics), and time (short-lived responses to immunological memory that can persist for decades) (9–14).

The quantitative resolution of immune repertoires has been fueled by the advent of high-throughput sequencing (2, 15–20). Since 2009, high-throughput adaptive immune receptor repertoire sequencing (AIRR-seq) has provided unprecedented molecular insight into the complexity of adaptive immunity by generating data sets of 100 millions to billions of reads (6, 21, 22). The exponential rise in immune repertoire data has correspondingly led to a large increase in the number of computational methods directed at dissecting repertoire complexity (**Figures 1** and **2**) (23). Immune repertoire sequencing has catalyzed the field of computational and systems immunology in the same way that genomics and transcriptomics have for systems and computational biology (23). To date, the computational methods that have been developed and applied to immune repertoires relate to (i) the underlying mechanisms of diversity generation, (ii) repertoire architecture, (iii) antibody evolution, and (iv) molecular convergence.

This review provides an overview of the computational methods that are currently being used to dissect the high-dimensional complexity of immune repertoires. We will treat only those methods that are downstream of data preprocessing although currently there is no consensus on standard operating preprocessing procedures, and please refer to recent reviews on these subjects (2, 17, 24). Specifically, this review centers on computational, mathematical, and statistical approaches used to analyze, measure, and predict immune repertoire complexity. The description of these methods will be embedded within the main areas of immune repertoire research. Given that the genetic structure of antibody and TCRs is very similar, the majority of the methods illustrated in this review can be applied both in the context of antibody and T-cell studies. Exceptions to this rule are stated explicitly.

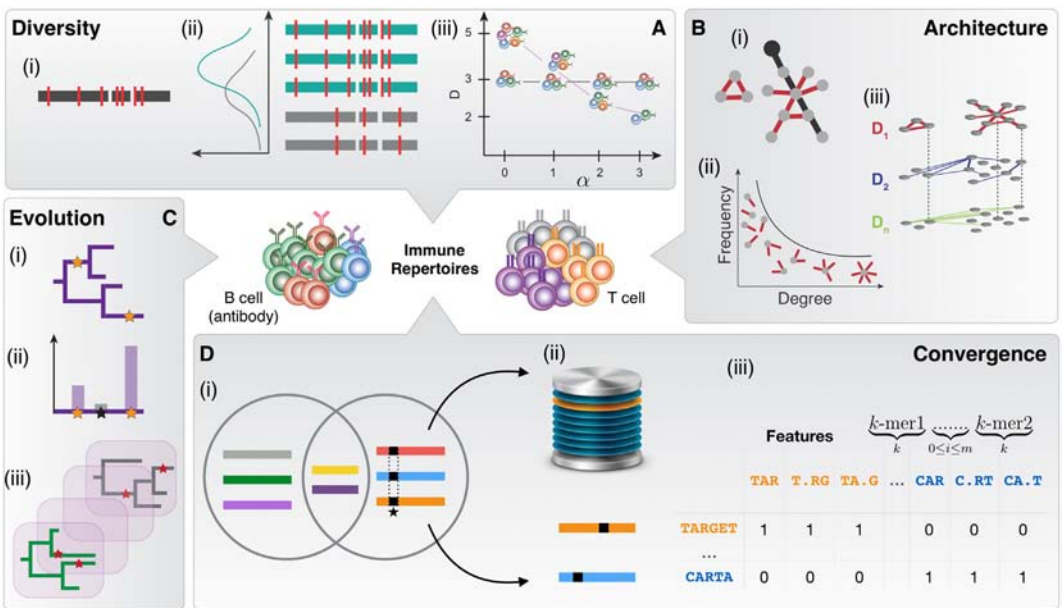


FIGURE 1 | The immune repertoire space is defined by diversity, architecture, evolution, and convergence. **(A)** Diversity measurements are based on (i) the accurate annotation of V (D) J segments using deterministic and probabilistic approaches with population-level or individualized germline gene reference databases. (ii) Probabilistic and hidden Markov models allow inference of recombination statistics. (iii) Measurement of clonotype diversity using diversity profiles. **(B)** Analysis of repertoire architecture relies predominantly on (i) clonal networks that are constructed by connecting nucleotide or amino acid sequence nodes by similarity edges. The sequence similarity between clones is defined *via* a string distance [e.g., Levenshtein distance (LD)], resulting in undirected Boolean networks for a given threshold (nucleotides/amino acids). An example of the global characterization of the network is the diameter, shown by black edges. An example of the local parameters of the network is the degree ($n = 1$) related to the individual clonal node in black. (ii) Degree distribution is a global characteristic of immune repertoire networks, which can be used for analyzing clonal expansion. (iii) Several similarity layers decompose the immune repertoire along its similarity layers. Layer D1 captures clonal nodes similar by edit distance 1 (1 nt/a.a. different), D2 of distance 2 and so forth. **(C)** Assessing evolution of antibody lineages. (i) Reconstruction of phylogenetic trees. Stars indicate somatic hypermutation. (ii) Probabilistic methods for the inference of mutation statistics in antibody lineage evolution. (iii) Simulation of antibody repertoire evolution for benchmarking antibody-tailored phylogenetic inference algorithms. **(D)** Naive and antigen-driven cross-individual sequence similarity and convergence in immune repertoires. (i) The Venn diagram shows sequences shared in the two repertoires (circles). Signature-like sequence features are highlighted by black squares. (ii) Database of convergent or antigen-specific immune receptor sequences. (iii) K-mer sequence decomposition and classification of immune receptor sequences.

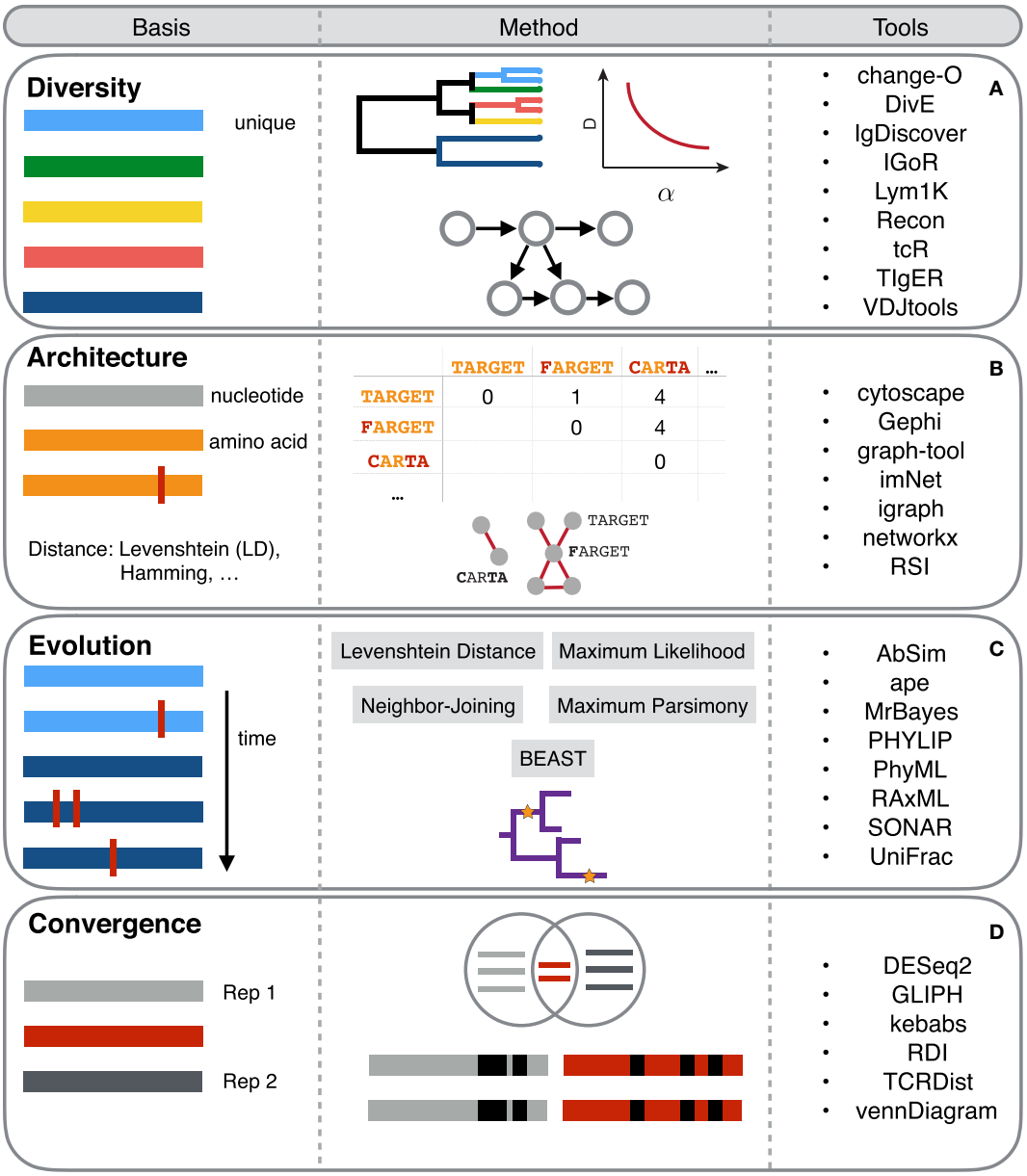


FIGURE 2 | An overview of selected computational tools used in immune repertoire analyses. Each horizontal colored bar colored bar in the *Basis* column represents a unique antibody or T-cell receptor (TCR) sequence. Vertical red bars represent sequence differences or somatic hypermutation. The *Method* column describes the general concept of the computational methods and how these are applied to immune repertoires. The *Tools* column highlights exemplary key resources for performing computational analysis in the respective analytical sections [rows **(A–D)**].

MEASURING IMMUNE REPERTOIRE DIVERSITY

The immense diversity is one of the key features of immune repertoires and enables broad antigen recognition breadth (**Figures 1A** and **2A**). The maximum theoretical amino acid diversity of immune repertoires is $\approx 10^{140}$ (calculated as $20^{110} \times 2$). The calculation takes into account the 20 unique amino acids, the 110 amino acids long variable region of immune receptors, and the 2 variable regions composing each receptor (IGV_L-IGV_H or TCRVα-TCRVβ) (25). However, this enormous diversity is restricted in humans and mice by a starting set of V, D, and J gene segments leading to a potential diversity of about 10^{13}–10^{18} (3–6, 26–30). Only a fraction of the potential diversity is represented at any point in time in any given individual: the number of B- and T-cells is restricted (human: $10^{11–12}$) and the number of different clones, depending on clone definitions, reaches about 10^9 in humans and $10^{6–7}$ in mice (3, 5, 6, 31). The study of immune repertoire diversity ranges from the study of (i) the diversity of the building blocks of immune repertoires (V,

D, and J segments) and antibody lineage reconstruction (ii) to the mathematical modeling of VDJ recombination and (iii) to the estimation of the theoretical and biologically available repertoire frequency diversity (32). Together, these subfields of repertoire diversity analysis have expanded our analytical and quantitative insight into the creation of naive and antigen-driven antigen receptor diversity.

Accurate quantification of repertoire diversity relies first and foremost on the correct annotation of sequencing reads. Read annotation encompasses multiple steps: (i) calling of V, D, and J segments, (ii) subdivision into framework (FR) and complementarity-determining regions (CDRs), (iii) identification of inserted and deleted nucleotides in the junction region, and (iv) the quantification of the extent of somatic hypermutation (for antibodies). VDJ annotation tools were recently reviewed by Greiff et al. and Yaari and Kleinstein (17, 24). An updated version is currently maintained on the B-T.CR forum.[1] The B-T.CR forum is an AIRR-seq community platform for community-edited Wiki pages related to data sets and analysis tools as well as scientific exchange on current relevant topics in AIRR-seq (33, 34).

Accurate antigen receptor germline gene genotyping is crucial for predicting adaptive immunity (personalized and precision medicine) in the genetically diverse human population (30, 35–38). All VDJ annotation tools rely, at least partly, on a reference database of germline gene alleles. A reference database that is not identical to that of the individual from which the sequencing data is being annotated bears the potential of inaccurate annotation. This could affect, for example, the accuracy of the calling of V, D, and J genes and alleles as well as the quantification of somatic hypermutation. Antibody gene allele variation has also been linked to differential effectiveness of the humoral immune response (30, 35). Indeed, an increasing number of human germline gene alleles—representing one or several single-nucleotide polymorphisms—has been recently detected (30, 37, 39–41). These discoveries call into question the widely adopted practice of using one central germline reference database containing a more or less static set of non-personalized germline gene alleles. To address this problem, Corcoran et al. developed a software package (IgDiscover), which employs a cluster identification approach to reconstruct *de novo* from an AIRR-seq data set the corresponding V-gene germline database—all without *a priori* knowledge of existing germline gene databases (36). By doing so, they detected extensive individual germline gene differences among rhesus macaques (36). Complementarily, Gadala-Maria et al. developed TiGER (Tool for Ig Genotype Elucidation via Rep-Seq), which detects novel alleles based on the mutation pattern analysis (37). In contrast to IgDiscover, TiGER uses initial VDJ allele assignments with existing databases and software. Extending the analysis of germline gene diversity to the population level, Yu et al. built Lym1K, which is a database that combines validated alleles with novel alleles found in the 1000 Genomes Project (42, 43). In addition to database-centered approaches, probabilistic annotation enabled the detection of novel IgV genes and led to the discovery that substitution and mutation processes are (although reproducible across individuals) segment and allele dependent, thus further refining VDJ annotation and downstream diversity measurement (4, 44–47).

As a direct application in fundamental immunology, the advent of AIRR-seq has enabled the opportunity to describe quantitatively the statistical properties of VDJ recombination. Indeed, the ability to generate large data sets allowed several studies to show evidence of biases in VDJ recombination, as some germline gene frequencies (and combinations thereof) were found to occur more often than others (6, 21, 48–50). To mathematically model the process of VDJ recombination in both B- and T-cells, Elhanati et al. and Murugan et al. have employed techniques borrowed from statistical physics (maximum entropy, Hidden Markov, and probabilistic models) (4, 5, 45) to uncover the amount of diversity information inherent to each part of antibody and TCR sequences (entropy decomposition). VDJ recombination probability inference was mostly performed on non-productive sequences (e.g., out-of-frame, containing stop codon) as these receptors were assumed to be exempt from selection, thus representing unselected products of the generation process (4).

The deep sequence coverage of AIRR-seq has also led to the discovery of public clones or clonotypes—sequences that are shared across two or more individuals (6, 51–54). The existence of naive and antigen-associated public clones signifies a predetermined reduction in *a priori* genetic and antigen-driven immune receptor diversity (6). Although the exact definition of what constitutes a "public clone" is debatable (55), advancements have been made in understanding the generation and structure of public B- and T-cell clonotypes. By quantifying VDJ recombination probabilities as described above, Elhanati et al. have suggested that the emergence of public clonotypes is a direct consequence of the underlying VDJ recombination bias (56). The inference of VDJ recombination statistics of naive B- and T-cell populations may be of use in vaccination studies for helping distinguish public *antigen-specific* clonotypes from *genetically (naive) predetermined* ones. If feasible, such an approach might render the need of a healthy control cohort for determining *naive* public clones superfluous (47, 57). Complementarily, Greiff et al. have demonstrated extensive VDJ recombination bias by support vector machine analysis. Specifically, it showed that both public and private clones possess predetermined sequence signatures independent of mouse strain, species, and immune receptor type (antibody, TCR). These sequence signatures were found in both naive and antigen-selected B-cell compartments, which might suggest that naive recombination bias exerts a stronger diversity-constricting effect than antigen-driven evolution (58).

While the above-described methods of immune repertoire diversity analysis are relatively new, the quantification and comparison of clonotype diversity have been already studied in the era preceding high-throughput sequencing platforms by borrowing and adapting from mathematical ecology (59–62). The first step to quantifying clonal repertoire diversity is the definition of clonotype. Definitions of clonotype used in the

[1] http://b-t.cr/t/list-of-v-d-j-annotation-software/18.

literature range from the exact amino acid CDR3 to clusters of (e.g., CDR3) sequences to the sequence of entire variable chain regions (IGV_L-IGV_H or TCRVα-TCRVβ) using methods ranging from likelihood-based lineage inference to distance-based measures. A complete list of clonotyping tools has been compiled on the B-T.CR forum.[2] The debate on what constitutes a clonotype is ongoing and beyond the scope of this review. The interested reader is kindly referred to two extensive reviews (17, 63) and a recent report by Nouri and Kleinstein, who have developed a flexible user-defined method for clonotype identification (64).

To measure clonotype diversity, diversity indices are used [detailed reviews on diversity indices have been recently published in Ref. (17, 24)]. Briefly, diversity indices enable the comparison of repertoire diversity by parameterizing the repertoire space. They thus overcome the problem of clonally distinct repertoires (65). Several dedicated software packages exist for diversity index calculations (66–69). Briefly, the Diversity ($^{\alpha}D$) of a repertoire of S clones is usually calculated as follows: $^{\alpha}D = \left(\sum_{i=1}^{S} f_i^{\alpha} \right)^{\frac{1}{1-\alpha}}$ (Hill-Diversity), where f_i is the frequency of the ith clone weighted by the parameter α. Special cases of this Diversity function correspond to popular diversity indices in the immune repertoire field: species richness ($\alpha = 0$), and the exponential Shannon-Weiner ($\alpha \to 1$), inverse Simpson ($\alpha = 2$), and Berger–Parker indices ($\alpha \to \infty$). The higher the value of alpha, the higher is the influence of the more abundant clones on $^{\alpha}D$. Thus, each $^{\alpha}D$ value captures a different region (clonal subset) of the clonal frequency distribution (65). Due to the mathematical properties of the Diversity function [Schur concavity (70)], different repertoires may yield *qualitatively* different $^{\alpha}D$ values depending on the Diversity index used [Figure 1 in Greiff et al. (65)]. Therefore, for any discriminatory diversity comparisons, at least two Diversity indices should be considered. Diversity *profiles*, which are collections (vectors) of several Diversity indices, have been suggested to be superior to *single* diversity indices, when comparing clonal diversity (65, 66, 71). Using hierarchical clustering, α-parametrized diversity profiles have been shown to faithfully capture the shape of a repertoire's underlying clonal frequency distribution, which represents the state of clonal expansion (65). Thus, diversity profiles can serve as a parameterized proxy for a repertoire's state of clonal expansion. In addition, Mora and Walczak showed that the Rényi entropy (the mathematical foundation of Hill-Diversity profiles) can be constructed, in some cases, from rank-frequency plots (72), thereby establishing a direct mathematical link between clonal frequency distribution and diversity indices. Another interesting novel diversity analysis method is the *clonal plane* and the *polyclonal monoclonal diversity* index developed by Afzal et al. (73). These two related mathematical concepts represent repertoire diversity in a coordinate system spanned by species richness and evenness. This allows a visually straightforward identification of polyclonal and oligoclonal samples.

Although clonal frequency distributions, in most cases, cannot be compared directly across individuals due to restricted clonal overlap, their mathematical description has been the object of several studies. Specifically, clonal frequency distributions were found to be power-law distributed, with a few abundant clones, and a large number of lowly abundant clones (65, 74–76). Furthermore, Schwab et al. showed analytically *via* numerical simulations that Zipf-like distributions, a subclass of power-law distributions arise naturally if fluctuating unobserved variables affect the system (e.g., a variable external antigen environment influencing the observed antibody repertoire) (77). Indeed, it could be shown that clonotype diversity (or state of clonal expansion) contains antigen-associated information on the host immune status (6, 65, 78).

Given the heavy-tailed distribution of clonal frequencies (large number of low-abundant clones), comprehensive sampling of repertoires is challenging to achieve, thereby hindering cross-sample diversity comparison (65, 74, 77, 79). Indeed, diversity indices are highly sensitive to sample size variation caused by varying PCR and sequencing accuracy and biological and technological sampling depth (60, 62, 80). In general, two main approaches are used to mitigate sampling effects. (i) For the comparison of any two repertoires of unequal sampling size, Venturi et al. devised the following strategy: (a) sequencing reads are drawn randomly n-times without replacement from the repertoire with higher sampling depth (higher cell number and/or higher sequencing depth). (b) The desired diversity measure for each bootstrapped immune repertoire data set is then computed. (c) From the distribution of n diversity measures, the median diversity measure is estimated and compared with the smaller data set. This approach, however, does not aim to estimate the true underlying diversity of a cellular compartment (e.g., B- and T-cell developmental stages, antigen-specific repertoire). (ii) Inferring the true diversity of a repertoire is equivalent to the "missing-species problem," which describes the challenge to estimate the number of clones ("species") that have been missed in the sampling step. The quantification of missing (or unseen) species may be performed using diversity index estimators (60, 81, 82). These estimators attempt to estimate the number of missing receptors based on a more or less narrow region of the clonal frequency distribution's tail. A dedicated diversity estimator, adapted to the microevolutionary and high-diversity case of immune repertoires, was published by Laydon et al. They developed a rarefaction-based method called DivE, for estimating total repertoire size (species richness) (82, 83), which they showed to be both superior to common estimators of species richness such as Chao1 (81, 83) and Good-Turing (60, 84) and capable of estimating a repertoire's underlying clonal frequency distribution. Complementarily, Kaplinsky and Arnaout developed a maximum likelihood (ML) clone-size distribution-independent algorithm called Recon (reconstruction of estimated clones from observed numbers) that does estimate not only species richness but also any Hill-diversity measure (80). In general, however, gold-standard procedures for estimating repertoire diversity in various sampling scenarios are non-existent. A meta-study benchmarking current diversity index estimators on simulated immune repertoires will be needed to establish reliable guidelines for diversity estimation.

[2]http://b-t.cr/t/list-of-b-cell-clonal-identification-software/22.

To compare differences between diversity profiles, one should also consider resampling strategies as implemented in the R package *Change-O* by Gupta et al. These allow the determination of confidence areas around each diversity profile (66, 85) in the presence of differently sized repertoires. Accurate diversity calculation in case of incomplete sampling is of special importance when gaining information on human repertoires, which are often restricted to the isolation of a limited number of B- and T-cells from peripheral blood (17, 83, 86, 87).

Although the quantification of diversity is one of the more mature subfields of computational repertoire immunology, numerous open questions remain: (i) diversity has been measured from many different perspectives (germline gene diversity, state of clonal expansion, clonal size), thus capturing different dimensions of the repertoire diversity space. Is it possible to devise a universal metric that synthesizing different aspects of immune repertoire diversity into one? Such a metric would be very useful for repertoire-based immunodiagnostics. (ii) Hidden Markov and Bayesian (probabilistic) approaches have been used for modeling VDJ recombination. Those approaches, however, capture only short-range sequence interactions. Therefore, recurrent neural network approaches might be more appropriate to model the immune repertoire sequence space given their ability to account for sequence interactions of arbitrary length (88, 89)? (iii) Finally, we still have only very superficial insight into the biological diversity of antigen-specific repertoires and the combination rules of IGV_L/IGV_H and TCRVα1/TCRVβ chains due to the lack of large-scale data (76, 90–93). Once more extensive data have become available, can we leverage machine learning to uncover the underlying structure of antigen-specific repertoires and the prediction rules of chain pairing? Uncovering these immunological prediction rules is crucial for the knowledge-based development of antibody and T-cell-based immunotherapeutics.

RESOLVING THE SEQUENCE SIMILARITY ARCHITECTURE OF IMMUNE REPERTOIRES

The entirety of similarity relations among immune receptor sequences is called the similarity architecture of an immune repertoire. Thus, unlike immune repertoire diversity, which is based on the frequency profiles of immune clones, sequence similarity architecture captures frequency-independent clonal sequence similarity relations. The similarity among immune receptors directly influences antigen recognition breadth: the more dissimilar receptors are, the larger is the antigen space covered. Given the genetic, cellular, and clonal restrictions of immune repertoire diversity, the similarity architecture of antibody and T-cell repertoires has been a longstanding question and has only recently begun to be resolved. Understanding the sequence architecture of immune repertoires is, for example, crucial in the context of antibody therapeutics discovery for the conception of naive antibody libraries and synthetic repertoires that recapitulate natural repertoires (94).

One powerful approach to interrogate and measure immune repertoire architecture is network analysis (**Figures 1B** and **2B**) (94–100). Networks allow interrogation of sequence similarity and thereby add a complementary layer of information to repertoire diversity analysis. Clonal networks are built by defining each clone (nucleotide or amino acid sequence) as a node (**Figure 1B**). An edge between clones is drawn if they satisfy a certain similarity condition, which is predefined *via* a string distance [e.g., Levenshtein distance (LD)], resulting in undirected Boolean networks (94–97, 99, 100). The default distance is usually 1 nucleotide or 1 amino acid difference, but larger distances have also been explored (94). Thus, the construction of clonal networks requires the calculation of an all-by-all distance matrix. While the complete distance matrix can be computed on a single machine with repertoires of clone sizes <10,000, it becomes computationally expensive in terms of time and memory to calculate networks of clone sizes that exceed 10^5 clones, which is the size of many repertoires in both mice and humans (3, 5, 6). Therefore, Miho et al. have developed a high-performance computing pipeline (*imNet*), which can compute distance matrices and construct corresponding large-scale repertoire networks (94). This method led to the biological insight that antibody repertoire networks are, in contrast to other systems (101, 102), resistant to subsampling, which is of great importance for the network analysis of human repertoires where limited access to B-cell populations and lymphoid organs restricts complete biological sampling (17, 86). Although networks of a few thousand nodes may be visualized using software suites such as igraph (103), networkx (104), gephi (105), and cytoscape (106), interpretation of the visual graphics is not informative for networks beyond the clonal size of 10^3 (94). Furthermore, visualization of networks provides only marginal quantitation of the network similarity architecture, thus limiting the quantitative understanding of immune repertoires. Graph properties and network analysis have been recently employed to quantify the network architecture of immune repertoires (94, 100). Architecture analytics may be subdivided into properties that capture the repertoire at the global level (generally one coefficient per network) and those that describe the repertoire at the clonal and thus local level (one coefficient per clone per repertoire, vector of size equal to the clone size) (94).

Global coefficients are, for example, degree distribution, clustering coefficient, diameter, and assortativity (94). The degree of a node is the number of its edges (i.e., the number of similar clones to a certain clone), and a repertoire's degree distribution quantifies the abundance of node degrees (i.e., clonal similarities) across clones of a repertoire. This degree distribution has been used to describe and classify the networks by type, such as power law (a few highly connected clones and many clones with few connections), which is reminiscent of antigen-driven clonal expansion, or exponential (more even degree distribution across clones, covering extensive sequence space), which is more reflective of naive repertoires (94). The degree distribution thus provides insights into the overall distribution of connectedness (clonal similarities) within a repertoire and its state of clonal sequence expansion. Local characterization allows for the interrogation and correlation of additional clonal-related features, such as frequency and antigen specificity, within the immune repertoire architecture. Local parameters are, for example, degree, authority, closeness,

betweenness, and PageRank (94). PageRank, for instance, measures the importance of the similarity between two CDR3 clones within the network. Detailed mathematical descriptions of available network parameters have been described elsewhere (94, 107, 108).

Complementary to networks, which provide a discrete characterization of repertoires, similarity indices, similarity indices have been devised that provide a continuous description of repertoire architecture by quantifying the similarity between all sequences of a repertoire (using distance metrics) on a scale ranging from 0 (zero similarity) to 100% (all sequences are 100% identical) (6, 109). In addition to sequence similarity, the index by Strauli and Hernandez takes the frequency of each sequence into account, thus normalizing sequencing similarity by the frequency of each of the pairwise compared sequences (109).

The assessment of repertoire architecture has only recently started to transition from the visual investigation of clusters of immune receptor sequences to the construction of large-scale networks and the truly quantitative analysis of entire repertoires across similarity layers (>1 amino acid/nucleotide differences). This advance enabled the discovery of fundamental properties of repertoire architecture such as reproducibility, robustness, and redundancy (94). Although the biological interpretation of the mathematical characterization of immune repertoire networks is at an early stage, the universal use of network analysis in the deconvolution of complex systems (107, 108) suggests a great potential in immune repertoire research. Many important questions remain: (i) How can network repertoire architecture be compared across individuals without condensing networks into network indices and potentially losing information? Thus, can discrete and continuous representation of repertoire architecture be merged into one comprehensive mathematical framework? (ii) Can the linking of networks across similarity layers serve to understand the dynamic and potential space of antigen-driven repertoire evolution (94)? (iii) Is the network structure that is observed on the antibody immunogenomic level also maintained on the phenotypic and immunoproteomics level of serum antibodies (110–116)?

RETRACING THE ANTIGEN-DRIVEN EVOLUTION OF ANTIBODY REPERTOIRES

Upon antigen challenge, B-cells expand and hypermutate their antibody variable regions, thus forming a B-cell lineage that extends from the naive unmutated B-cells, to somatically hypermutated memory B-cells (25), to terminally differentiated plasma cells (11). Somatic hypermutation is unique to B-cells and absent in T-cells. Retracing antibody repertoire evolution enables insights into how vaccines (78) and pathogens shape the humoral immune response (117–119).

To infer the ancestral evolutionary relationships among individual B-cells, lineage trees are constructed from the set of sequences belonging to a clonal lineage (**Figures 1C** and **2C**). A clonal lineage is defined as the number of receptor sequences originating from the same recombination event. For building a lineage tree, a common preprocessing step is to group together all sequences with identical V and J genes and CDR3 length. Schramm et al. published a software for the ontogenetic analysis of antibody repertoires, which is designed to enable the automation of antibody repertoire lineage analysis. Importantly, it provides interfaces to phylogenetic inference programs such as BEAST and DNAML (120).

In antibody repertoire phylogenetics, there is no consensus as to which phylogenetic method is optimal for the inference of lineage evolution (17, 121). Most of the current phylogenetic methods rely on assumptions that may be true for species evolution but might be invalid for antibody evolution. One prominent example is the assumption that each site mutates independently of the neighboring nucleotides, which is not the case in antibody evolution (121). In addition, antibodies evolve on time scales that differ by several orders of magnitudes from those of species. These two factors likely decrease the accuracy of clade prediction (clade: set of descendent sequences that all share a common ancestor), thus potentially impacting antibody phylogenetic studies.

Several phylogenetic methods, such as LD, neighbor joining (NJ), maximum parsimony (MP), ML, and Bayesian inference (BEAST), have been used for delineating the evolution of B-cell clonal lineages from antibody repertoire sequencing data (85, 122–124). For general information regarding the methods, refer to the review by Yang and Rannala (125). Briefly, both LD and NJ are distance-based methods that rely upon an initial all-by-all distance matrix calculation and have been implemented in many computational platforms (Clustal, T-REX) and R packages (ape, phangorn) (126–129). Even in the event $>10^5$ sequences per sample, the distance matrix calculation in phylogenetics poses less of a problem than in network analysis since a sample's sequences are grouped by lineage members of identical V–J gene and CDR3 length, thus reducing computational complexity. The relatively short computation time of distance-based methods renders them particularly useful for initial data exploration (125). MP attempts to explain the molecular evolution by non-parametrically selecting the shortest possible tree that explains the data (24). MP trees can be produced using several available tools (e.g., PAUP, TNT, PHYLIP, Rphylip) (130–133). Both ML and BEAST infer lineage evolution using probabilistic methods, which can incorporate biologically relevant parameters such as transition/transversion rate and nucleotide frequencies. A variety of ML tools have been developed (e.g., PhyML, RAxML, and MEGA) (134–136). While multiple phylogenetic tools utilizing Bayesian methods exist (137, 138), this review focuses on BEAST given its recurrent use in antibody repertoire studies (120, 124, 139–141). BEAST traditionally employs a Markov chain Monte Carlo algorithm to explore the tree parameter space. This computationally expensive process limits the practical number of sequences per lineage tree to $<10^3$. Despite the extensive computational requirements (both in memory and in run time), BEAST has the advantage of producing time-resolved phylogenies and inferring somatic hypermutation rates (138, 139). The BEAST framework shows, therefore, the highest scientific benefit when applied to experiments examining antibody evolution within the same host across multiple sampling time points (124), as inferred mutation rates and tree heights (duration of evolution) are reported in calendar time.

Yermanos et al. have compared five of the most common phylogenetics reconstruction methods for antibody repertoire analysis in terms of their absolute accuracy and their concordance in clade assignment using both experimental and simulated antibody sequence data (139). Correctly inferring the clades of a phylogenetic tree is crucial for describing the evolutionary relationship between clonally selected and expanded B-cells (i.e., memory B-cells) that belong to a given lineage (i.e., derived from a naive B-cell). Phylogenetic trees inferred by the methods tested (LD, NJ, ML, MP, BEAST) resulted in different topologies as measured by both clade overlap (number of internal nodes sharing the same descendant sequences) and treescape metric (comparison of the placement of the most recent common ancestor of each pair of tips in two trees) (142). These results suggest caution in the interpretation and comparison of results from the phylogenetic reconstruction of antibody repertoire evolution (139).

The accurate reconstruction of antibody phylogenetic trees is tightly linked to the detailed understanding of the physical and temporal dynamics of somatic hypermutation along antigen-driven antibody sequence evolution. Mutation statistics can be inferred probabilistically to account for the fact that the likelihood of mutation is not uniformly distributed over the antibody VDJ region (46, 47). For example, there is a preference to mutate particular DNA motifs called hotspots (length: 2–7 bp) and concentrated in the CDRs over others (coldspots) (4, 121, 143, 144). To uncover the sequence-based rules of somatic hypermutation targeting, Yaari et al. developed S5F, an antibody-specific mutation model. This model provides an estimation of the mutability and mutation preference for each nucleotide in the VDJ region of the heavy chain based on the four surrounding nucleotides (two on either side). The estimated profiles could explain almost half of the variance in observed mutation patterns and were highly conserved across individuals (121). Cui et al. have, in addition, reported two new models that add to the heavy-chain S5F model: the light-chain mouse RS5NF and the light-chain human S5F L chain model (145). In addition, Sheng et al. investigated the intrinsic mutation frequency and substitution bias of somatic hypermutations at the amino acid level by developing a method for generating gene-specific substitution profiles (146). This method revealed gene-specific substitution profiles that are unique to each human V-gene and also highly consistent between human individuals.

The existence of hotspot and coldspot mutation motifs violates the standard assumption of likelihood-based phylogenetics, which is that evolutionary changes at different nucleotide or codon sites are statistically independent. Furthermore, since hotspot motifs are, by definition, more mutable than non-hotspot motifs, their frequency within the B-cell lineage may decrease over time as they are replaced with more stable motifs (147). To explicitly parameterize the effect of biased mutation within a phylogenetic substitution model, Hoehn et al. developed a model that can partially account for the effect of context-dependent mutability of hotspot and coldspot motifs and explicitly model descent from a known germline sequence (148). The resulting model showed a substantially better fit to three well-characterized lineages of HIV-neutralizing antibodies, thus being potentially useful for analyzing the temporal dynamics of antibody mutability in the context of chronic infection. In addition, Vieira et al. assessed the evidence for consistent changes in mutability during the evolution of B-cell lineages (140). By using Bayesian phylogenetic modeling, they showed that mutability losses were about 60% more frequent than gains (in both CDRs and FRs) in anti-HIV antibody sequences (140).

Although computational methods tailored to the phylogenetic analysis of antibody evolution are slowly beginning to surface, many important problems remain. (i) First approaches in coupling clonal expansion information to the inference of phylogenetic trees have been developed (149). Will these additional layers of information enable a better prediction of antibody evolution? (ii) There has been progress in comparing the differences of antibody repertoires in the context of phylogenetic trees using the UniFrac distance measure (150, 151). Briefly, for a given pair of samples, UniFrac measures the total branch length that is unique to each sample. The comparison of tree *topologies*, however, remains a challenge. This is because each lineage tree is composed of a different number of sequences, and there are thousands, if not more, of simultaneously evolving lineages within a single host. Although methods exists for the comparison of unlabeled phylogenetic trees by, for instance, means of their Laplacian spectra (152), their application and ability to extract meaningful biological conclusions have not yet been realized. (iii) It is unclear to what extent antibody evolution differs between different acute and chronic viral infections, or different antigens. Specifically, is it possible to relate antigen-driven convergence and affinity (6, 50, 117) to phylogenetic antigen-specific signatures (153)?

DISSECTING NAIVE AND ANTIGEN-DRIVEN REPERTOIRE CONVERGENCE

Convergence (overlap) of immune repertoires describes the phenomenon of identical or similar immune receptor sequences shared by two or more individuals. Specifically, sequence convergence can either mean that (i) clones (public clonotypes, entire clonal sequence or clonotype cluster) or (ii) motifs (sequence substrings) are shared. Several researchers in the field have endeavored to quantify the extent of naive and antigen-driven repertoire convergence using a large variety of computational approaches that quantify cross-individual sequence similarity (6, 53, 78, 117, 154–156) (**Figures 1D** and **2D**). Repertoire convergence may be of substantial importance for the prediction and manipulation of adaptive immunity (6).

The simplest way to quantify sequence convergence is by clonotype overlap among pairwise samples expressed as a percentage normalized by the clonal size of either one or both of the samples compared (6, 48, 157). In case clonotypes are treated not as single sequences but clusters of sequences, clusters were defined as shared between samples if each sample contributed at least one sequence to the cluster (156). Overlap indices such as Morisita-Horn (158) add additional information to the measurement of clonal overlap by integrating the clonal frequency

of compared clones (62, 159, 160). A parameterized version of the Morisita-Horn index, similar to the Hill-diversity, may be used to weigh certain clonal abundance ranges differently (60). Rubelt and Bolen expanded on the idea of an overlap index by incorporating both binned sequence features (e.g., clone sequences, germline genes) and their frequency for measuring the impact of heritable factors on VDJ recombination and thymic selection. Their Repertoire Dissimilarity Index consists of a non-parametric Euclidian-distance-based bootstrapped subsampling approach, which enables the quantification of the average variation between repertoires (50, 161). Importantly, it accounts for variance in sequencing depth between samples. Another clone-based approach was developed by Emerson et al. who mined public TCRβ clonotypes in CMV-positive and CVM-negative individuals to predict their CMV status. To this end, they identified CMV-associated clonotypes by using Fisher's exact test. Subsequently, these clonotypes were used within the context of a probabilistic classifier to predict an individual's CMV status. The classifier used dimensionality reduction and feature selection to mitigate the influence of the variance of HLA types across individuals because the distribution of TCRβ clones is HLA dependent (154).

Moving from the clonal to the subsequence level, several groups compared the average distance between repertoires based on their entire sequence diversity (without predetermining feature bins). Specifically, Yokota et al. developed an algorithm for comparing the similarity of immune repertoires by projecting the high-dimensional intersequence relations, calculated from pairwise sequence alignments, onto a low-dimensional space (162). Such low-dimensional embedding of sequence similarity has the advantage of enabling the identification of those sequences that contribute most to intersample (dis)similarity. As previously described, Strauli and Hernandez quantified sequence convergence between repertoires in response to influenza vaccination not only by incorporating genetic distance (Needleman-Wunsch algorithm) but also by incorporating the frequency of each clonal sequence (109). Their approach relies on a statistical framework called functional data analysis (FDA), which is often used for gene expression analysis. In their implementation, FDA models each sample as a continuous function over sampling time points and is thus suitable for the analysis of sequence convergence over a time course experiment. The FDA framework has the advantage of accounting for uneven time point sampling and measurement error, both of which are common characteristics of immune repertoire data sets (2, 17). Bürckert et al. also employed a method borrowed from gene expression analysis (DESeq2) (163) to select for clusters of CDR3s, which are significantly overrepresented within different cohorts of immunized animals (164). These clusters exhibited convergent antigen-induced CDR3 signatures with stereotypic amino acid patterns seen in previously described tetanus toxoid and measles-specific CDR3 sequences.

Given the high-dimensional complexity of the immune repertoire sequence space, sequence distance-based approaches might not suffice for covering the entire complexity of sequence convergence. A greater portion of the sequence space may be covered by sequence-based machine learning (artificial intelligence). Here, the idea is that sequence signatures and motifs are shared between individuals belonging to a predefined class (e.g., different immune status). Sun et al. discriminated the TCRβ repertoire of mice immunized with and without ovalbumin with 80% accuracy by deconstructing it into overlapping amino acid k-mers (165). Sun Cinelli et al. used a one-dimensional Bayesian classifier for the selection of features, which were subsequently used for support vector machine analysis (166). As a third machine learning alternative, Greiff et al. leveraged gapped-k-mers and support vector machines for the classification of public and private clones with 80% accuracy from antibody and TCR repertoires of human and mice. This study used overlapping k-mers to construct sequence prediction profiles, which highlight those convergent sequence regions that contribute most to the identity of a class (public/private clones but also, e.g., also different immune states and antigen specificities) (58). Beyond k-mers, several groups have exploited the addition of additional information such as physicochemical properties (Atchley and Kidera factors) to provide more extensive information to machine learning algorithms (167–171). Finally, a machine learning independent approach using local search graph theory for the detection of disease-associated k-mers was recently published by Apeltsin et al. (172).

One of the longest standing challenges in immunology is whether it is possible to predict antigen specificity from the sequence of the immune receptor (2, 15, 173–175). Sequence-dependent prediction implies that immune receptor sequences specific to one antigen share exclusive sequence signatures (motifs) or have higher intraclass than interclass similarity (class = antigen). Two investigations towards sequenced-based specificity prediction using sequence similarity (sequence distance) approaches have recently been reported (155, 176). In one example, Dash et al. developed a distance measure called TCRDist, which is guided by structural information on pMHC binding (155). Two TCRs sequences were compared by computing a similarity-weighted Hamming distance between CDR sequences, including an additional loop between CDR2 and CDR3. TCRDist was used to detect clusters of highly similar, antigen-specific groups of TCRs that were shared across different mouse or human samples. To predict the antigen specificity of a TCR, it was assigned to the cluster to which it had the highest similarity (as based on the TCRDist), resulting in highly accurate prediction (155). By using a similar approach, Glanville et al. developed GLIPH, a tool that identifies TCR specificity groups using a three-step procedure: (i) determining of shared motifs and global similarity, (ii) clustering based on local and global relationships between TCRs, and (iii) analyzing the enrichment for common V-gene, CDR3 lengths, clonal expansion, shared HLA alleles in recipients, motif significance, and cluster size. This approach yielded also highly accurate prediction of antigen-specific TCRs and led to the design of synthetic TCRs (not existing in biological data) that retained antigen specificity (176).

One of the biggest bottlenecks of learning the underlying principles of antigen-driven repertoire convergence is the scarcity of antigen-specific sequence data. This is not only a problem for machine learning but also a problem for network-based approaches, where one wishes to map antigen-specific information onto generated networks (94, 100). To address this issue for T-cells, Shugay et al.

(VDJDB) and Tickotsky et al. (McPAS-TCR) have built dedicated and curated databases. VDJDB gathers >20,000 unique TCR sequences from different species associated with their epitope (>200) and MHC context (177). McPAS-TCR contains more than 5,000 pathogen-associated TCRs from humans and mice (178). For antibodies, Martin has conceived Abysis, which encompasses >5,000 sequences of known function (from literature) from many species (>15) along with, where available, PDB 3D structure information (179). Finally, the Immune Epitope Database has also started capturing epitope-specific antibody and TCR information (>20,000 and >2,000 epitopes) (180).

Significant progress in the understanding of antigen-associated signatures has been made. However, several long-standing questions remain to be answered. (i) The emergence of antigen-driven convergence and phylogenetic evolution are inherently linked. Is it feasible to model both phenomena in a unified computational environment similarly to recent efforts in coupling phylogenetics with the understanding of somatic hypermutation patterns (140, 148)? (ii) Can recently developed models for the inference of VDJ recombination patterns and selection factors be applied to the analysis of antigen-associated sequence signatures (4, 56)? (iii) Do more advanced sequence-based machine learning techniques such as deep neural networks, capable of capturing long-range sequence interactions (out of reach for k-mer-based approaches), improve modeling of the epitope and paratope space (89, 181–186)?

CONCLUSION

The toolbox of computational immunology for the study of immune repertoires has reached an impressive richness leading to remarkable insights into B- and T-cell development and selection (6, 52, 56, 187, 188), disease, infection, and vaccine profiling (78, 85, 117, 189–192), propelling forward the fields of immunodiagnostics and immunotherapeutics (65, 118, 193). Here, we have discussed computational, mathematical, and statistical methods in the light of underlying assumptions and limitations. Indeed, although considerably matured over the last few years, the field still faces several important and scientifically interesting problems. (i) There exist only few platforms to benchmark computational tools, thus hindering the standardization of methodologies. Recently, a consortium of scientists working in AIRR-seq has convened to establish and implement consensus protocols and simulation frameworks[3] (2, 17, 33, 34, 43, 194, 195). (ii) With the exponential increase of both bulk and single-cell data (90, 196), the scalability of computational tools is becoming progressively important. Although advances in this regard have been made in sequence annotation, clonotype clustering, and network construction (64, 94, 197, 198), further efforts especially in the field of phylogenetics are necessary to infer the evolution of large-scale antibody repertoires (139). (iii) Although there exist many approaches, which capture parts of the immune repertoire complexity, a computational approach for the synthesis of many dimensions of the repertoire space at once is missing thus hindering a high-dimensional understanding of the adaptive immune response. (iv) Very few attempts exist yet, which aim to link immune receptor and transcriptomics data (199, 200). Recently, computational tools have been developed that can extract immune receptor sequences from bulk and single-cell transcriptomic data (197, 200–204). Linking immune repertoire and transcriptome may provide a deeper understanding of how antibody and T-cell specificity are regulated on the genetic level with profound implications for synthetic immunology (205–207). (v) Many methods capture a static space of repertoires, but few methods create *predictive* quantitative knowledge. Increasing the predictive performance of computational methods will help in the antibody discovery from display libraries and immunizations and the design of vaccines and immunodiagnostics (15, 19, 208–210).

AUTHOR CONRIBUTIONS

VG and STR conceived and designed the review. All authors wrote the review.

FUNDING

This work was funded by the Swiss National Science Foundation (Project #: 31003A_170110, to SR), SystemsX.ch – AntibodyX RTD project (to SR); European Research Council Starting Grant (Project #: 679403 to SR). The professorship of STR is made possible by the generous endowment of the S. Leslie Misrock Foundation. We are grateful to ETH Foundation for the Pioneer Fellowship to Enkelejda Miho.

[3]http://airr.irmacs.sfu.ca/.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302:575–81. doi:10.1038/302575a0
2. Wardemann H, Busse CE. Novel approaches to analyze immunoglobulin repertoires. *Trends Immunol* (2017) 38(7):471–82. doi:10.1016/j.it.2017.05.003
3. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106:20216–21. doi:10.1073/pnas.0909775106
4. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. *Phil Trans R Soc Lond B Biol Sci* (2015) 370:20140243. doi:10.1098/rstb.2014.0243
5. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* (2012) 109:16161–6. doi:10.1073/pnas.1212755109
6. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S, et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep* (2017) 19:1467–78. doi:10.1016/j.celrep.2017.04.054
7. Burnet FM. Theories of immunity. *Perspect Biol Med* (1960) 3:447–58. doi:10.1353/pbm.1960.0034
8. Ahmed R, Gray D. Immunological memory and protective immunity: understanding their relation. *Science* (1996) 272:54–60. doi:10.1126/science.272.5258.54
9. Hammarlund E, Lewis MW, Carter SV, Amanna I, Hansen SG, Strelow LI, et al. Multiple diagnostic techniques identify previously vaccinated

individuals with protective immunity against monkeypox. *Nat Med* (2005) 11:1005–11. doi:10.1038/nm1273
10. Amanna IJ, Carlson NE, Slifka MK. Duration of humoral immunity to common viral and vaccine antigens. *N Engl J Med* (2007) 357:1903–15. doi:10.1056/NEJMoa066092
11. Manz RA, Thiel A, Radbruch A. Lifetime of plasma cells in the bone marrow. *Nature* (1997) 388:133–4. doi:10.1038/40540
12. Landsverk OJB, Snir O, Casado RB, Richter L, Mold JE, Réu P, et al. Antibody-secreting plasma cells persist for decades in human intestine. *J Exp Med* (2017) 214(2):309–17. doi:10.1084/jem.20161590
13. Halliley JL, Tipton CM, Liesveld J, Rosenberg AF, Darce J, Gregoretti IV, et al. Long-lived plasma cells are contained within the CD19−CD38hiCD138+ subset in human bone marrow. *Immunity* (2015) 43(1):132–45. doi:10.1016/j.immuni.2015.06.016
14. Pollok K, Mothes R, Ulbricht C, Liebheit A, Gerken JD, Uhlmann S, et al. The chronically inflamed central nervous system provides niches for long-lived plasma cells. *Acta Neuropathol Commun* (2017) 5:88. doi:10.1186/s40478-017-0487-8
15. Calis JJA, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol* (2014) 35:581–90. doi:10.1016/j.it.2014.09.004
16. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32:158–68. doi:10.1038/nbt.2782
17. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol* (2015) 36:738–49. doi:10.1016/j.it.2015.09.006
18. Baum PD, Venturi V, Price DA. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* (2012) 42:2834–9. doi:10.1002/eji.201242999
19. Robinson WH. Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. *Nat Rev Rheumatol* (2014) 11:171–82. doi:10.1038/nrrheum.2014.220
20. Cobey S, Wilson P, Matsen FA. The evolution within us. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140235. doi:10.1098/rstb.2014.0235
21. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. High-throughput sequencing of the *zebrafish* antibody repertoire. *Science* (2009) 324:807–10. doi:10.1126/science.1170020
22. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A public database of memory and naive B-cell receptor sequences. *PLoS One* (2016) 11:e0160853. doi:10.1371/journal.pone.0160853
23. Kidd BA, Peters LA, Schadt EE, Dudley JT. Unifying immunology with informatics and multiscale biology. *Nat Immunol* (2014) 15:118–27. doi:10.1038/ni.2787
24. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* (2015) 7:121. doi:10.1186/s13073-015-0243-2
25. Janeway CA, Murphy K. *Janeway's Immunobiology*. 8th Revised Edition. Taylor & Francis (2011).
26. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) 92(4):530–46. doi:10.1016/j.ajhg.2013.03.004
27. Johnston CM, Wood AL, Bolland DJ, Corcoran AE. Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. *J Immunol* (2006) 176:4221–34. doi:10.4049/jimmunol.176.7.4221
28. Malissen M, Minard K, Mjolsness S, Kronenberg M, Goverman J, Hunkapiller T, et al. Mouse T cell antigen receptor: Structure and organization of constant and joining gene segments encoding the β polypeptide. *Cell* (1984) 37:1101–10. doi:10.1016/0092-8674(84)90444-6
29. Arden B, Clark SP, Kabelitz D, Mak TW. Human T-cell receptor variable gene segment families. *Immunogenetics* (1995) 42:455–500. doi:10.1007/BF00172176
30. Watson CT, Glanville J, Marasco WA. The individual and population genetics of antibody immunity. *Trends Immunol* (2017) 38(7):459–70. doi:10.1016/j.it.2017.04.003
31. Trepel F. Number and distribution of lymphocytes in man. A critical analysis. *J Mol Med* (1974) 52:511–5.
32. Granato A, Chen Y, Wesemann DR. Primary immunoglobulin repertoire development: time and space matter. *Curr Opin Immunol* (2015) 33:126–31. doi:10.1016/j.coi.2015.02.011
33. Breden F, Prak Luning TE, Peters B, Rubelt F, Schramm CA, Busse CE, et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front Immunol* (2017) 8:1418. doi:10.3389/fimmu.2017.01418
34. Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* (2017) 18(12):1274–8. doi:10.1038/ni.3873
35. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* (2016) 6:20842. doi:10.1038/srep20842
36. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
37. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* (2015) 112:E862–70. doi:10.1073/pnas.1417683112
38. Ralph DK, Matsen FA IV. Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. *Q-Bio* (2017). Available from http://arxiv.org/abs/1711.05843
39. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) 184:6986–92. doi:10.4049/jimmunol.1000445
40. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, et al. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol Baltim* (2012) 188:1333–40. doi:10.4049/jimmunol.1102097
41. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol Immunol* (2017) 87:12–22. doi:10.1016/j.molimm.2017.03.012
42. Yu Y, Ceredig R, Seoighe C. A Database of human immune receptor alleles recovered from population sequencing data. *J Immunol* (2017) 198(5):2202–10. doi:10.4049/jimmunol.1601710
43. Watson CT, Matsen FA, Jackson KJL, Bashir A, Smith ML, Glanville J, et al. Comment on "a database of human immune receptor alleles recovered from population sequencing data". *J Immunol* (2017) 198:3371–3. doi:10.4049/jimmunol.1700306
44. Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res* (2004) 32:W435–40. doi:10.1093/nar/gkh412
45. Elhanati Y, Marcou Q, Mora T, Walczak AM. repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics* (2016) 32:1943–51. doi:10.1093/bioinformatics/btw112
46. Ralph DK, Matsen FA IV. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput Biol* (2016) 12:e1004409. doi:10.1371/journal.pcbi.1004409
47. Marcou Q, Mora T, Walczak AM. IGoR: a tool for high-throughput immune repertoire analysis. *Q-Bio* (2017).Available from: http://arxiv.org/abs/1705.08246
48. Glanville J, Kuo TC, von Büdingen H-C, Guey L, Berka J, Sundar PD, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci U S A* (2011) 108:20066–71. doi:10.1073/pnas.1107498108
49. Reddy ST, Ge X, Miklos AE, Hughes RA, Kang SH, Hoi KH, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* (2010) 28:965–9. doi:10.1038/nbt.1673
50. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. *Nat Commun* (2016) 7:11112. doi:10.1038/ncomms11112

51. Shugay M, Bolotin DA, Putintseva EV, Pogorelyy MV, Mamedov IZ, Chudakov DM. Huge overlap of individual TCR beta repertoires. *T Cell Biol* (2013) 4:466. doi:10.3389/fimmu.2013.00466
52. Covacu R, Philip H, Jaronen M, Almeida J, Kenison JE, Darko S, et al. System-wide analysis of the T cell response. *Cell Rep* (2016) 14:2733–44. doi:10.1016/j.celrep.2016.02.056
53. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* (2014) 24:1603–12. doi:10.1101/gr.170753.113
54. Collins AM, Jackson KJL. On being the right size: antibody repertoire formation in the mouse and human. *Immunogenetics* (2017). doi:10.1007/s00251-017-1049-8
55. Castro R, Navelsaker S, Krasnov A, Du Pasquier L, Boudinot P. Describing the diversity of Ag specific receptors in vertebrates: contribution of repertoire deep sequencing. *Dev Comp Immunol* (2017) 75:28–37. doi:10.1016/j.dci.2017.02.018
56. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci U S A* (2014) 111:9875–80. doi:10.1073/pnas.1409572111
57. Pogorelyy MV, Minervina AA, Chudakov DM, Mamedov IZ, Lebedev YB, Mora T, et al. Method for identification of condition-associated public antigen receptor sequences. *Q-Bio* (2017). Available from: http://arxiv.org/abs/1709.09703
58. Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, et al. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J Immunol* (2017) 199:2985–97. doi:10.4049/jimmunol.1700594
59. Jost L. Entropy and diversity. *Oikos* (2006) 113:363–75. doi:10.1111/j.2006.0030-1299.14714.x
60. Rempala GA, Seweryn M. Methods for diversity and overlap analysis in T-cell receptor populations. *J Math Biol* (2013) 67:1–30. doi:10.1007/s00285-012-0589-7
61. Venturi V, Kedzierska K, Tanaka MM, Turner SJ, Doherty PC, Davenport MP. Method for assessing the similarity between subsets of the T cell receptor repertoire. *J Immunol Methods* (2008) 329:67–80. doi:10.1016/j.jim.2007.09.016
62. Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J Immunol Methods* (2007) 321:182–95. doi:10.1016/j.jim.2007.01.019
63. Hershberg U, Prak ETL. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Phil Trans R Soc Lond B Biol Sci* (2015) 370:20140239. doi:10.1098/rstb.2014.0239
64. Nouri N, Kleinstein SH. Performance-optimized partitioning of clonotypes from high-throughput immunoglobulin repertoire sequencing data. *bioRxiv* (2017). doi:10.1101/175315
65. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* (2015) 7:49. doi:10.1186/s13073-015-0169-8
66. Gupta NT, Heiden JV, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) 31(20):3356–8. doi:10.1093/bioinformatics/btv359
67. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. *Vegan: Community Ecology Package*. (2015). Available from: http://CRAN.R-project.org/package=vegan
68. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. tcR: An R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics* (2015) 16:175. doi:10.1186/s12859-015-0613-1
69. Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, et al. VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput Biol* (2015) 11:e1004503. doi:10.1371/journal.pcbi.1004503
70. Solomon DL. *Unit CUB, Biometrics CUD of, Biology CUD of BS and C. Biometrics Unit Technical Reports: Number BU-573-M: A Comparative Approach to Species Diversity*. (1975). Available from: http://ecommons.library.cornell.edu/handle/1813/32672
71. Snir O, Mesin L, Gidoni M, Lundin KEA, Yaari G, Sollid LM. Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J Immunol* (2015) 194:5703–12. doi:10.4049/jimmunol.1402611
72. Mora T, Walczak AM. Renyi entropy, abundance distribution and the equivalence of ensembles. *ArXiv Prepr ArXiv160305458* (2016). Available from: http://arxiv.org/abs/1603.05458
73. Afzal S, Gil-Farina I, Gabriel R, Ahmad S, von Kalle C, Schmidt M, et al. Systematic comparative study of computational methods for T-cell receptor sequencing data analysis. *Brief Bioinform* (2017) 1–13. doi:10.1093/bib/bbx111
74. Mora T, Walczak AM, Bialek W, Callan CG. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci U S A* (2010) 107:5405–10. doi:10.1073/pnas.1001705107
75. Oakes T, Heather JM, Best K, Byng-Maddick R, Husovsky C, Ismail M, et al. Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Front Immunol* (2017) 8:1267. doi:10.3389/fimmu.2017.01267
76. Grigaityte K, Carter JA, Goldfless SJ, Jeffery EW, Hause RJ, Jiang Y, et al. Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire. *bioRxiv* (2017). doi:10.1101/213462
77. Schwab DJ, Nemenman I, Mehta P. Zipf's law and criticality in multivariate data without fine-tuning. *Phys Rev Lett* (2014) 113:068102. doi:10.1103/PhysRevLett.113.068102
78. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* (2014) 16:105–14. doi:10.1016/j.chom.2014.05.013
79. Bolkhovskaya OV, Zorin DY, Ivanchenko MV. Assessing T cell clonal size distribution: a non-parametric approach. *PLoS One* (2014) 9:e108658. doi:10.1371/journal.pone.0108658
80. Kaplinsky J, Arnaout R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* (2016) 7:11881. doi:10.1038/ncomms11881
81. Chao A, Shen T-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ Ecol Stat* (2003) 10:429–43. doi:10.1023/A:1021993627070
82. Laydon DJ, Melamed A, Sim A, Gillet NA, Sim K, Darko S, et al. Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput Biol* (2014) 10:e1003646. doi:10.1371/journal.pcbi.1003646
83. Laydon DJ, Bangham CRM, Asquith B. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Phil Trans R Soc Lond B Biol Sci* (2015) 370:20140291. doi:10.1098/rstb.2014.0291
84. Good IJ. The population frequencies of species and the estimation of population parameters. *Biometrika* (1953) 40:237–64. doi:10.1093/biomet/40.3-4.237
85. Stern JNH, Yaari G, Heiden JAV, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med* (2014) 6:248ra107. doi:10.1126/scitranslmed.3008879
86. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res* (2011) 21:790–7. doi:10.1101/gr.115428.110
87. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, et al. An atlas of B-cell clonal distribution in the human body. *Nat Biotechnol* (2017) 35:879–84. doi:10.1038/nbt.3942
88. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* (1997) 9:1735–80. doi:10.1162/neco.1997.9.8.1735
89. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol* (2016) 12:878. doi:10.15252/msb.20156651
90. Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. Single-cell transcriptomics to explore the immune system in health and disease. *Science* (2017) 358:58–63. doi:10.1126/science.aan6828
91. DeKosky B. Paired VH:VL analysis of naïve B cell repertoires and comparison to antigen-experienced B cell repertoires in healthy human donors. *Decoding*

the Antibody Repertoire, Springer Theses (Springer International Publishing) (2017). p. 41–57. https://www.nature.com/articles/nm.3743
92. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31:166–69. doi:10.1038/nbt.2492
93. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-throughput pairing of T cell receptor α and β sequences. *Sci Transl Med* (2015) 7:ra131–301. doi:10.1126/scitranslmed.aac5624
94. Miho E, Greiff V, Roskar R, Reddy ST. The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis. *bioRxiv* (2017). doi:10.1101/124578
95. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) 23:1874–84. doi:10.1101/gr.154815.113
96. Ben-Hamo R, Efroni S. The whole-organism heavy chain B cell repertoire from *zebrafish* self-organizes into distinct network features. *BMC Syst Biol* (2011) 5:27. doi:10.1186/1752-0509-5-27
97. Chang Y-H, Kuan H-C, Hsieh TC, Ma KH, Yang C-H, Hsu W-B, et al. Network signatures of IgG immune repertoires in hepatitis B associated chronic infection and vaccination responses. *Sci Rep* (2016) 6:26556. doi:10.1038/srep26556
98. Hoehn KB, Gall A, Bashford-Rogers R, Fidler SJ, Kaye S, Weber JN, et al. Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Phil Trans R Soc Lond B Biol Sci* (2015) 370:20140241. doi:10.1098/rstb.2014.0241
99. Lindner C, Thomsen I, Wahl B, Ugur M, Sethi MK, Friedrichsen M, et al. Diversification of memory B cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nat Immunol* (2015) 16:880–8. doi:10.1038/ni.3213
100. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I, et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *Elife* (2017) 6:e22057. doi:10.7554/eLife.22057
101. Lee SH, Kim P-J, Jeong H. Statistical properties of sampled networks. *Phys Rev E* (2006) 73:016102. doi:10.1103/PhysRevE.73.016102
102. Sethu H, Chu X. A new algorithm for extracting a small representative subgraph from a very large graph. *Phys* (2012). Available from: http://arxiv.org/abs/1207.4825
103. Csardi G, Nepusz T. The igraph software package for complex network research, complex system. *InterJournal* (2006) 1695.
104. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using networkx. In: Varoquaux G, Vaught T, Millman J, editors. *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA (2008). p. 11–5.
105. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. *ICWSM* (2009) 8:361–2.
106. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* (2003) 13:2498–504. doi:10.1101/gr.1239303
107. Albert R, Jeong H, Barabasi A-L. Error and attack tolerance of complex networks: article: nature. *Nature* (2000) 406:378–82. doi:10.1101/187120
108. Barabási A-L. *Network science*. Boston, USA: Cambridge University Press (2016).
109. Strauli NB, Hernandez RD. Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Med* (2016) 8:60. doi:10.1186/s13073-016-0314-z
110. Wine Y, Boutz DR, Lavinder JJ, Miklos AE, Hughes RA, Hoi KH, et al. Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proc Natl Acad Sci U S A* (2013) 110:2993–8. doi:10.1073/pnas.1213737110
111. Wine Y, Horton AP, Ippolito GC, Georgiou G. Serology in the 21st century: the molecular-level analysis of the serum antibody repertoire. *Curr Opin Immunol* (2015) 35:89–97. doi:10.1016/j.coi.2015.06.009
112. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* (2014) 111:2259–64. doi:10.1073/pnas.1317793111
113. Iversen R, Snir O, Stensland M, Kroll JE, Steinsbø Ø, Korponay-Szabó IR, et al.. Strong clonal relatedness between serum and gut IgA despite different plasma cell origins. *Cell Rep* (2017) 20:2357–67. doi:10.1016/j.celrep.2017.08.036
114. Chen J, Zheng Q, Hammers CM, Ellebrecht CT, Mukherjee EM, Tang H-Y, et al. Proteomic analysis of pemphigus autoantibodies indicates a larger, more diverse, and more dynamic repertoire than determined by B cell genetics. *Cell Rep* (2017) 18:237–47. doi:10.1016/j.celrep.2016.12.013
115. VanDuijn MM, Dekker LJ, Van IJcken JWF, Sillevis Smitt PAE, Luider TM. Immune repertoire after immunization as seen by next-generation sequencing and proteomics. *Front Immunol* (2017) 8:1286. doi:10.3389/fimmu.2017.01286
116. Berger CT, Greiff V, Mehling M, Fritz S, Meier MA, Hoenger G, et al. Influenza vaccine response profiles are affected by vaccine preparation and preexisting immunity, but not HIV infection. *Hum Vaccin Immunother* (2015) 11:391–6. doi:10.1080/21645515.2015.1008930
117. Wang C, Liu Y, Cavanagh MM, Saux SL, Qi Q, Roskin KM, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc Natl Acad Sci U S A* (2015) 112:500–5. doi:10.1073/pnas.1415875112
118. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc Natl Acad Sci U S A* (2013) 110:6470–5. doi:10.1073/pnas.1219320110
119. Hoehn KB, Fowler A, Lunter G, Pybus OG. The diversity and molecular evolution of B-cell receptors during infection. *Mol Biol Evol* (2016) 33:1147–57. doi:10.1093/molbev/msw015
120. Schramm CA, Sheng Z, Zhang Z, Mascola JR, Kwong PD, Shapiro L. SONAR: a high-throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of B cell transcripts. *B Cell Biol* (2016) 7:372. doi:10.3389/fimmu.2016.00372
121. Yaari G, Heiden JV, Uduman M, Gadala-Maria D, Gupta N, Stern JN, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput Immunoglobulin sequencing data. *Front B Cell Biol* (2013) 4:358. doi:10.3389/fimmu.2013.00358
122. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree©: creating immunoglobulin variable region gene lineage trees. *J Immunol Methods* (2008) 338:67–74. doi:10.1016/j.jim.2008.06.006
123. Andrews SF, Kaur K, Pauli NT, Huang M, Huang Y, Wilson PC. High preexisting serological antibody levels correlate with diversification of the influenza vaccine response. *J Virol* (2015) 89(6):3308–17. doi:10.1128/JVI.02871-14
124. Wu X, Zhang Z, Schramm CA, Joyce MG, Do Kwon Y, Zhou T, et al. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* (2015) 161:470–85. doi:10.1016/j.cell.2015.03.004
125. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* (2012) 13:303–14. doi:10.1038/nrg3186
126. Schliep KP. Phangorn: phylogenetic analysis in R. *Bioinformatics* (2011) 27:592–3. doi:10.1093/bioinformatics/btq706
127. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* (2004) 20:289–90. doi:10.1093/bioinformatics/btg412
128. Boc A, Diallo AB, Makarenkov V. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res* (2012) 40:W573–9. doi:10.1093/nar/gks485
129. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. *Bioinformatics* (2007) 23:2947–8. doi:10.1093/bioinformatics/btm404
130. Swofford D, Begle DP. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1, March 1993*. Illinois: Center for Biodiversity, Natural History Survey (1993).
131. Giribet G. TNT: Tree analysis using New Technology. *Syst Biol* (2005) 54:176–8. doi:10.1080/10635150590905830
132. Felsenstein J. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* (1989) 5:164–6. doi:10.1111/j.1096-0031.1989.tb00562.x
133. Revell LJ, Chamberlain SA. Rphylip: an R interface for PHYLIP. *Methods Ecol Evol* (2014) 5:976–81. doi:10.1111/2041-210X.12233
134. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* (2014) 30:1312–3. doi:10.1093/bioinformatics/btu033

135. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* (2005) 33:W557–9. doi:10.1093/nar/gki352
136. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* (2016) 33:1870–4. doi:10.1093/molbev/msw054
137. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinforma Oxf Engl* (2003) 19:1572–4. doi:10.1093/bioinformatics/btg180
138. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* (2014) 10:e1003537. doi:10.1371/journal.pcbi.1003537
139. Yermanos A, Greiff V, Krautler NJ, Menzel U, Dounas A, Miho E, et al. Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics* (2017) 33(24):3938–46. doi:10.1093/bioinformatics/btx533
140. Vieira MC, Zinder D, Cobey S. Selection and neutral mutations drive pervasive mutability losses in long-lived B cell lineages. *bioRxiv* (2017). doi:10.1101/163741
141. Pinheiro A, de Mera IG, Alves PC, Gortázar C, de la Fuente J, Esteves PJ. Sequencing of modern lepus VDJ genes shows that the usage of VHn genes has been retained in both oryctolagus and lepus that diverged 12 million years ago. *Immunogenetics* (2013) 65:777–84. doi:10.1007/s00251-013-0728-3
142. Kendall M, Colijn C. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol Biol Evol* (2016) 33:2735–43. doi:10.1093/molbev/msw124
143. Yeap L-S, Hwang JK, Du Z, Meyers RM, Meng F-L, Jakubauskaitė A, et al. Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* (2015) 163:1124–37. doi:10.1016/j.cell.2015.10.042
144. Betz AG, Rada C, Pannell R, Milstein C, Neuberger MS. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proc Natl Acad Sci U S A* (1993) 90:2385–8. doi:10.1073/pnas.90.6.2385
145. Cui A, Niro RD, Heiden JAV, Briggs AW, Adams K, Gilbert T, et al. A Model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J Immunol* (2016) 197(9):3566–74. doi:10.4049/jimmunol.1502263
146. Sheng Z, Schramm CA, Kong R, NISC Comparative Sequencing Program, Mullikin JC, Mascola JR, et al. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front Immunol* (2017) 8:537. doi:10.3389/fimmu.2017.00537
147. Sheng Z, Schramm CA, Connors M, Morris L, Mascola JR, Kwong PD, et al. Effects of darwinian selection and mutability on rate of broadly neutralizing antibody evolution during HIV-1 infection. *PLoS Comput Biol* (2016) 12:e1004940. doi:10.1371/journal.pcbi.1004940
148. Hoehn KB, Lunter G, Pybus OG. A phylogenetic codon substitution model for antibody lineages. *Genetics* (2017) 206:417–27. doi:10.1534/genetics.116.196303
149. DeWitt WS III, Mesin L, Victora GD, Minin VN, Matsen FA IV. Using genotype abundance to improve phylogenetic inference. *Q-Bio* (2017). Available from: http://arxiv.org/abs/1708.08944
150. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* (2005) 71:8228–35. doi:10.1128/AEM.71.12.8228-8235.2005
151. de Bourcy CFA, Angel CJL, Vollmers C, Dekker CL, Davis MM, Quake SR. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc Natl Acad Sci U S A* (2017) 114(5):1105–10. doi:10.1073/pnas.1617959114
152. Lewitus E, Morlon H. Characterizing and comparing phylogenies from their laplacian spectrum. *Syst Biol* (2016) 65:495–507. doi:10.1093/sysbio/syv116
153. Horns F, Vollmers C, Dekker CL, Quake SR. Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. *bioRxiv* (2017). doi:10.1101/145052
154. Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* (2017) 49:659–65. doi:10.1038/ng.3822
155. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* (2017) 547:89–93. doi:10.1038/nature22383
156. Galson JD, Trück J, Fowler A, Clutterbuck EA, Münz M, Cerundolo V, et al. Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine* (2015) 2(12):2070–9. doi:10.1016/j.ebiom.2015.11.034
157. Chen H. *VennDiagram: Generate High-Resolution Venn and Euler Plots*. (2016). Available from: https://CRAN.R-project.org/package=VennDiagram.
158. Morisita M. Measuring of the dispersion of individuals and analysis of the distributional patterns. *Mem Fac Sci Kyushu Univ Ser E* (1959) 2:5–23.
159. Dziubianau M, Hecht J, Kuchenbecker L, Sattler A, Stervbo U, Rödelsperger C, et al. TCR Repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell–related pathology. *Am J Transplant* (2013) 13:2842–54. doi:10.1111/ajt.12431
160. Rempala GA, Seweryn M, Ignatowicz L. Model for comparative analysis of antigen receptor repertoires. *J Theor Biol* (2011) 269:1–15. doi:10.1016/j.jtbi.2010.10.001
161. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. The repertoire dissimilarity index as a method to compare lymphocyte receptor repertoires. *BMC Bioinformatics* (2017) 18:155. doi:10.1186/s12859-017-1556-5
162. Yokota R, Kaminaga Y, Kobayashi TJ. Quantification of inter-sample differences in T-cell receptor repertoires using sequence-based information. *Front Immunol* (2017) 8:1500. doi:10.3389/fimmu.2017.01500
163. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* (2014) 15:550. doi:10.1186/s13059-014-0550-8
164. Bürckert J-P, Dubois ARSX, Faison WJ, Farinelle S, Charpentier E, Sinner R, et al. Functionally convergent B cell receptor sequences in transgenic rats expressing a Human B cell repertoire in response to tetanus toxoid and measles antigens. *Front Immunol* (2017) 8:1834. doi:10.3389/fimmu.2017.01834
165. Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E, et al. Specificity, privacy, and degeneracy in the CD4 T cell receptor repertoire following immunization. *Front Immunol* (2017) 8:430. doi:10.3389/fimmu.2017.00430
166. Sun Cinelli M, Best K, Heather JM, Reich-Zeliger S, Shifrut E, Friedman N, et al. Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics* (2017) 33:951–5. doi:10.1093/bioinformatics/btw771
167. Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A* (2005) 102:6395–400. doi:10.1073/pnas.0408677102
168. Thomas N, Best K, Cinelli M, Reich-Zeliger S, Gal H, Shifrut E, et al. Tracking global changes induced in the CD4 T cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinforma Oxf Engl* (2014) 30(22):3181–8. doi:10.1093/bioinformatics/btu523
169. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem* (1985) 4:23–55. doi:10.1007/BF01025492
170. Konishi H, Komura D, Katoh H, Atsumi S, Koda H, Yamamoto A, et al. Capturing the difference in humoral immunity between normal and tumor environments from RNA sequences of B-cell receptors using supervised machine learning. *bioRxiv* (2017):187120.
171. Ostmeyer J, Christley S, Rounds WH, Toby I, Greenberg BM, Monson NL, et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics* (2017) 18:401. doi:10.1186/s12859-017-1814-6
172. Apeltsin L, Wang S, Büdingen H-C, Sirota M. A haystack heuristic for autoimmune disease biomarker discovery using next-gen immune repertoire sequencing data. *Sci Rep* (2017) 7:5338. doi:10.1038/s41598-017-04439-5
173. Torkamani A, Andersen KG, Steinhubl SR, Topol EJ. High-definition medicine. *Cell* (2017) 170:828–43. doi:10.1016/j.cell.2017.08.007
174. Boyd SD, Crowe JE Jr. Deep sequencing and human antibody repertoire analysis. *Curr Opin Immunol* (2016) 40:103–9. doi:10.1016/j.coi.2016.03.008
175. Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief Bioinform* (2017):bbw138. doi:10.1093/bib/bbw138

176. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* (2017) 547:94–8. doi:10.1038/nature22976
177. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* (2018) 46(D1):D419–27. doi:10.1093/nar/gkx760
178. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* (2017) 33(18):2924–9. doi:10.1093/bioinformatics/btx286
179. Martin ACR. Protein sequence and structure analysis of antibody variable domains. In: Kontermann R, Dübel S, editors. *Antibody Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg (2016). p. 33–51. Available from: http://link.springer.com/10.1007/978-3-642-01147-4_3
180. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The Immune Epitope Database (IEDB) 3.0. *Nucleic Acids Res* (2015) 43:D405–12. doi:10.1093/nar/gku938
181. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* (2017). doi:10.1101/142760
182. Jurtz VI, Rosenberg Johansen A, Nielsen M, Armenteros A, Juan J, Nielsen H, et al. An introduction to deep learning on biological sequence data – examples and solutions. *Bioinformatics* (2017) 33(22):3685–90. doi:10.1093/bioinformatics/btx531
183. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat Med* (2016) 22:1456–64. doi:10.1038/nm.4224
184. Snir O, Chen X, Gidoni M, du Pré MF, Zhao Y, Steinsbø Ø, et al. Stereotyped antibody responses target posttranslationally modified gluten in celiac disease. *JCI Insight* (2017) 2:93961. doi:10.1172/jci.insight.93961
185. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. *Mol Pharm* (2016) 13:1445–54. doi:10.1021/acs.molpharmaceut.5b00982
186. Greiff V, Redestig H, Luck J, Bruni N, Valai A, Hartmann S, et al. A minimal model of peptide binding predicts ensemble properties of serum antibodies. *BMC Genomics* (2012) 13:79. doi:10.1186/1471-2164-13-79
187. Becattini S, Latorre D, Mele F, Foglierini M, Gregorio CD, Cassotta A, et al. Functional heterogeneity of human memory CD4+ T cell clones primed by pathogens or vaccines. *Science* (2014) 347:400–6. doi:10.1126/science.1260668
188. Kaplinsky J, Li A, Sun A, Coffre M, Koralov SB, Arnaout R. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc Natl Acad Sci U S A* (2014) 111:E2622–9. doi:10.1073/pnas.1403278111
189. Ghraichy M, Galson JD, Kelly DF, Trück J. B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. *Immunology* (2018) 153(2):145–60. doi:10.1111/imm.12865
190. Khavrutskii IV, Chaudhury S, Stronsky SM, Lee DW, Benko JG, Wallqvist A, et al. Quantitative analysis of repertoire-scale immunoglobulin properties in vaccine-induced B-cell responses. *Front Immunol* (2017) 8:910. doi:10.3389/fimmu.2017.00910
191. Galson JD, Trück J, Clutterbuck EA, Fowler A, Cerundolo V, Pollard AJ, et al. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med* (2016) 8:68. doi:10.1186/s13073-016-0322-z
192. Ellebedy AH, Jackson KJL, Kissick HT, Nakaya HI, Davis CW, Roskin KM, et al. Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat Immunol* (2016) 17:1226–34. doi:10.1038/ni.3533
193. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee J-Y, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) 13:691–700. doi:10.1016/j.chom.2013.05.008
194. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, Pevzner PA. Reconstructing antibody repertoires from error-prone immunosequencing reads. *J Immunol* (2017) 199(9):3369–80. doi:10.4049/jimmunol.1700485
195. Safonova Y, Lapidus A, Lill J. IgSimulator: a versatile immunosequencing simulator. *Bioinformatics* (2015) 31(19):3213–5. doi:10.1093/bioinformatics/btv326
196. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol* (2016) 35(3):203–14. doi:10.1016/j.tibtech.2016.09.010
197. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* (2015) 12:380–1. doi:10.1038/nmeth.3364
198. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *J Immunol* (2017) 198(6):2489–99. doi:10.4049/jimmunol.1601850
199. Brown SD, Raeburn LA, Holt RA. Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Med* (2015) 7:125. doi:10.1186/s13073-015-0248-x
200. Rizzetto S, Koppstein DN, Samir J, Singh M, Reed JH, Cai CH, et al. B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *bioRxiv* (2017):181156. doi:10.1101/181156
201. Mangul S, Mandric I, Yang HT, Strauli N, Montoya D, Rotman J, et al. Profiling adaptive immune repertoires across multiple human tissues by RNA sequencing. *bioRxiv* (2016):089235. doi:10.1101/089235
202. Lindeman I, Emerton G, Sollid LM, Teichmann S, Stubbington MJT. BraCeR: Reconstruction of B-cell receptor sequences and clonality inference from single-cell RNA-sequencing. *bioRxiv* (2017):185504. doi:10.1101/185504
203. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat Methods* (2016) 13:329–32. doi:10.1038/nmeth.3800
204. Li B, Li T, Pignon J-C, Wang B, Wang J, Shukla SA, et al. Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat Genet* (2016) 48:725–32. doi:10.1038/ng.3581
205. Geering B, Fussenegger M. Synthetic immunology: modulating the human immune system. *Trends Biotechnol* (2015) 33:65–79. doi:10.1016/j.tibtech.2014.10.006
206. Roybal KT, Lim WA. Synthetic immunology: hacking immune cells to expand their therapeutic capabilities. *Annu Rev Immunol* (2017) 35:229–53. doi:10.1146/annurev-immunol-051116-052302
207. Jiang N. Immune engineering: from systems immunology to engineering immunity. *Curr Opin Biomed Eng* (2017) 1:54–62. doi:10.1016/j.cobme.2017.03.002
208. Liu XS, Mardis ER. Applications of immunogenomics to cancer. *Cell* (2017) 168:600–12. doi:10.1016/j.cell.2017.01.014
209. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, et al. By-passing in vitro screening—next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* (2010) 38:e193. doi:10.1093/nar/gkq789
210. Parola C, Neumeier D, Reddy ST. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. *Immunology* (2018) 153(1):31–41. doi:10.1111/imm.12838

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. EM is the founder of aiNET GmbH.

Copyright © 2018 Miho, Yermanos, Weber, Berger, Reddy and Greiff. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 01 March 2018
doi: 10.3389/fimmu.2018.00329

Coupling of Single Molecule, Long Read Sequencing with IMGT/HighV-QUEST Analysis Expedites Identification of SIV gp140-Specific Antibodies from scFv Phage Display Libraries

Seung Yub Han[1], Alesia Antoine[2], David Howard[1], Bryant Chang[1], Woo Sung Chang[1], Matthew Slein[1], Gintaras Deikus[2], Sofia Kossida[3], Patrice Duroux[3], Marie-Paule Lefranc[3], Robert P. Sebra[2], Melissa L. Smith[2] and Ismael Ben F. Fofana[1]**

[1] Biology Department, Boston College, Chestnut Hill, MA, United States, [2] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Icahn Institute of Genomics and Multiscale Biology, New York, NY, United States, [3] The international ImMunoGeneTics information system® (IMGT®), Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), UMR CNRS, Montpellier University, Montpellier, France

OPEN ACCESS

Edited by:
Prabakaran Ponraj,
Intrexon,
United States

Reviewed by:
Paolo Casali,
The University of Texas Health Science Center San Antonio, United States
Jiang Zhu,
The Scripps Research Institute, United States

***Correspondence:**
Melissa L. Smith
melissa.smith@mssm.edu;
Ismael Ben F. Fofana
ismael.fofana@bc.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: 22 September 2017
Accepted: 06 February 2018
Published: 01 March 2018

Citation:
Han SY, Antoine A, Howard D, Chang B, Chang WS, Slein M, Deikus G, Kossida S, Duroux P, Lefranc M-P, Sebra RP, Smith ML and Fofana IBF (2018) Coupling of Single Molecule, Long Read Sequencing with IMGT/HighV-QUEST Analysis Expedites Identification of SIV gp140-Specific Antibodies from scFv Phage Display Libraries. Front. Immunol. 9:329. doi: 10.3389/fimmu.2018.00329

The simian immunodeficiency virus (SIV)/macaque model of human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome pathogenesis is critical for furthering our understanding of the role of antibody responses in the prevention of HIV infection, and will only increase in importance as macaque immunoglobulin (IG) gene databases are expanded. We have previously reported the construction of a phage display library from a SIV-infected rhesus macaque (*Macaca mulatta*) using oligonucleotide primers based on human IG gene sequences. Our previous screening relied on Sanger sequencing, which was inefficient and generated only a few dozen sequences. Here, we re-analyzed this library using single molecule, real-time (SMRT) sequencing on the Pacific Biosciences (PacBio) platform to generate thousands of highly accurate circular consensus sequencing (CCS) reads corresponding to full length single chain fragment variable. CCS data were then analyzed through the international ImMunoGeneTics information system® (IMGT®)/HighV-QUEST (www.imgt.org) to identify variable genes and perform statistical analyses. Overall the library was very diverse, with 2,569 different IMGT clonotypes called for the 5,238 IGHV sequences assigned to an IMGT clonotype. Within the library, SIV-specific antibodies represented a relatively limited number of clones, with only 135 different IMGT clonotypes called from 4,594 IGHV-assigned sequences. Our data did confirm that the IGHV4 and IGHV3 gene usage was the most abundant within the rhesus antibodies screened, and that these genes were even more enriched among SIV gp140-specific antibodies. Although a broad range of VH CDR3 amino acid (AA) lengths was observed in the unpanned library, the vast majority of SIV gp140-specific antibodies demonstrated a more uniform VH CDR3 length (20 AA). This uniformity was far less apparent when VH CDR3 were classified according to their clonotype (range: 9–25 AA), which we believe is more relevant for specific antibody identification. Only 174 IGKV and 588 IGLV clonotypes were identified within the VL sequences associated with SIV gp140-specific VH. Together, these data strongly suggest that the combination of

SMRT sequencing with the IMGT/HighV-QUEST querying tool will facilitate and expedite our understanding of polyclonal antibody responses during SIV infection and may serve to rapidly expand the known scope of macaque V genes utilized during these responses.

Keywords: antibody, simian immunodeficiency virus, rhesus macaque, PacBio sequencing, single chain fragment variable library, phage display, International ImMunoGeneTics information system/HighV-QUEST

INTRODUCTION

Nonhuman primates are an important animal model for numerous human diseases, as there is great similarity between the human and macaque genomes (1–7). In addition, macaque immunoglobulin (IG) genes are likely those most closely related to human IG genes among available human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) animal models (8–11). As a result, the variable heavy (VH) and variable light (VL) domains of macaque antibody heavy (H) and light (L) chains can be generated using polymerase chain reaction (PCR) conditions and oligonucleotide primers based on human IG nucleotide sequences. This has been shown thus far for VH rather than VL genes, although the use of rhesus specific primers for amplification may function for both variable domains (9, 11, 12).

The simian immunodeficiency virus (SIV)/macaque model of AIDS has been extensively studied and provides the most accurate reflection of HIV pathogenesis across all available animal models (5, 13–16). Virus-specific antibodies are abundantly produced during the course of HIV/SIV infection in humans or macaques (5, 17–20). These antibodies, which primarily target the envelope glycoprotein (Env) on the surface of HIV/SIV virions, generally do not provide protection due in part to their inability to neutralize the virus. In fact, the Env surface glycoprotein employs multiple strategies to shield its neutralization-sensitive epitopes, such as the CD4 and the coreceptor-binding sites, as well as the fusion peptide on the Membrane Proximal External Region (MPER). Env trimer oligomerization, the presence of hypervariable loops and extensive glycosylation are all components of a complex escape mechanism that limits the potential potency of antibody-mediated neutralization (5, 21).

Despite this, many potent HIV neutralizing antibodies have been isolated and characterized (22–26). Unique structural features of these antibodies, including extensive somatic mutations and unusually long VH domain complementarity determining region three (VH CDR3), have been associated with the development of potent neutralizing activity (19). However, roles for the VL domains and other complementarity determining regions in conferring potent HIV neutralizing activity have not been excluded (27–29).

Despite the extensive knowledge gained from studying this panel of naturally occurring neutralizing antibodies, stimulating the development of broadly neutralizing antibodies (bnAbs) by vaccination has remained a critical roadblock in the development of an efficacious HIV vaccine. The failures of past vaccine immunogens to do so have been, in part, attributed to the inability of the various vaccine antigens to engage B cells that express the proper germline receptors (30–33). However, progress in this area is being made. For example, very recently anti-HIV bnAbs were elicited in the cow (*Bos taurus*), an unusual experimental animal for HIV-related research (34). The authors were motivated to take this approach due to the inherently long VH CDR3, which characterize cow antibodies. The rarity of such long VH CDR3 in the human IG repertoire raises the question of whether induction of such long VH CDR3 bnAbs will ever be achieved in humans *via* vaccination. In light of these questions and challenges, many have also turned their focus to the possibility of eliciting protective non-neutralizing antibodies (nnAbs), which may also be capable of preventing HIV infection. Interestingly, the most successful clinical HIV vaccine trial to date, RV144, found a correlation with limited protection and non-neutralizing Env-binding antibodies that target the V1-V2 loop (35–38). Similarly, the most successful pre-clinical HIV vaccine, which uses live-attenuated strains of SIVmac239 found only low titers of neutralizing antibodies in protected macaques; suggesting instead a role for nnAbs in this model (39). Unlike nAb activity, which appears to be conferred by a single or few dominant, protective monoclonal antibodies (mAbs) in a given individual; the characterization of protective nnAbs will likely require large-scale analysis of polyclonal antibodies.

We have previously reported the construction of a phage display library from a SIV-infected rhesus macaque (11). This library was generated by PCR amplification of the VH and VL chains using primers corresponding to the human IG gene sequences. Our prior screening of mAbs from this library relied on handpicking bacterial clones after biopanning with SIV Env gp140, followed by Sanger sequencing. This approach was inefficient and generated only a few dozen sequences for analysis, likely severely underrepresenting the repertoire present. The screening of this library would be greatly improved with the use of next generation sequencing (NGS) technologies, such as Illumina, however often NGS platforms are limited in the length of the reads (≤400 bp), covering at most one VH or VL domain per read. This limitation has been addressed through the use of alternate high throughput NGS platforms such as the Ion Torrent Personal Genome Machine (PGM) S5 and the Pacific Bioscience (PacBio) RSII and Sequel systems (40, 41). Using the PGM-S5 system, He et al. generated 900 bp sequencing reads to identify precursors and lineage intermediates of HIV-1 bnAbs from a phage display library. Although single chain fragment variables (scFvs) from this combinatorial library did not necessarily represent authentic VH–VL pairing, the authors were

Abbreviations: SIV, simian immunodeficiency virus; Env, envelope; scFv, single chain fragment variable; IMGT®, the international ImMunoGeneTics information system; bp, base pair; IG, immunoglobulin; CDR, complementarity determining region; AA, amino acid.

able to validate their data using biopanning onto a native-like gp140 trimer and comparison with previously characterized bnAb lineages (42–49). Using the PacBio RSII system, Hemadou et al. generated long reads (>800 bp) covering full length scFvs following *in vivo* panning in an animal model of atherosclerosis. Subsequently, they analyzed the sequencing data using International ImMunoGeneTics information system (IMGT)/HighV-QUEST combined with a novel scFv functionality tool for simultaneous characterization of VH and VL chains from individual scFvs. Here, using our previously characterized phage display library (11), we tested the validity of PacBio sequencing and IMGT/HighV-QUEST analysis (www.imgt.org) combined with scFv functionality for the identification and characterization of SIV-specific antibodies.

MATERIALS AND METHODS

Phage Display Library Preparation

Construction of a scFv phage display library using archived spleen biopsies from a SIV-infected rhesus macaque (Mm333-95) has been previously reported (11, 50). The animal was housed at the New England Primate Research Center of Harvard Medical School, and given care in accordance with standards of the Association for Assessment and Accreditation of Laboratory Animal Care and the Harvard Medical School Animal Care and Use Committee. The study was approved by the Harvard Medical Area Standing Committee on Animals, within the Office for Research Subject Protection at Harvard Medical School, and conducted according to the principles described in the *Guide for the Care and Use of Laboratory Animals* (51).

In the current study, we aimed to evaluate our sequencing and analysis pipelines for the characterization of SIV-specific antibodies. For that reason, the previously generated library was used (11) as it allowed for the most direct comparison of the new pipelines with the prior gold standard method of handpicking colonies representing selected SIV-specific scFvs and screening by Sanger sequencing.

In the first round, antibody variable domains, VH and VL, were amplified by PCR using oligonucleotide primers corresponding to the human IG sequences. In the second round, VH and VL products were linked together using external primers corresponding to the 5′ (RSC-F: gaggaggaggaggaggaggcggggc-ccaggcggccgagctc) and 3′ (RSC-B: gaggaggaggaggaggagcctggc-cggcctggccactagtg) regions of the VL and VH PCR products, respectively. This fusion was facilitated by the addition of a linker sequence to the internal PCR primers corresponding to 3′ and 5′ regions in VL and VH PCR products, respectively. The resultant scFv (VL-linker-VH) products were cloned into the phagemid vector pComb3xSS. XL1 Blue *Escherichia coli* were transformed with recombinant pComb3xSS-scFv DNA by electroporation using a Gene Pulser Xcell (Bio-Rad, Hercules, CA, USA). The phage library preparation was obtained after amplification by culturing in the presence of VCSM13 helper phage (Agilent, Santa Clara, CA, USA). Biopanning of the library using immobilized SIV Env gp140 was also previously described (11).

scFv DNA Preparation from Library and Sub-Library

Total DNA was extracted from the bacterial pellets obtained during preparation of the unpanned library and the fourth round SIV Env gp140-panned library using the PureLink DNA maxi-preparation kit (Invitrogen, Carlsbad, CA, USA). scFv DNA was then PCR amplified from the extracted DNA using the same external primers (RSC-F and RSC-B) that had been selected for initial library construction. Samples were amplified in triplicate, with three different concentrations (25, 50, and 75 ng) of DNA as initial template input. High-fidelity Platinum SuperFi polymerase (Life Technologies, Carlsbad, CA, USA) was used to prepare the reaction mixture as follows: 1 µl template DNA (25, 50, or 75 ng), 2 µl (200 pmol) 5′ Primer (RSC-F), 2 µl (200 pmol) 3′ Primer (RSC-B), and 45 µl Platinum SuperFi Master mix. PCR reactions were performed under the following conditions: heated to 94°C for 5 min, subjected to 15 cycles of: 94°C for 15 s, 56°C for 15 s, 72°C for 2 min, followed by a 10-min extension at 72°C. Five microliters of each reaction were evaluated for successful amplification on a 1% agarose gel. The triplicate reactions were then pooled together before performing a PCR cleanup using the QIAquick PCR purification Kit (Qiagen, Valencia, CA, USA).

SMRTbell Library Preparation and Sequencing

PacBio SMRTbell library preparation and sequencing has previously been described (52). Briefly, SMRTbell libraries were prepared following the manufacturer's protocol and using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA). A total of 250 ng of AMPure PB bead-purified scFv amplicon was added directly to the DNA damage repair step of the Amplicon Template Preparation and Sequencing protocol (http://www.pacb.com/wp-content/uploads/2015/09/Unsupported-Amplicon-Template-Preparation-Sequencing.pdf). Following construction, SMRTbell library quality and quantity were assessed using both the Agilent 12000 DNA Kit and the 2100 Bioanalyzer System (Santa Clara, CA, USA), as well as the Qubit dsDNA High Sensitivity Assay kit and Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Sequencing primer annealing and P6 polymerase binding were performed using the recommended 20:1 primer:template ratio and 10:1 polymerase:template ratio, respectively. scFv SMRTbell libraries were loaded onto SMRT cells at a concentration of 50 pM. SMRT sequencing was performed on the PacBio RS II system using the C4 sequencing kit with magnetic bead loading and 6-h movies. Circular consensus sequencing (CCS) reads were generated using Arrow and the CCS2 protocol as a part of SMRTLink version 4.0; CCS reads were filtered by both quality (Q30, 99.9% accuracy) and size, retaining only sequences that ranged from 700 to 1,200 bp based on expected scFv size distribution. Resultant FASTA files were used for downstream analyses. These filtered CCS reads have been submitted to the Sequence Read Archive under submission SUB3223332 entitled "*Macaca mulatta*-derived SIV-gp140-specific scFv circular consensus sequences" and can be retrieved using accession number SRP125114.

IMGT/HighV-QUEST Analysis

International ImMunoGeneTics information system/HighV-QUEST analysis was performed *via* the IMGT web portal (53–57). The CCS FASTA files were analyzed using IMGT/HighV-QUEST program version 1.5.5 with the advanced scFb functionality (57). Resultant data files obtained from IMGT/HighV-QUEST were further analyzed using the statistical and clonotype analysis tool that uses the IMGT/V-QUEST version 3.4.7 with advanced scFv functionality (57). An IMGT clonotype (AA) (55) is defined as a unique V-(D)-J rearrangement (with the IMGT gene and allele names determined by IMGT/HighV-QUEST at the nucleotide level) and a unique CDR3-IMGT AA in-frame junction sequence (C104, W118 for IGHV and C104, F118 for IGKV and IGLV) (see IMGT/HighV-QUEST documentation). Data filtering was applied with the following criteria to be fulfilled for each of the two V domains: (i) >85% of identity of the V-REGION of the V domain with the V-REGION of the closest germline IMGT gene and allele (58) and (ii) in-frame V-(D)-J junction. Filtered sequences were then analyzed to identify the closest V, D (for VH) and J IMGT genes and alleles, in order to characterize the amino acid (AA) junction and to give a complete description of the scFv with IMGT labels, using the IMGT/V-QUEST algorithm for scFv, implemented in IMGT/HighV-QUEST. Here, we report the statistical and IMGT clonotype analysis for each domain for the unpanned library (Pan0) and the fourth round SIV Env gp140 panned library (Pan4). In order to ensure that the breadth and depth of the data generated by a single SMRT cell per sample was not limiting the characterization of the complex scFv pools, a comparative analysis was done pooling data generated from three SMRT cells run for a single sample.

scFv Recovery from the Sub-Library

Two clones were selected from the fourth round of gp140-panning based on extent of VH CDR3 representation among the sequences generated by IMGT analysis. In addition, these clones also contained a new VH CDR3 AA sequence compared to previously identified clones (11). Recovery of scFv from libraries has previously been described (59). Briefly, overlapping primers were design within the VH CDR3. For clone selection one (S1), the 5′ primer (S1-F: GCG AGA GGC TCC AAA CAA TTT TGT AGT) and 3′ primer (S1-R: ACT ACA AAA TTG TTT GGA GCC TCT CGC) were used while for clone selection two (S2), the 5′ primer (S2-F: CCT CTC CCC GAC TGG GCT GAT TAT AAG) and 3′ primer (S2-R: CTT ATA ATC AGC CCA GTC GGG GAG AGG) were selected. PCR was conducted as described above, using the Platinum HiFi Mastermix (Life Technologies, Carlsbad, CA, USA). 1 µl of DpnI digested product was used to transform TOP10 F′ *E. coli* (Thermo Fisher Scientific, Waltham, MA, USA) and the subsequently transformed bacteria were plated on LB/Agar medium supplemented with 50 µg/ml of Carbenicillin (AmericanBio, Natick, MA, USA). Positive clones were verified by Sanger sequencing. The recovered sequence was also analyzed using IMGT/V-QUEST(www.imgt.org) with scFv functionality, as described above.

Following this selective cloning, binding specificity to SIV Env gp140 of these clones were confirmed after induction of the bacterial culture as previously described (8, 11, 20). Briefly, bacterial colonies were cultured with SB medium supplemented with 50 µg/ml of Carbenicillin (AmericanBio, Natick, MA, USA). After 5–8 h of culture at 37°C and 250 rpm, scFv expression was induced by addition of Isopropyl β-D-1-thiogalactopyranoside (IPTG) at 2 mM. The induction was performed at 37°C overnight. Culture supernatant was clarified by centrifugation at 12,000 rpm in a microcentrifuge for 5 min. 50 µl of supernatant was tested by enzyme-linked immunosorbent assay (ELISA) for binding to SIV Env gp120 and/or gp140. ELISA plates were also coated with BSA as a negative control antigen.

RESULTS

SMRT Sequencing

We have previously described the construction of a phage display library derived from a rhesus macaque infected with SIV. This work generated 32 unique scFv sequences after four rounds of panning onto SIV Env gp140, all of which were characterized to target the gp41 region of Env (11). Here, we aimed to perform a large-scale deep sequencing-driven analysis of the same scFv library and compare the diversity of the unpanned library (Pan0) to the library after round 4 of gp140 selection (Pan4). DNA extracted from bacterial pellets obtained after Pan0 and Pan4 culturing were used as templates to amplify the collection of scFv sequences under examination (**Figure 1**).

As expected, we obtained DNA amplicons of ~800 bp representing the variable domains VH (~ 400 bp), VL (~350 bp) and the short (21 bp) or long linker (54 bp) (**Figure 1**). We also obtained additional and less pronounced products between 800 and 1,700 bp (**Figures 1A,B**) with the 1,700 bp fragment corresponding to scFv dimers. Furthermore, peaks at 50 and 17,000 bp were also observed and corresponded to the lower and upper ladder of the Agilent 12000 DNA Kit, respectively (**Figure 1B**). In order to focus our analyses on highly accurate, full-length scFv fragments, SMRT sequencing data were processed using the latest CCS2 algorithm and filtered by size (700–1,200 bp) and accuracy (99.9%, QV30) (**Figure 1C**). **Table 1** summarizes the total number of sequences and IMGT clonotype-respective assignments of the VH and VL domains to the IGHV, IGKV, and IGLV genes observed. The total number of sequences assigned to an IMGT clonotype was independent of PCR template DNA concentration and similar between Pan0 (3,837, 3,159, and 5,238) and Pan4 (4,594, 4,594, and 3,884). As would be expected following antigen-specific selection, the analysis of IGHV sequences resulted in many more distinct clonotypes in Pan0 (1,887, 1,685, and 2,569) than Pan4 (135, 141, and 127). The number of observed out-of-frame sequences was similar between Pan0 (117, 90, and 115) and Pan4 (106, 115, and 80). The number of "other category" sequences was far higher for Pan0 (112, 76, and 103) than Pan4 (1, 0, and 0) suggesting that the Pan4 products contain more productive sequences than were present in the unpanned library. A total of 4,640 (Pan0) and 6,957 (Pan4) IGKV sequences and a total of 4,068 (Pan0) and 7,484 (Pan4) IGLV sequences were assigned to an IMGT clonotype. A more diverse pool of clonotypes was observed with Pan 0, consisting of 1,651 IGKV and 1,573 IGLV variants. The Pan4 sublibrary was far less diverse with only 147 and 588 different clonotypes for IGKV and IGLV, respectively.

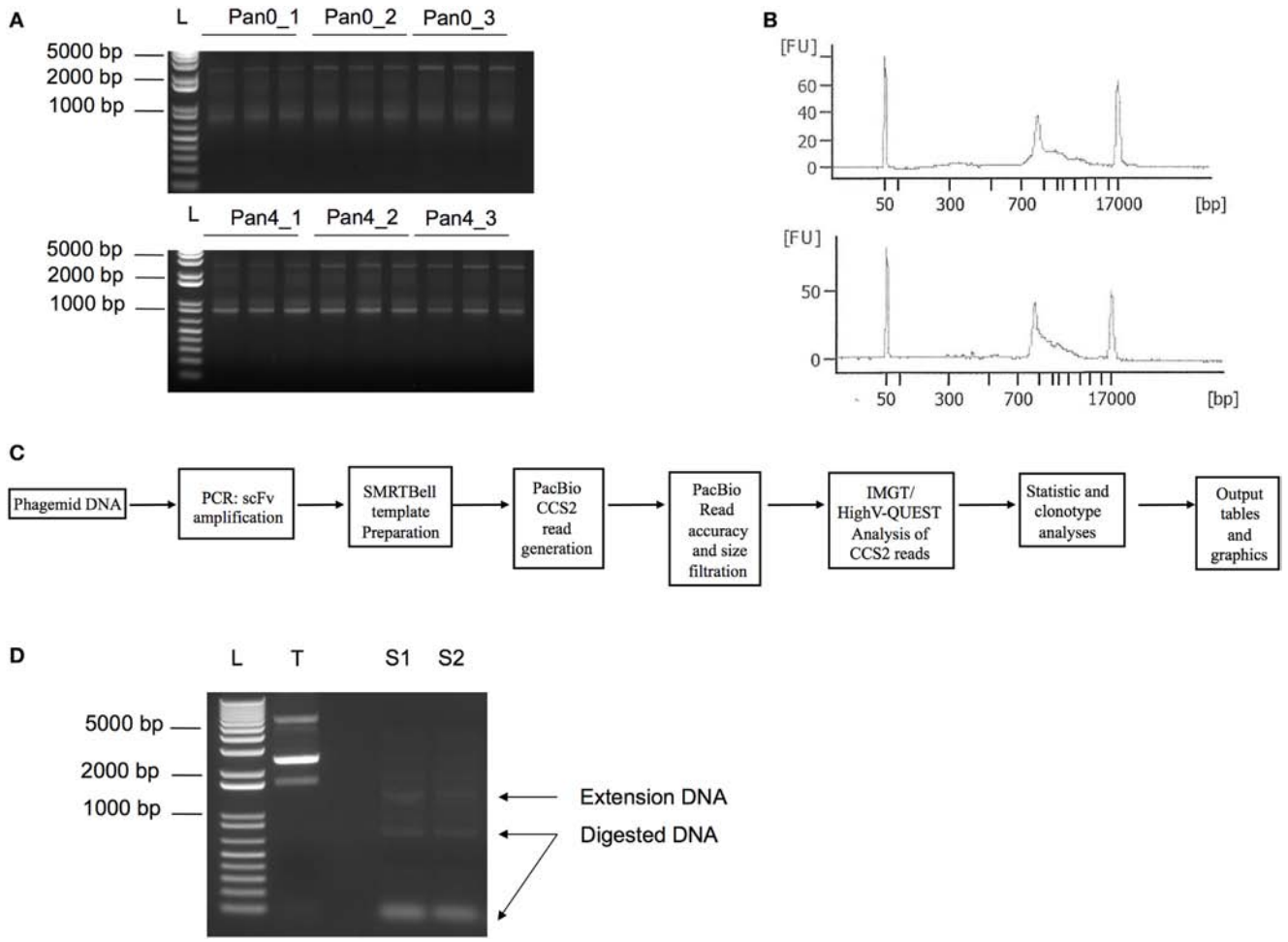


FIGURE 1 | PCR amplification of scFv libraries. scFv DNA was amplified using human IG-specific PCR primers. **(A)** Triplicate amplified samples were generated using 25, 50, and 75 ng of input DNA template for Pan0_1, Pan0_2, or Pan0_3 and Pan4_1, Pan4_2, or Pan4_3, respectively. Pan0 correspond to the unpanned library while Pan4 represents fourth round selection onto gp140. **(B)** DNA traces generated using the DNA12000 kit on the Agilent Bioanalyzer and represent the distribution of scFv amplicons based on size. **(C)** Key steps in SMRT sequencing, IMGT/HighV-QUEST and clonotype analysis. **(D)** PCR amplification of selected scFv clones chosen for functional validation, S1 and S2, followed by Dpn1 digestion. T, control non-amplified and non-digested DNA; L, 1 Kb plus DNA ladder; bp, base pair; FU, florescence units; scFv, single chain fragment variable; SMRT, single molecule, real-time.

scFv Recovery

A manual inspection of the VH CDR3 AA sequences in the newly assigned clonotypes allowed for the verification of the presence of clonotypes previously identified by handpicking and Sanger screening of clones obtained from Pan4 (11). As more than 100 additional clonotypes were observed, the validity of the PacBio-generated sequences was verified by testing the SIV-specificity of two newly identified clones belonging to two different IMGT IGHV clonotypes. Of the selected (S1 and S2) clonotypes, S1 IGHV was not abundantly represented in the pooled data, with only 3 sequences observed, while S2 IGHV was represented at a higher proportion, with 102 sequences. Two S1 and S2 clones were recovered by PCR using DNA Pan4 as the template (**Figure 1D**). The identities of the recovered clones were confirmed by Sanger sequencing and the clonotype characteristics were analyzed using IMGT/V-QUEST with scFv functionality (**Table 2**). Somatic hypermutation (SHM) frequencies were determined using nucleotide identity with the closest germline as obtained with IMGT/V-QUEST. To confirm functional activity, the recovered clones were evaluated by ELISA to characterize binding to SIV Env (**Table 2**).

V-D-J Assignments

Data files derived from the IMGT/HighV-QUEST analysis of the PacBio CCS2 reads were submitted to the statistical and IMGT clonotype analysis using the new scFv functionality tool on the IMGT/HighV-QUEST web portal (45) (www.imgt.org). The sequence analysis pipeline is provided in **Figure 1C**.

IGHV

Pan0, which reflects a broad representation of total antibody genes in this macaque, was highly diverse with a single gene (IGHV1-1, IGHV4-2, and IGHV7-1) representing the IGHV1, IGHV4, and IGHV7 subgroups, two genes (IGHV2-1 and IGHV2-2) for the IGHV2 subgroup and 10 genes (IGHV3-5, IGHV3-6, IGHV3-7,

TABLE 1 | IGHV, IGKV, and IGLV genes, sequence number and IMGT clonotype (AA) assignment.

	Pan0_1	Pan0_2	Pan0_3	Batch Pan0	Pan4_1	Pan4_2	Pan4_3	Batch Pan4
PacBio Output seq Nb	8,706	7,335	11,599	27,640	15,036	14,986	13,231	43,253
Nb of seq assigned to an IMGT clonotype (AA) for IGHV	3,837	3,159	5,238	12,269	4,594	4,593	3,884	13,087
Nb of different IMGT clonotypes (AA) for IGHV	1,887	1,685	2,569	5,313	135	141	127	247
Nb of out-of-frame seq for IGHV	117	90	115	322	106	115	80	301
Nb of seq of other categories for IGHV	112	76	103	292	1	0	0	1
Nb of seq assigned to an IMGT clonotype (AA) for IGKV	3,511	2,818	4,640	11,205	6,957	7,062	5,974	19,995
Nb of different IMGT clonotypes (AA) for IGKV	1,280	1,099	1,651	3,263	147	168	150	280
Nb of out-of-frame seq for IGKV	126	96	172	409	216	248	214	678
Nb of seq of other categories for IGKV	54	29	50	137	2	6	3	11
Nb of seq assigned to an IMGT clonotype (AA) for IGLV	3,110	2,531	4,068	9,910	7,484	7,252	6,504	21,502
Nb of different IMGT clonotypes (AA) for IGLV	1,268	1,077	1,573	3,205	588	548	535	1,078
Nb of out-of-frame seq for IGLV	209	169	235	641	581	570	488	1,639
Nb of seq of other categories for IGLV	67	60	111	240	18	22	27	67

Filtered CCS reads (≥99.9% accuracy and 700–1,200 bp in length) generated from SMRT sequencing data were analyzed using IMGT/HighV-QUEST in order to determine V, D, J gene sequence, allele assignment and functionality (53, 54). IMGT/HighV-QUEST output files were processed through IMGT statistical and clonotype analysis to determine the number of assigned sequences and clonotypes. The number of sequences assigned to an IMGT clonotype corresponds to in-frame sequences [C104, W118 for VH and C104, F118 for VL (V-KAPPA or V-LAMBDA)] from "1 copy" + "More than 1," "Single allele" and "Several alleles" (or genes) (41). Number of in-frame "other categories" sequences were not assigned to a "single gene." An IMGT clonotype (AA) was defined by a unique V-(D)-J rearrangement (with IMGT gene and allele names determined by IMGT/HighV-QUEST at the nucleotide level) and a unique CDR3-IMGT AA in-frame junction sequence (55, 56). Pan0 correspond to the unpanned library while Pan4 represent fourth round of selection onto SIV Env gp140. Batch0 and Batch4 represent the analysis of pooled data from three independent SMRT cells for Pan0 and Pan4, respectively. Nb, number; seq, sequence. PCR template concentration of 25, 50, and 75 ng are represented by Pan0_1, Pan0_2, and Pan0_3 or Pan4_1, Pan4_2, and Pan4_3, respectively.

TABLE 2 | Characteristic of two scFv clones recovered from Pan4.

scFv	S1	S2
IGHV gene and allele	IGHV1-1*01 F	IGHV3-5*01 F
IGHV SHM (%)	20.7	7.14
VH CDR3 AA sequence	ARGSKQFCSSSYCSVGFDY	AAEPLPDDWADWADYKKYGLDY
IGLV gene and allele	IGLV2S1*01 F	IGLV1-10*01 F
IGLV SHM (%)	5.65	6.07
V-LAMBDA CDR3 AA sequence	SSYAGSNTFLF	AAWDDSLSGWIF
Target	gp41	gp41

V genes and alleles, and complementarity determining regions 3 (VH CDR3 and V-LAMBDA CDR3) were determined using IMGT/V-QUEST with scFv functionality (57). Somatic hypermutation frequencies [SHM (%)] represent the nucleotide sequence divergence from the closest germline in IMGT/V-QUEST (54). scFv binding specificity was determined by enzyme-linked immunosorbent assay (ELISA).
F, functional; AA, amino acid.

IGHV3-9, IGHV3-10, IGHV3-11, IGHV3-12, IGHV3-14, IGHV3-21, and IGHV3-22) for IGHV3 (**Figure 2A**). Individually, IGHV4-2 was the most represented gene in the Pan0 dataset, with 1,734 (33.14%) assigned sequences. From a subgroup perspective, IGHV3 was the most represented with a total of 2,165 (41.31%) assigned sequences. Within IGHV3, three genes, IGHV3-6, IGHV3-9, and IGHV3-7, were the most abundant with 562, 498, and 462 assigned sequences, respectively. Overall, the number of assigned IMGT clonotypes was proportional to the number of sequences, as observed for IGHV4-2 (36.97%), IGHV3-6 (9.71%), IGHV3-9 (10.37%), and IGHV3-7 (8.54%).

For Pan4, a total of 9 IGHV genes were observed, compared to 14 in Pan0. IGHV3 remained the most represented subgroup with 6 genes (IGHV3-5, IGHV3-6, IGHV3-7, IGH3-9, and IGHV3-14), followed by IGHV2 with 2 genes (IGHV2-1 and IGHV2-2) and IGHV1 and IGHV4 with a single gene each (IGHV1-1, IGHV4-2) identified. The ranking in allelic representation was generally similar, with the striking exception of a switch between the 2 most represented genes, including 3,814 (83.05%) sequences for IGHV3-9 and 671 (14.61%) sequences for IGHV4-2. IGHV3-6 and IGHV3-7, which were well represented in Pan0, were reduced to only 2 (0.04%) and 35 (0.76%) sequences, respectively, in Pan4. IGHV3-5 was the third most represented in Pan4 with 49 (1.07%) sequences. The number of different IMGT clonotypes was again observed to be relatively proportional to the number of assigned sequences. However, IGHV3-9 showed less diversity with 42 (31.57%) distinct IMGT clonotypes compared to IGHV4-2 with 58 (43.60%) different clonotypes.

For IGHD, Pan0 contained a very diverse representation of IGHD genes (**Figure 2B**), with multiple genes for IGHD1 (8), IGHD2 (6), IGHD3 (4), IGHD4 (3), IGHD5 (3), and IGHD6 (6). In the case of IGHD7, only a single gene (IGHD7-1) was called, using 34 sequences assigned to this clonotype. With this exception, all other IGHD families were represented with a higher number of assigned sequences (from 327 to 1,331) with clonotype analyses. The number of assigned IMGT clonotypes was also highly diverse in regard to clonotypes called, with the most abundant genes having the highest numbers of IMGT clonotypes. The results obtained from Pan4 were strikingly different. The IMGT clonotype assignment of IGHD sequences was almost exclusively

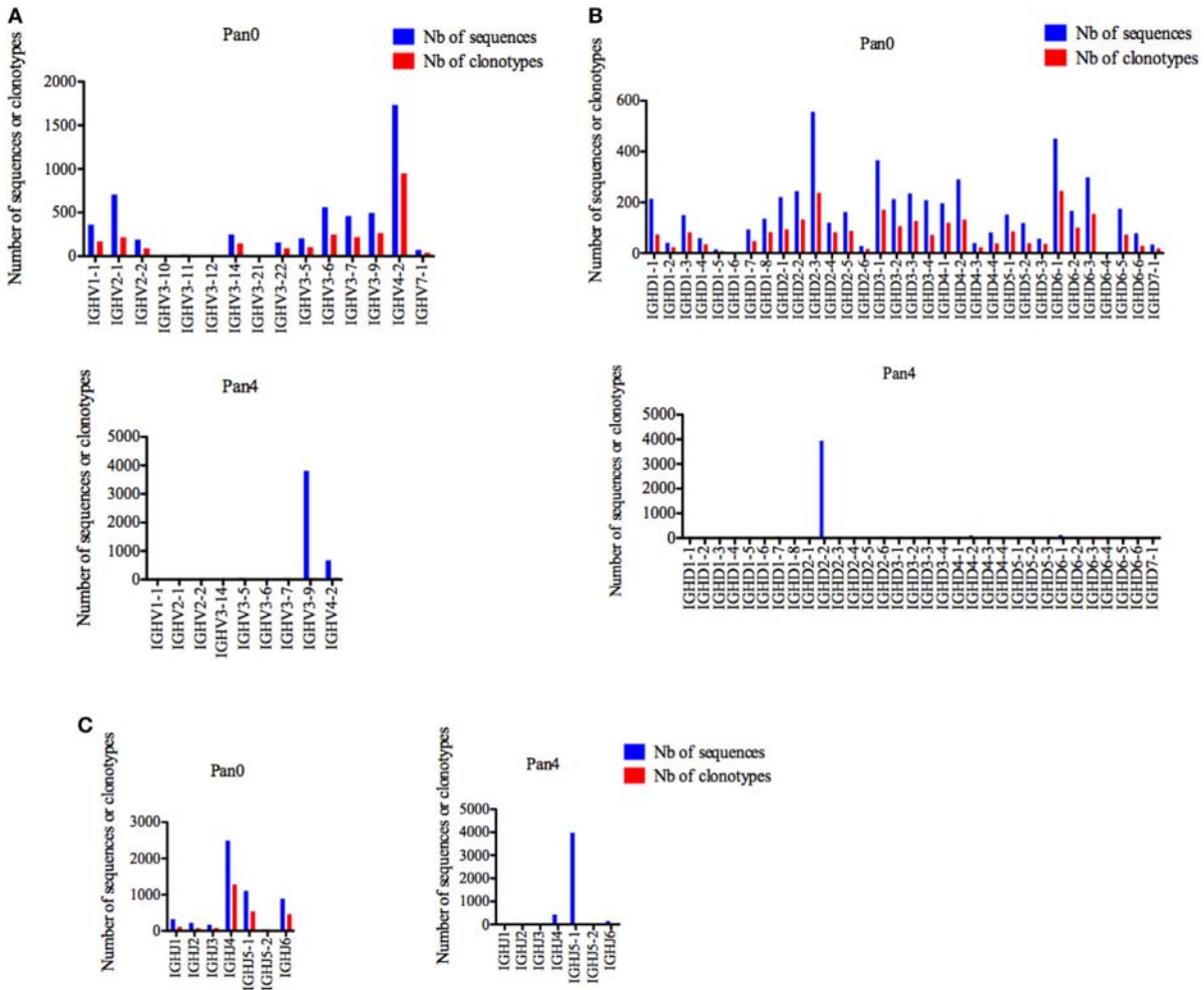


FIGURE 2 | Number of IGH sequences and clonotypes, per IGHV, IGHD, and IGHJ gene, in Pan0 and Pan4 single chain fragment variable libraries. A International ImMunoGeneTics information system (IMGT) clonotype [amino acid (AA)] was defined as a unique V-(D)-J rearrangement using IMGT gene and allele names as determined by IMGT/HighV-QUEST at the nucleotide level, and a unique, in frame variable heavy CDR3-IMGT AA junction (C104, W118) (56). **(A)** IGHV; **(B)** IGHD; **(C)** IGHJ. Nb: number.

dominated by IGHD2-2, which was comprised of 3,948 (85.93%) sequences. IGHD6-1 and IGHD4-2 were the next most abundant genes present with only 2.72 and 2.31% of the assigned sequences, respectively. Despite the enriched representation in regards to sequence number, the IGHD2-2 clonotypes seen were not particularly diverse, encompassing only 22 (16.29%) of the detected differential IMGT clonotypes, while IGHD6-1 and IGHD4-2 were represented by 11 (8.15%) and 8 (5.92%) clonotypes, respectively. Surprisingly, IGHD3-3, which had only 67 (1.45%) assigned sequences, included the most variable selection of IMGT clonotypes at 23 (17%).

Similarly, a diverse collection of IGHJ genes was observed in Pan0 (**Figure 2C**). IGHJ4 was the most abundantly represented with 2,491 (47.57%) sequences. The next most abundant were IGJ5-1 and IGHJ6 with 1,108 (21.16%) and 891 (17%) sequences, respectively. As was observed for prior Pan0 analysis, the number of different IMGT clonotypes was proportional to the most represented genes. However, in regard to Pan4, IGHJ genes were again highly enriched for the IGHJ5-1 gene, represented by 3,978 (86.59%) of the assigned sequences while IGHJ4 was only a minor group with 433 (9.42%) assigned sequences. Overall, the variety of clonotypes seen was generally low, but the most represented genes included the majority of different clonotypes.

We next turned our attention to VH CDR3 lengths, as these have been previously correlated with anti-HIV antibody activity. A variety of VH CDR3 lengths were observed in the Pan0 library, ranging from 5 AA and up to 31 AA-long (**Figure 3**). Within these VH, 78.29% had a CDR3 size of 17 AA or less, and 82.14% had a CDR3 of 18 AA or less. A similar trend was observed with the number of IMGT clonotypes, where the most prevalent VH CDR3 lengths were represented by the most diverse pool of clonotypes. In the Pan4 libraries, the VH CDR3 length extended from 9 to 25 AA, however, the vast majority of the assigned sequences (3,769, 82.04%) had a CDR3 of 20 AA. The next most abundant length observed was 15 AA with 316 (6.88%) assigned sequences, followed by 18 AA with 148 (3.22%). The numbers of

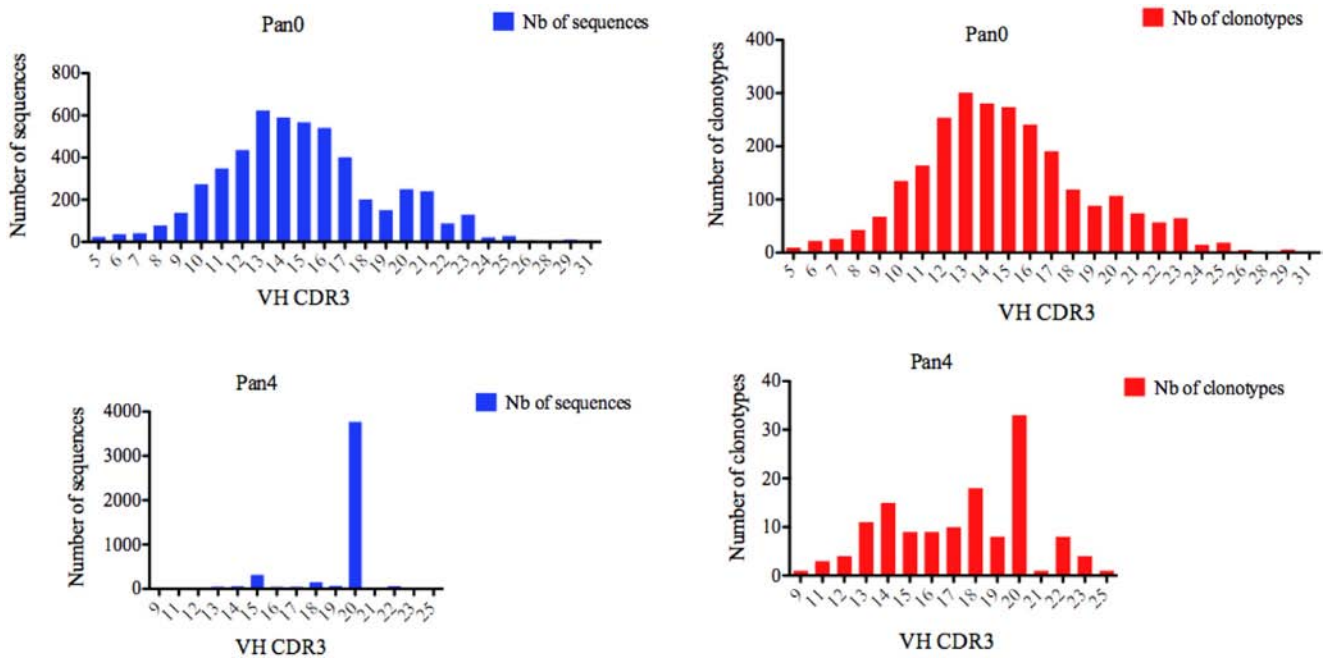


FIGURE 3 | Number of IGH sequences and clonotypes distributed by VH CDR3 length. VH CDR3-IMGT length was determined by IMGT/HighV-QUEST in productive (those with in-frame junctions, C104, W118) scFv sequences. The numbers of clonotypes per VH CDR3 length are plotted in Pan0 and Pan4 libraries. Nb: number.

different clonotypes contained within these VH CDR3 size bins were usually low. For example, the 20 AA-long VH CDR3, which represented 82.04% of the total assigned sequences, showed relatively poor diversity with only 33 (24%) of the different IMGT clonotypes represented. This was followed by the next most clonotypically diverse VH CDR3 length bins, including 15 AA CDR3 (11.11%) and 18 AA CDR3 (13.33%). Consequently, we observed a more diversified range of VH CDR3 lengths when considering clonotype distribution rather than the simple number of sequences (**Figure 3**) per VH CDR3 AA length.

IGKV

A diversified representation of IGKV genes was observed within Pan0, which identified 69 unique IGKV genes (**Figure 4A**). The most abundantly represented were IGKV3-8, IGKV3-9, IGKV3-3, and IGKV2S4 with 732 (16.24%), 408 (9.05%), 383 (8.50%), and 316 (7.01%) sequences assigned, respectively. As previously observed with IGHV, the number of different IMGT clonotypes was directly proportional to the gene abundance described above. A less diverse pool of 23 IGKV genes was obtained from the Pan4 data. These genes were dominated by the IGKV3-8 and IGKV3-9 genes with 4,525 (65.04%) and 2,043 (29.37%) of assigned sequences, respectively. The next most abundant gene characterized was IGKV7-1 with 163 (2.34%) assigned sequences. The number of different clonotypes was proportional and dominated by IGKV7-1 and IGKV1-14 with 82 (55.78%) and 16 (10.88%), respectively; IGKV7-1 contained 13 (8.84%) different clonotypes.

Five different IGKJ genes (1–5) were detected in Pan0 (**Figure 4B**), with the most abundant being IGKJ2, IGKJ4, and IGKJ1 with 1,993 (43.03%), 1,263 (27.26%), and 1,227 (26.49%), respectively. The number of different clonotypes observed was also highest for IGKJ2, IGKJ4, and IGKJ1 genes with 629 (38.21%), 549 (33.35%), and 390 (23.69%), respectively. For Pan4, 4 IGKJ genes were detected (**Figure 4B**), with IGKJ2 being the most abundant with 6,596 (94.83%) assigned sequences. A similar abundance was observed for the number of different IMGT clonotypes within IGKJ2, which was shown to have 102 (69.86%) assigned clonotypes.

For Pan0, V-KAPPA CDR3 lengths ranged from 5 AA to 29 AA long (**Figure 5**), with the most represented lengths being 9 and 8 AA, comprising 3,690 (79.52%) and 794 (17.11%) assigned sequences, respectively. A similar relationship was observed for the number of different IMGT clonotypes within these V-KAPPA CDR3 length bins, with 1,390 (29.95%) and 185 (11.2%) for 9 and 8 AA-long CDR3, respectively. In Pan4, the observed V-KAPPA CDR3 were between 8 and 10 with 9 and 8 AA-long CDR3 lengths dominating the pool as both the most represented in sequence space, as well as the most diverse, containing 122 (83.00%) and 23 (15.64) different clonotypes, respectively.

IGLV

Similar to IGKV, IGLV gene usage was also shown to be highly diverse in the Pan0 library, including 48 unique IGLV genes (**Figure 6A**). The most abundant genes were IGLV1-15, IGL8-1, and IGLV10-1 with 531 (13.15%), 413 (10.23%), and 344 (8.52%) sequences, respectively, and the number of clonotypes was proportional to that of the observed gene abundances. However, in the Pan4 sub-library only 28 different IGLV genes were observed (**Figure 6A**). IGLV1-15 and IGLV6-5 were the most abundant. IGLV1-15, the most represented gene in Pan0, was further enriched

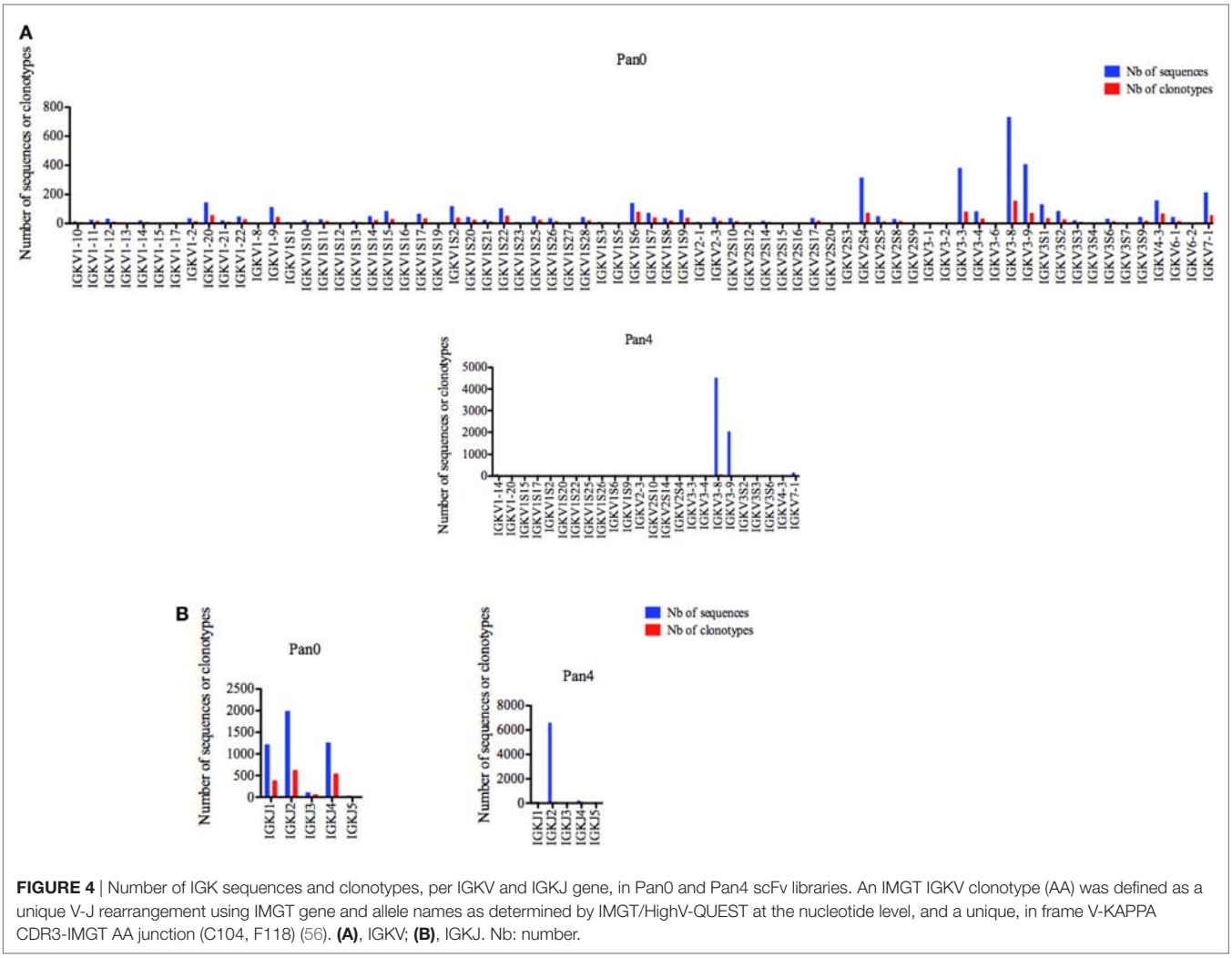


FIGURE 4 | Number of IGK sequences and clonotypes, per IGKV and IGKJ gene, in Pan0 and Pan4 scFv libraries. An IMGT IGKV clonotype (AA) was defined as a unique V-J rearrangement using IMGT gene and allele names as determined by IMGT/HighV-QUEST at the nucleotide level, and a unique, in frame V-KAPPA CDR3-IMGT AA junction (C104, F118) (56). **(A)**, IGKV; **(B)**, IGKJ. Nb: number.

in Pan4 with 2,350 (31.44%) assigned sequences. IGLV6-5, which represented only 0.71% of the scFv from the Pan0 library, was the most abundant sequence observed in Pan4 with 2,939 (39.33%) of the assigned sequences. As seen within the other IG genes, the number of clonotypes was proportional to the relative representation of each gene in the pool.

Six different IGLJ genes were identified within the Pan0 library (**Figure 6B**). IGLJ6 and IGLJ7 were the most dominant, with 1,708 (42.49%) and 793 (19.73%) sequences assigned, respectively. In the Pan4 library, 5 different IGLJ genes were observed, missing only the presence of the IGLJ4 gene as compared to the Pan0 library. Despite this similar level of diversity, the majority of sequences were IGLJ3 and IGLJ6 with 3,511 (47.09%) and 1,636 (21.94%), respectively. Once again, the number of clonotypes was again proportional to the most represented genes.

Single chain fragment variable from the Pan0 library contained V-LAMBDA CDR3 lengths ranging from 8 to 13 AA (**Figure 7**), with the majority of CDR3 lengths in the pool comprised of 11 (2,308 sequences, 56.73%), 10 (1,013 sequences, 24.90%), and 9 (608 sequences, 14.94%) AA. V-LAMBDA CDR3 lengths within the Pan4 library were observed to vary from 9 to 28 AA, but were dominated by CDR3 lengths of 11 and 10 AA, representing 4,443 (59.36%) and 2,988 (39.92%) of the assigned sequences, respectively. In both Pan0 and Pan4, the overall number of clonotypes was proportional to the assigned genes sequences; specifically in Pan4 the 11 AA V-LAMBDA CDR3 group contained the highest number of diverse clonotypes, followed by those with 10 AA CDR3, with 473 (80.44%) and 93 (15.81%) respectively.

VH-VL Combinations

A clear advantage of the SMRT sequencing over other NGS approaches is the generation of longer reads, which for phage display libraries should allow the analysis of full length scFv and identification and interrogation of productive VH-VL combinations. To address this, the IMGT/HighV-QUEST output data were manually inspected to identify VL combinations for the IGHV clonotypes corresponding to the two recovered clones. The S1 VH was associated with three productive IGVL genes, including two sequences from IGLV1-15 and one from IGLV1-10. The S2 VH was associated with 102 productive VL sequences,

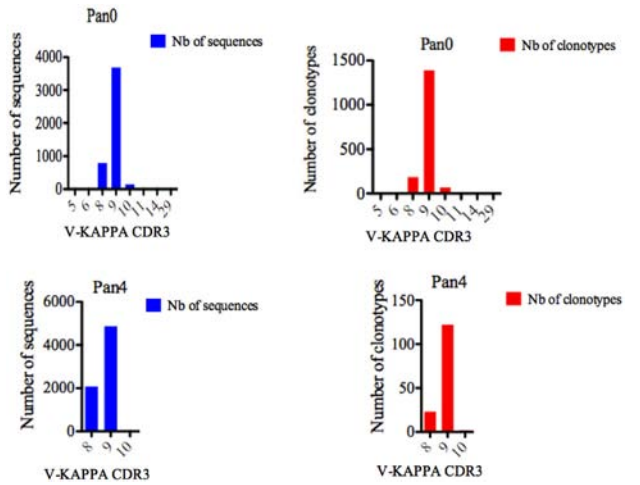


FIGURE 5 | Number of IGK sequences and clonotypes, distributed by V-KAPPA CDR3 length. V-KAPPA CDR3-IMGT length was determined by IMGT/HighV-QUEST in productive (those with in-frame junctions, C104, F118) single chain fragment variable sequences. The numbers of clonotypes per V-KAPPA CDR3 length are plotted in Pan0 and Pan4 libraries. Nb: number.

including 76 IGLV2S1, 8 IGLV6-5, 5 IGKV3-8, 3 IGLV1-15, 2 IGLV1S6, and 1 sequence assigned to other genes (IGKV3-9, IGLV1-10, IGLV1-15, IGLV1-7, IGLV1S3, IGLV2-3, IGLV2-7, and IGLV6-1).

Batch Analysis of Triplicate SMRT Cell Outputs

When screening a highly variable library, a limitation of the PacBio RSII system is the relatively low read depth per SMRT cell, as compared to other NGS technologies. We observed this when running single SMRT cells per sample in our experiments. However, this can be countered by running multiple SMRT cells per sample to increase read depth and the ability to detect low level variants within a population. As the IMGT statistical and clonotype analysis tool functions with batched data, including multiple SMRTcells (up to one million sequences), we performed a batch analysis for triplicate, pooled samples from Pan0 and Pan4 (**Table 1**). In this batched dataset, for IGHV, the number of sequences assigned to an IMGT clonotype was increased (two- to fourfold) for Pan0 and almost threefold in Pan4 as compared to those sequences generated with a single SMRT cell per sample (Pan0_1, Pan0_2, and Pan0_3 or Pan4_1, Pan4_2, and Pan4_3). The number of IMGT clonotypes was also approximately twofold of that called using single SMRT cell data for both Pan0 and Pan4. Similar increases in the number of sequences and clonotypes were also observed for IGKV and IGLV when combining outputs from triplicate SMRT cells for both Pan0 and Pan4 (**Table 1**). These observations highlight the need to consider increasing depth of coverage when screening highly variable libraries using single molecule sequencing approaches, either through batching multiple cells of data or generating sequence data using a higher throughput instrument (i.e., PacBio Sequel system).

DISCUSSION

Understanding the role of polyclonal nnAbs in the inhibition of HIV-1 infection will require large-scale analyses of virus-specific antibodies. These analyses should evaluate the functionality of the antibodies, as well as characterize their structure, including V-(D)-J gene or allele usage and CDR3 lengths. Techniques that rely on antigen-based selection of single B cells or screening of single colonies following phage display generally result in a low number of sequences and are not always suited to examine the expected level of complexity in such a system. Broadly used NGS technologies, such as Illumina, can generate very high numbers of sequence reads, but are limited in read length and can cover no more than one variable domain (VH or VL) per read, limiting data interpretation and the ability to evaluate the impact of specific VH–VL pairing.

Here, we evaluated the ability of SMRT sequencing technology to generate full length scFv sequencing reads for the analysis of phage display libraries. We obtained thousands of highly accurate (≥99.9%) sequence reads, each containing productive VH-VL combinations. As previously reported for this rhesus macaque (11), SIV Env gp140-specific scFv were dominated by the usage of IGHV4-2 and IGHV3-9. However, we observed far fewer IGHV1-1-containing scFvs in the Pan4 library than expected. We also observed more IGHV3-9-contaning scFv sequences than those with IGHV4-2, despite the latter being more abundant in the unpanned library. Together, these data suggest that the SMRT sequencing approach will be useful for the characterization of scFv libraries as more comprehensive rhesus V gene databases are being developed (60, 61).

The libraries examined were very diverse and included a great number of distinct clonotypes that were, in general, proportional to the number of sequences obtained for each gene. Interestingly, SIV Env gp140-specific scFvs were characterized by a more limited number of clonotypes, as has been previously reported for both HIV-1 and SIV antibodies (12, 22, 23, 32, 62). Specifically, the most dramatic examples were observed with IGHD and IGHJ genes, in which more than 85% of sequences contained IGHD2-2 and IGHJ5-1. It will be necessary and informative to determine whether these patterns are similar in other experimentally infected animals, particularly in those immunized with SIVmac239Δnef, which show protection from SIVmac239 challenge infection. Though less pronounced in terms of clonotype distribution, SIV Env gp140-specific scFvs presented an average of VH CDR3 size of 20 AA (representing around 80% of total sequences), which was far greater than that seen in the Pan0 library, containing the same proportion of sequences for VH CDR3 ranging between 5 and 17 AA. It is difficult to interpret the relevance of these results to nnAbs in particular, as the plasma of this rhesus macaque also contained a high titer of neutralizing antibodies. Understanding the impact of VH CDR3 length on antibody functionality and disease outcome will definitely require a similar analysis across a greater number of animals.

Additional high-quality sequences were obtained for V-LAMBDA antibodies containing a diversified pool of genes from the major IGLV subgroups. IGVL genes from SIV Env gp140-specific scFvs were generally derived from among the most

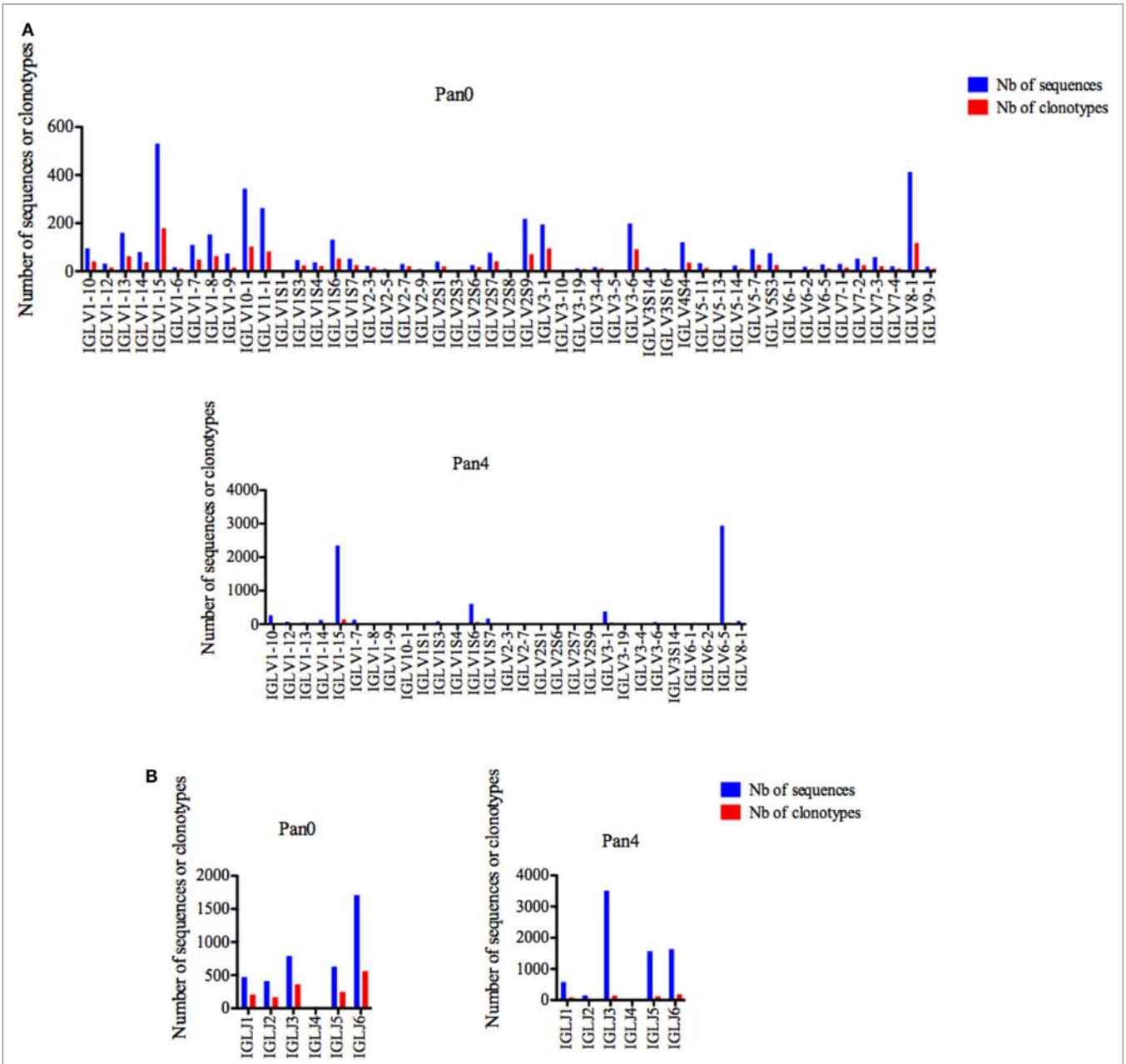


FIGURE 6 | Number of IGL sequences and clonotypes, per IGLV and IGLJ gene, in Pan0 and Pan4 single chain fragment variable libraries. An International ImMunoGeneTics information system (IMGT) IGLV clonotype (AA) was defined as a unique V-J rearrangement using IMGT gene and allele names as determined by IMGT/HighV-QUEST at the nucleotide level, and a unique, in frame V-LAMBDA CDR3-IMGT AA junction (C104, F118) (56). **(A)**, IGLV; **(B)**, IGLJ. Nb: number.

represented genes in the Pan0 library, with the exception of IGLV6-5, which showed a ~40-fold increase in representation in the SIV Env gp140-specific scFv library, compared to that from the total, unpanned IG repertoire. Due to the combinatory nature of VH-VL association of our phage display libraries, further analyses will certainly be required to clarify and understand the potential contribution of diverse IGKV and IGLV genes to the overall antibody functionality.

It has also been observed that the IGHV1 (34.4%) and IGHV4 (26.15%) subgroups were the most dominant among naïve IgM libraries in human adults but that these representations could be altered during early (40 days) or advanced (8 months) HIV-1 infection with differences between IgM and IgG frequencies, as well as between peripheral blood mononuclear cells (PBMC) and bone marrow (BM) compartments (63). During early HIV-1 infection, IGHV1 and IGHV3 (respectively, 40.96% in PBMC and 51.22% in BM IgM libraries, and, respectively, 35.13% in BM and 33.8% in PBMC IgG libraries) were the most dominant while IGHV3 and IGHV4 (respectively, 38.42% in PBMC and 23.69% in BM IgM libraries), and specifically IGHV4 in IgG libraries (PBMC

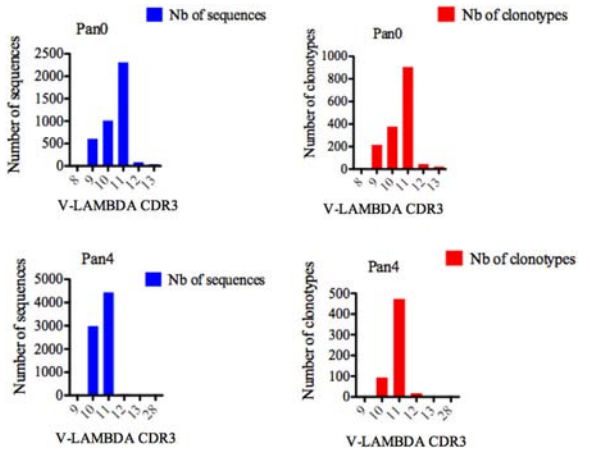


FIGURE 7 | Number of IGL sequences and clonotypes, distributed by V-LAMBDA CDR3 length. V-LAMBDA CDR3-IMGT length was determined by IMGT/HighV-QUEST in productive (those with in-frame junctions, C104, F118) single chain fragment variable sequences. The numbers of clonotypes per V-LAMBDA CDR3 length are plotted in Pan0 and Pan4 libraries. Nb: number.

27.89% and BM 28.32%) were the second most dominant (63). During the more advanced (8 month) stage of infection, IGHV3 (57.10% in PBMC, 59.31% in BM IgM libraries) and IGHV1 (53.03% in PBMC and 46.29% in BM IgG libraries) were the most dominant while IGHV4 and IGHV1 (respectively, 17.18% in PBMC and 15.98% in BM IgM libraries), and IGHV3 (22.96% in PBMC, 29.15% in BM IgG libraries) were the second most dominant (63). The complexity of these data suggest that analyses of multiple animals and infection time points are needed, as well as a comparative analyses of these data compared to those generated using scFv libraries constructed with rhesus specific-primers to ensure that any rhesus-specific IG populations are included (9, 12). These additions, as well as leveraging the growing rhesus databases (60, 61) will be necessary moving forward to ensure the most comprehensive analysis of polyclonal antibody responses during SIV infection in rhesus macaques.

There are several caveats that are worth noting regarding the approaches used to generate the V-(D)-J composition and CDR3 lengths of SIV-specific antibodies identified and characterized in this study.

The first caveat surrounds the relatively low read depth generated by the SMRT sequencing approach. As mentioned earlier in the text, He et al. generated over one million raw reads (750–950 bp) using the PGM-S5 system in a similar study investigating HIV-specific scFv antibodies. Interestingly, these data demonstrated similar sequence output size from the IMGT statistical and clonotype analysis. Hemadou et al. analyzed more than 450,000 reads using a SMRT sequencing strategy, but by combining the data from 15 SMRT cells to increase depth. In the current work, we did observe an increase (two- to fourfold) of individual sequences or clonotypes called when 3 SMRT cells of data were combined prior to clonotype analysis. While the current data are already hypothesis generating, the increased information gained by increasing read depth underscores the need for future, similar studies to batch data from multiple RSII SMRT cells or to move this type of study to the higher throughput Sequel system, which generates 400,000–500,000 single molecule reads per chip, drastically increasing the single molecule read depth available.

The next caveat relates to our use of human-based primers for the construction of phage display library under examination. Overall, the rhesus macaque genome is highly homologous to that of the human (~93%) (64). In particular, Sundling et al. found a high level of homology (~92%) between the human and macaque functional VH regions, with genes clustered according to family distribution rather than species when examined by phylogenetic analysis (65). However, it is noted that these authors also identified unique macaque-specific antibody gene families. Consequently, the phage display library used in the current study is likely biased toward gene families with high homology to the corresponding human genes. It is the focus of our immediate future work to recapitulate these data and expand upon them with the use of macaque-specific primers (9, 12, 62, 65–67). For example, Dai et al. observed a diverse repertoire of VH CDR3 length by tracking lineages of CD4-binding site-directed mAbs in macaques immunized with an HIV-1 trimer vaccine (62). Our VH CDR3 results differed from these observations, which may be due to this type of primer bias, or due to the fact that our antibodies are targeting the MPER region of Env, rather than the receptor-binding site. In fact, our variable light chain results were comparable to those of Dai et al., where it was observed that the 1 or 2 CDR3 AA length was most abundant for V-KAPPA CDR3 and V-LAMBDA CDR3. A related caveat pertains generally to the characterization of all combinatorial phage display approaches, mainly that our libraries do not represent authentic VH–VL pairing. Methods to clone paired VH-VL regions from single macaque B cells have been developed (9, 66). These studies relied on Sanger sequencing while other macaque-based studies did not apply NGS to characterize authentic VH–VL pairs (62, 65). Moving forward, the use of combinatorial phage display technology should be complemented with single cell-based methodologies, including droplet emulsion-based approaches, to resolve relevant VH–VL pairing (68–72).

Another caveat that could impact our V-(D)-J characterization of SIV gp140-specific antibodies was the use of the IMGT database for classification and analyses, as it possesses an incomplete representation of macaque IG germline sequences. In their study Dai et al. address this concern by comparing antibody V-(D)-J compositions obtained *via* their in-house "CS germline-gene database" to that of the IMGT database, observing a more diversified VH family composition (mostly VH1, VH2, and VH3) using their in-house tool, while IMGT-generated data were skewed toward VH1 and VH4. Although we exclusively used IMGT analyses in this study, we also found that VH3 and VH4 family genes were among the most represented within the SIV gp140-specific antibodies. These two families were followed by VH1 albeit to a lesser level than what was observed by Dai et al. or our previous study (11). It is possible that the differences between our results and those from Dai et al. were individual specific, or due to the varied panning antigens used. Using Sanger sequencing, we have previously observed VH4-skewed usage in gp120-

specific antibodies from the same animal examined in this current study (11). Most importantly, we are currently working with IMGT to expand their macaque IG germline database, as to eliminate this caveat in future studies. To note, Dai et al. also observed similar patterns of VH CDR3 lengths and SHM levels between IMGT and their own "CS germline-gene database" results. Here, we also observed a wide range of VH CDR3 lengths in the context of differential clonotype identities, which is probably more relevant for functional antibody classification.

Lastly, engineering of native-like SIV Env scaffolds may be necessary for the identification of highly functional antibodies, particularly those with potent neutralization activity. For instance, we have previously demonstrated the relative stability of non-engineered soluble SIVmac239 Env gp140 trimer and its ability to deplete SIV-infected macaque plasma of antibodies with neutralization activity against the neutralization sensitive SIVmac316 isolate (20). The depletion was less effective against neutralization resistant SIVmac239, although this was expected as the plasma did not display much neutralization activity against SIVmac239 in the first place (20). Subsequent follow up proved challenging, as we were unable to isolate potent SIV neutralizing antibodies although the phage display library had been constructed from an SIV-infected macaque with unusually high neutralization titers (11). This failure may have be due to the use of soluble SIVmac239 Env gp140 trimer, which likely would be improved through specific engineering as has been described for HIV-1 (22, 23, 30, 31, 33, 73).

AUTHOR CONTRIBUTIONS

IF, RS, MLS, and M-PL designed the study and wrote the manuscript. IF, SH, DH, BC, WC, and MS performed scFv DNA preparation. AA and GD performed PacBio the SMRT sequencing. IF performed scFv recovery from sublibrary. SK, PD, and M-PL developed and assisted with the IMGT analysis tools and web portal (www.imgt.org). All other authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Arthur Lavoie and Karthik Kalyan for maintaining the IMGT/HighV-QUEST web portal and the entire IMGT team in developing databases and tools.

FUNDING

IF was supported by a Boston college fund for undergraduate research. This work was granted access to the HPC@LR and to the High-Performance Computing (HPC) resources of the Centre Informatique National de l'Enseignement Supérieur (CINES) and to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) under the allocation [036029] (2010-2017) made by GENCI (Grand Equipement National de Calcul Intensif).

REFERENCES

1. Burbacher TM, Grant KS. Nonhuman primates as animal models for toxicology research. *Curr Protoc Toxicol* (2001) Chapter 1:Unit1.1. doi:10.1002/0471140856.tx0101s00
2. Lynch RM, Yamamoto T, Mcdermott AB. HIV vaccine research and discovery in the nonhuman primates model: a unified theory in acquisition prevention and control of SIV infection. *Curr Opin HIV AIDS* (2013) 8(4):288–94. doi:10.1097/COH.0b013e328361cfe8
3. Geisbert TW, Pushko P, Anderson K, Smith J, Davis KJ, Jahrling PB. Evaluation in nonhuman primates of vaccines against Ebola virus. *Emerg Infect Dis* (2002) 8(5):503–7. doi:10.3201/eid0805.010284
4. Scanga CA, Flynn JL. Modeling tuberculosis in nonhuman primates. *Cold Spring Harb Perspect Med* (2014) 4(12):a018564. doi:10.1101/cshperspect.a018564
5. Sato S, Johnson W. Antibody-mediated neutralization and simian immunodeficiency virus models of HIV/AIDS. *Curr HIV Res* (2007) 5(6):594–607. doi:10.2174/157016207782418515
6. Davis AS, Taubenberger JK, Bray M. The use of nonhuman primates in research on seasonal, pandemic and avian influenza, 1893-2014. *Antiviral Res* (2015) 117:75–98. doi:10.1016/j.antiviral.2015.02.011
7. Keitel WA, Treanor JJ, El Sahly HM, Evans TG, Kopper S, Whitlow V, et al. Evaluation of a plasmid DNA-based anthrax vaccine in rabbits, nonhuman primates and healthy adults. *Hum Vaccin* (2009) 5(8):536–44. doi:10.4161/hv.5.8.8725
8. Barbas CF, Burton DR, Scott JK, Silverman GJ III. *Phage Dispaly: A Laboratory Manual*. New York: Cold Spring Harbor Laboratory Press (2001).
9. Sundling C, Phad G, Douagi I, Navis M, Karlsson Hedestam GB. Isolation of antibody V(D)J sequences from single cell sorted rhesus macaque B cells. *J Immunol Methods* (2012) 386(1–2):85–93. doi:10.1016/j.jim.2012.09.003
10. Andris-Widhopf J, Steinberger P, Fuller R, Rader C, Barbas CF III. Generation of human scFv antibody libraries: PCR amplification and assembly of light- and heavy-chain coding sequences. *Cold Spring Harb Protoc* (2011) 2011(9):1139–50. doi:10.1101/pdb.prot065573pdb.prot065573
11. Ita S, Agostinho MR, Sullivan K, Yub Han S, Akleh R, Johnson WE, et al. Analysis of SIVmac envelope-specific antibodies selected through phage display. *AIDS Res Hum Retroviruses* (2017) 33(8):869–79. doi:10.1089/AID.2016.0247
12. Sundling C, Zhang Z, Phad GE, Sheng Z, Wang Y, Mascola JR, et al. Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *J Immunol* (2014) 192(8):3637–44. doi:10.4049/jimmunol.1303334
13. Johnson WE, Lifson JD, Lang SM, Johnson RP, Desrosiers RC. Importance of B-cell responses for immunological control of variant strains of simian immunodeficiency virus. *J Virol* (2003) 77(1):375–81. doi:10.1128/JVI.77.1.375-381.2003
14. Laird ME, Igarashi T, Martin MA, Desrosiers RC. Importance of the V1/V2 loop region of simian-human immunodeficiency virus envelope glycoprotein gp120 in determining the strain specificity of the neutralizing antibody response. *J Virol* (2008) 82(22):11054–65. doi:10.1128/JVI.01341-08
15. Laird ME, Desrosiers RC. Infectivity and neutralization of simian immunodeficiency virus with FLAG epitope insertion in gp120 variable loops. *J Virol* (2007) 81(20):10838–48. doi:10.1128/JVI.00831-07
16. Adnan S, Reeves RK, Gillis J, Wong FE, Yu Y, Camp JV, et al. Persistent low-level replication of SIVΔnef drives maturation of antibody and CD8 T cell responses to induce protective immunity against vaginal SIV infection.*PLoS Pathog* (2016) 12(12):e1006104. doi:10.1371/journal.ppat.1006104
17. Pantophlet R, Burton DR. GP120: target for neutralizing HIV-1 antibodies. *Annu Rev Immunol* (2006) 24(1):739–69. doi:10.1146/annurev.immunol.24.021605.090557
18. Kwong PD, Mascola JR, Nabel GJ. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nat Rev Immunol* (2013) 13(9):693–701. doi:10.1038/nri3516
19. Kwong PD, Mascola JR. Human antibodies that neutralize HIV-1: identification, structures, and B Cell ontogenies. *Immunity* (2012) 37(3):412–25. doi:10.1016/j.immuni.2012.08.012
20. Fofana IB, Colantonio AD, Reeves RK, Connole MA, Gillis JM, Hall LR, et al. Flow cytometry based identification of simian immunodeficiency virus Env-specific B lymphocytes. *J Immunol Methods* (2011) 370(1–2):75–85. doi:10.1016/j.jim.2011.05.010S0022-1759(11)00112-8

21. Mascola JR, Montefiori DC. The role of antibodies in HIV vaccines. *Annu Rev Immunol*(2010)28(1):413–44.doi:10.1146/annurev-immunol-030409-101256
22. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TY, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* (2011) 333(6049):1633–7. doi:10.1126/science.1207227
23. Scheid JF, Mouquet H, Feldhahn N, Seaman MS, Velinzon K, Pietzsch J, et al. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* (2009) 458(7238):636–40. doi:10.1038/nature07930
24. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* (2014) 509(7498):55–62. doi:10.1038/nature13036nature13036
25. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* (2010) 329(5993):856–61. doi:10.1126/science.1187659science.1187659
26. Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, Goss JL, et al. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* (2009) 326(5950):285–9. doi:10.1126/science.1178746
27. Gallerano D, Cabauatan CR, Sibanda EN, Valenta R. HIV-specific antibody responses in HIV-infected patients: from a monoclonal to a polyclonal view. *Int Arch Allergy Immunol* (2015) 167(4):223–41. doi:10.1159/000438484
28. Pejchal R, Doores KJ, Walker LM, Khayat R, Huang PS, Wang SK, et al. A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. *Science* (2011) 334(6059):1097–103. doi:10.1126/science.1213256
29. West AP, Diskin R, Nussenzweig MC, Bjorkman PJ. Structural basis for germ-line gene usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120. *Proc Natl Acad Sci U S A* (2012) 109(30):E2083–90. doi:10.1073/pnas.1208984109
30. McGuire AT, Glenn JA, Lippy A, Stamatatos L. Diverse recombinant HIV-1 Envs fail to activate B cells expressing the germline B cell receptors of the broadly neutralizing anti-HIV-1 antibodies PG9 and 447-52D. *J Virol* (2014) 88(5):2645–57. doi:10.1128/JVI.03228-13JVI.03228-13
31. Jardine J, Julien JP, Menis S, Ota T, Kalyuzhniy O, McGuire A, et al. Rational HIV immunogen design to target specific germline B cell receptors. *Science* (2013) 340(6133):711–6. doi:10.1126/science.1234150science.1234150
32. Hoot S, McGuire AT, Cohen KW, Strong RK, Hangartner L, Klein F, et al. Recombinant HIV envelope proteins fail to engage germline versions of anti-CD4bs bNAbs. *PLoS Pathog* (2013) 9(1):e1003106. doi:10.1371/journal.ppat.1003106PPATHOGENS-D-12-02139
33. McGuire AT, Hoot S, Dreyer AM, Lippy A, Stuart A, Cohen KW, et al. Engineering HIV envelope protein to activate germline B cell receptors of broadly neutralizing anti-CD4 binding site antibodies. *J Exp Med* (2013) 210(4):655–63. doi:10.1084/jem.20122824jem.20122824
34. Sok D, Le KM, Vadnais M, Saye-Francisco KL, Jardine JG, Torres JL, et al. Rapid elicitation of broadly neutralizing antibodies to HIV by immunization in cows. *Nature* (2017) 548(7665):108–11. doi:10.1038/nature23301
35. Pollara J, Bonsignori M, Moody MA, Liu P, Alam SM, Hwang KK, et al. HIV-1 vaccine-induced C1 and V2 Env-specific antibodies synergize for increased antiviral activities. *J Virol* (2014) 88(14):7715–26. doi:10.1128/JVI.00156-14
36. Zolla-Pazner S, deCamp AC, Cardozo T, Karasavvas N, Gottardo R, Williams C, et al. Analysis of V2 antibody responses induced in vaccinees in the ALVAC/AIDSVAX HIV-1 vaccine efficacy trial. *PLoS One* (2013) 8(1):e53629. doi:10.1371/journal.pone.0053629
37. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, et al. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N Engl J Med* (2012) 366(14):1275–86. doi:10.1056/NEJMoa1113425
38. Karasavvas N, Billings E, Rao M, Williams C, Zolla-Pazner S, Bailer RT, et al. The Thai Phase III HIV Type 1 Vaccine trial (RV144) regimen induces antibodies that target conserved regions within the V2 loop of gp120. *AIDS Res Hum Retroviruses* (2012) 28(11):1444–57. doi:10.1089/aid.2012.0103
39. Alpert MD, Harvey JD, Lauer WA, Reeves RK, Piatak M Jr, Carville A, et al. ADCC develops over time during persistent infection with live-attenuated SIV and is associated with complete protection against SIVmac251 challenge. *PLoS Pathog* (2012) 8(8):e1002890. doi:10.1371/journal.ppat.1002890
40. He L, Lin X, de Val N, Saye-Francisco KL, Mann CJ, Augst R, et al. Hidden lineage complexity of glycan-dependent HIV-1 broadly neutralizing antibodies uncovered by digital panning and native-like gp140 trimer. *Front Immunol* (2017) 8:1025. doi:10.3389/fimmu.2017.01025
41. Hemadou A, Giudicelli V, Smith ML, Lefranc MP, Duroux P, Kossida S, et al. Pacific biosciences sequencing and IMGT/HighV-QUEST analysis of full-length single chain fragment variable from an in vivo selected phage-display combinatorial library. *Front Immunol* (2017) 8:1796. doi:10.3389/fimmu.2017.01796
42. Doores KJ. The HIV glycan shield as a target for broadly neutralizing antibodies. *FEBS J* (2015) 282(24):4679–91. doi:10.1111/febs.13530
43. Crispin M, Doores KJ. Targeting host-derived glycans on enveloped viruses for antibody-based vaccine design. *Curr Opin Virol* (2015) 11:63–9. doi:10.1016/j.coviro.2015.02.002
44. Doores KJ, Kong L, Krumm SA, Le KM, Sok D, Laserson U, et al. Two classes of broadly neutralizing antibodies within a single lineage directed to the high-mannose patch of HIV envelope. *J Virol* (2015) 89(2):1105–18. doi:10.1128/JVI.02905-14
45. Walker LM, Simek MD, Priddy F, Gach JS, Wagner D, Zwick MB, et al. A limited number of antibody specificities mediate broad and potent serum neutralization in selected HIV-1 infected individuals. *PLoS Pathog* (2010) 6(8):11–2. doi:10.1371/journal.ppat.1001028
46. Mouquet H, Scharf L, Euler Z, Liu Y, Eden C, Scheid JF, et al. Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proc Natl Acad Sci U S A* (2012) 109(47):E3268–77. doi:10.1073/pnas.1217207109
47. Sok D, Doores KJ, Briney B, Le KM, Saye-Francisco KL, Ramos A, et al. Promiscuous glycan site recognition by antibodies to the high-mannose patch of gp120 broadens neutralization of HIV. *Sci Transl Med* (2014) 6:236. doi:10.1126/scitranslmed.3008104
48. Garces F, Lee JH, de Val N, de la Pena AT, Kong L, Puchades C, et al. Affinity maturation of a potent family of HIV antibodies is primarily focused on accommodating or avoiding glycans. *Immunity* (2015) 43(6):1053–63. doi:10.1016/j.immuni.2015.11.007
49. Kong L, Lee JH, Doores KJ, Murin CD, Julien JP, McBride R, et al. Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. *Nat Struct Mol Biol* (2013) 20(7):796–803. doi:10.1038/nsmb.2594
50. Sato S, Yuste E, Lauer WA, Chang EH, Morgan JS, Bixby JG, et al. Potent antibody-mediated neutralization and evolution of antigenic escape variants of simian immunodeficiency virus strain SIVmac239 in vivo. *J Virol* (2008) 82(19):9739–52. doi:10.1128/JVI.00871-08JVI.00871-08
51. National Research Council (US) Committee for the Update of the Guide for the Care and Use of Laboratory Animals. *Guide for the Care and Use of Laboratory Animals*. Washington, DC: National Academies Press (2011).
52. Smith ML, Murrell B, Eren K, Ignacio C, Landais E, Weaver S, et al. Rapid sequencing of complete env genes from primary HIV-1 samples. *Virus Evol* (2016) 2(2):8. doi:10.1093/ve/vew018
53. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP. 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* (2012) 7(1):339. doi:10.1038/nmat3328
54. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
55. Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) 4:2333. doi:10.1038/ncomms3333
56. Lefranc M-P. Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of immunoinformatics. *Front Immunol* (2014) 5:22. doi:10.3389/fimmu.2014.00022
57. Giudicelli V, Duroux P, Kossida S, Lefranc MP. IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. *BMC Immunol* (2017) 18(1):35. doi:10.1186/s12865-017-0218-8
58. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin

and T cell receptor nucleotide sequences. *Nucleic Acids Res* (2006) 34:D781–4. doi:10.1093/nar/gkj088
59. Sasso E, Paciello R, D'Auria F, Riccio G, Froechlich G, Cortese R, et al. One-step recovery of scFv clones from high-throughput sequencing-based screening of phage display libraries challenged to cells expressing native claudin-1. *Biomed Res Int* (2015) 2015:703213. doi:10.1155/2015/703213
60. Francica JR, Sheng Z, Zhang Z, Nishimura Y, Shingai M, Ramesh A, et al. Analysis of immunoglobulin transcripts and hypermutation following SHIVAD8 infection and protein-plus-adjuvant immunization. *Nat Commun* (2015) 6:6565. doi:10.1038/ncomms7565
61. Corcoran MM, Phad GE, Vázquez Bernat N, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
62. Dai K, He L, Khan SN, O'Dell S, McKee K, Tran K, et al. Rhesus macaque B-cell responses to an HIV-1 trimer vaccine revealed by unbiased longitudinal repertoire analysis. *MBio* (2015) 6(6):e1375–1315. doi:10.1128/mBio.01375-15
63. Xiao M, Prabakaran P, Chen W, Kessing B, Dimitrov DS. Deep sequencing and Circos analyses of antibody libraries reveal antigen-driven selection of Ig VH genes during HIV-1 infection. *Exp Mol Pathol* (2013) 95(3):357–63. doi:10.1016/j.yexmp.2013.10.004
64. Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* (2007) 316(5822):222–34. doi:10.1126/science.1139247
65. Sundling C, Li Y, Huynh N, Poulsen C, Wilson R, O'Dell S, et al. High-resolution definition of vaccine-elicited B cell responses against the HIV primary receptor binding site. *Sci Transl Med* (2012) 4:142. doi:10.1126/scitranslmed.3003752
66. Meng W, Li L, Xiong W, Fan X, Deng H, Bett AJ, et al. Efficient generation of monoclonal antibodies from single rhesus macaque antibody secreting cells. *MAbs* (2015) 7(4):707–18. doi:10.1080/19420862.2015.1051440
67. Margolin DH, Saunders EH, Bronfin B, de Rosa N, Axthelm MK, Goloubeva OG, et al. Germinal center function in the spleen during simian HIV infection in rhesus monkeys. *J Immunol* (2006) 177(2):1108–19. doi:10.4049/jimmunol.177.2.1108
68. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32(2):158–68. doi:10.1038/nbt.2782
69. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2015) 21(1):86–91. doi:10.1038/nm.3743
70. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31(2):166–9. doi:10.1038/nbt.2492
71. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113(19):E2636–45. doi:10.1073/pnas.1525510113
72. Mcdaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G. Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat Protoc* (2016) 11(3):429–42. doi:10.1038/nprot.2016.024
73. Mouquet H, Warncke M, Scheid JF, Seaman MS, Nussenzweig MC. Enhanced HIV-1 neutralization by antibody heteroligation. *Proc Natl Acad Sci U S A* (2012) 109(3):875–80. doi:10.1073/pnas.1120059109

Conflict of Interest Statement: 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.

Copyright © 2018 Han, Antoine, Howard, Chang, Chang, Slein, Deikus, Kossida, Duroux, Lefranc, Sebra, Smith and Fofana. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 02 March 2018
doi: 10.3389/fimmu.2018.00413

Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification

Jeliazko R. Jeliazkov[1†], Adnan Sljoka[2†], Daisuke Kuroda[3,4], Nobuyuki Tsuchimura[2], Naoki Katoh[2], Kouhei Tsumoto[3,5] and Jeffrey J. Gray[1,6,7,8*]*

[1] Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD, United States, [2] Department of Informatics, School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo, Japan, [3] Department of Bioengineering, School of Engineering, The University of Tokyo, Tokyo, Japan, [4] Medical Device Development and Regulation Research Center, School of Engineering, The University of Tokyo, Tokyo, Japan, [5] Laboratory of Medical Proteomics, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, [6] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, United States, [7] Institute for NanoBioTechnology, Johns Hopkins University, Baltimore, MD, United States, [8] Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, United States

OPEN ACCESS

Edited by:
Gregory C. Ippolito,
University of Texas at Austin,
United States

Reviewed by:
Roy Mariuzza,
University of Maryland,
College Park, United States
Oana Izabela Lungu,
Signature Science, LLC,
United States

***Correspondence:**
Adnan Sljoka
adnanslj@gmail.com;
Jeffrey J. Gray
jgray@jhu.edu

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

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: 06 December 2017
Accepted: 14 February 2018
Published: 02 March 2018

Citation:
Jeliazkov JR, Sljoka A, Kuroda D, Tsuchimura N, Katoh N, Tsumoto K and Gray JJ (2018) Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. Front. Immunol. 9:413. doi: 10.3389/fimmu.2018.00413

Antibodies can rapidly evolve in specific response to antigens. Affinity maturation drives this evolution through cycles of mutation and selection leading to enhanced antibody specificity and affinity. Elucidating the biophysical mechanisms that underlie affinity maturation is fundamental to understanding B-cell immunity. An emergent hypothesis is that affinity maturation reduces the conformational flexibility of the antibody's antigen-binding paratope to minimize entropic losses incurred upon binding. In recent years, computational and experimental approaches have tested this hypothesis on a small number of antibodies, often observing a decrease in the flexibility of the complementarity determining region (CDR) loops that typically comprise the paratope and in particular the CDR-H3 loop, which contributes a plurality of antigen contacts. However, there were a few exceptions and previous studies were limited to a small handful of cases. Here, we determined the structural flexibility of the CDR-H3 loop for thousands of recent homology models of the human peripheral blood cell antibody repertoire using rigidity theory. We found no clear delineation in the flexibility of naïve and antigen-experienced antibodies. To account for possible sources of error, we additionally analyzed hundreds of human and mouse antibodies in the Protein Data Bank through both rigidity theory and B-factor analysis. By both metrics, we observed only a slight decrease in the CDR-H3 loop flexibility when comparing affinity matured antibodies to naïve antibodies, and the decrease was not as drastic as previously reported. Further analysis, incorporating molecular dynamics simulations, revealed a spectrum of changes in flexibility. Our results suggest that rigidification may be just one of many biophysical mechanisms for increasing affinity.

Keywords: antibody repertoires, affinity maturation, complementarity determining regions, conformational flexibility, rigidity theory, pebble game algorithm, RosettaAntibody, molecular dynamics simulations

INTRODUCTION

Antibodies are proteins produced by the B cells of jawed vertebrates that play a central role in the adaptive immune system. They recognize a variety of pathogens and induce further immune response to protect the organism from external perturbation. Molecules that are bound by antibodies are referred to as antigen and are recognized by the antibody variable domain (Fv), which is comprised of a variable heavy (V_H) and light (V_L) domain. To overcome the challenge of recognizing a vast array of targets—the number of antigens being far greater than the number of antibody germline genes—antibodies rely on combinatoric and genetic mechanisms that increase sequence diversity (1–3). Starting from a limited array of germline genes, a naïve antibody is generated by productive pairing of a randomly recombined V_H, assembled from V-, D-, and J-genes on the heavy locus, and randomly recombined V_L, assembled from V- and J-genes on the kappa and lambda loci (1). Next, in a process known as affinity maturation, iterations of somatic hypermutation are followed by selection to evolve the antibody in specific response to a particular antigen. This evolution results in the gradual accumulation of mutations across the entire antibody, with higher mutation rates in the six complementarity determining regions (CDRs) than in the framework regions (FRs) (4, 5). The CDRs are hypervariable loops comprising a binding interface on the Fv domain beta-sandwich framework, with three loops contributed by each chain; the light chain CDRs are denoted as L1, L2, and L3 and the heavy chain CDRs are denoted as H1, H2, and H3. The five non-H3 CDRs can be readily classified into a discrete amount of canonical structures (6–10) because they possess limited diversity in both sequence and structure. The CDR-H3 on the other hand is the focal point of V(D)J recombination, resulting in exceptional diversity of both structure and sequence. While all CDRs contribute to antigen binding, the diverse CDR-H3 is often the most important CDR for antigen recognition (11–15). Thus, to understand the role of B cells in adaptive immunity and how they evolve antibodies capable of binding specific antigens, we must first understand the effects of affinity maturation on the CDRs, and in particular on the CDR-H3.

Over the past 20 years, the effects of affinity maturation have been studied with an assortment of experimental and computational methods. X-ray crystallography has been used to compare antigen-inexperienced (naïve) and antigen-experienced (mature) antibodies with both antigen present and absent. Analysis of the catalytic antibodies 48G7, AZ-28, 28B4, and 7G12 showed a 1.2 Å average increase in Cα root-mean-square deviation (RMSD) of the CDR-H3 upon antigen binding in the naïve over that of the mature antibody, whereas motion in the other CDRs varied (16–20). Beyond structural studies, surface plasmon resonance has been used to assess the energetics and association/dissociation rate constants of antibody–antigen binding. Manivel et al. studied a panel of 14 primary (naïve) and 11 secondary (mature) response anti-peptide antibodies, observing that affinity maturation resulted in increases in the association rate and corresponding changes in the entropy of binding (21). Schmidt et al. saw the opposite when studying a broadly neutralizing influenza virus antibody, observing that affinity maturation resulted primarily in a decrease in the dissociation rate, with little effect on the association rate (22). Isothermal calorimetry (ITC) has also been used to determine antigen-binding energetic, including the enthalpic and entropic contributions. For nine anti-fluorescein antibodies, including 4-4-20 and eight anti-MPTS antibodies, ITC results revealed diverse effects of affinity maturation: 14 of 17 mature antibodies bound antigen in an enthalpically favorable and entropically unfavorable manner, yet 3 of 17 showed the opposite, with entropically favorable and enthalpically unfavorable binding energetics (23, 24). Three-pulse photon echo peak shift (3PEPS) spectroscopy has been used to quantify dynamics of chromophore-bound antibodies on short timescales of femto- to nanoseconds. 3PEPS spectroscopy results from a panel of 18 antibodies showed that mature antibodies can possess a range of motions from small rearrangements such as side-chain motions to large rearrangements such as loop motions (23–25). In a specific comparison of naïve vs. mature, for the 4-4-20 antibody, the mature antibody was found to have smaller motions, i.e., to be more rigid, than naïve (23–28). Antibody dynamics have also been studied by hydrogen–deuterium exchange mass spectroscopy (HDX-MS), which in contrast to 3PEPS probes timescales of seconds to hours. Comparison of three naïve and mature anti-HIV antibodies showed changes in CDR-L2/H2, but not in CDR-H3 dynamics (29). Finally, molecular dynamics (MD) simulations have been used to study antibody dynamics on intermediate timescales of nano- to microseconds. MD simulations showed rigidification and reduction of CDR-H3 loop motion upon maturation for seven studied naïve/mature antibodies, with two exceptions, depending on the specific study (22, 28, 30–34). In an orthogonal protein design approach to examine the CDR-H3 loop flexibility, Babor et al. and Willis et al. found that naive antibody structures are more optimal for their sequences, when considering multiple CDR-H3 loop conformations (35, 36). In sum, past studies focusing on the effects of affinity maturation on CDRs have found evidence suggesting that mature antibodies have more structural rigidity and less conformational diversity than their naïve counterparts (16, 18, 19, 23–27).

With recent growth in the number of antibody structures deposited in the Protein Data Bank (PDB) and development of homology models from high-throughput sequencing of paired V_H–V_L genes in B cells, we now have the datasets necessary to test the rigidity hypothesis on a large scale. Prior studies, usually focused on a few antibodies at time, generally support the hypothesis that affinity maturation rigidifies the CDR-H3 loop. Thus, we hypothesize that this effect should be observable in a repertoire-scale study of thousands of antibodies. We first analyzed thousands of recently determined RosettaAntibody homology models of the most common antibody sequences found in the human peripheral blood cell repertoire (37). We estimated the structural flexibility of the CDR-H3 loop by applying graph theoretical techniques based on mathematical rigidity theory, namely the Floppy Inclusions and Rigid Substructure Topography (FIRST) and extensions of the Pebble Game (PG) algorithms to determine backbone degrees of freedom (DOFs). Surprisingly, we found no difference in the CDR-H3 loop flexibility of the naïve and mature antibody repertoires. We considered alternative explanations for our results, which were incongruent with past studies, by

expanding our analysis to a large set of antibody crystal structures, including several previously characterized antibodies, and extending our methods to include other measures of flexibility, such as B-factors and MD simulations. By all analysis methods, we found mixed results: some antibodies' CDR-H3 loops were more flexible after affinity maturation whereas others' became less flexible. In summary, we find that while affinity maturation can modulate antibody binding activity by reducing CDR-H3 structural flexibility, it does not necessarily do so.

MATERIALS AND METHODS

Immunomic Repertoire Modeling

Briefly, RosettaAntibody is an antibody modeling approach that assembles homologous structural regions into a rough model and then refines the model through gradient-based energy minimization, side-chain repacking, rigid-body docking, and *de novo* loop modeling of the CDR-H3. The approach is fully detailed in Ref. (38, 39). In a typical simulation, ~1,000 models are generated and the 10 lowest-energy models are retained. The immunomic repertoire we analyzed is from DeKosky et al. (37). In that study, models were generated for each of the ~1,000 most frequently occurring naïve and mature antibody sequences from two donors (a total of ~20,000 models representing the ~2,000 most frequent antibodies).

Structural Rigidity Determination

The flexibility or rigidity of the CDR-H3 loop backbone was determined by using several extensions of the PG algorithm (40–43), initially developed in Ref. (40), and method FIRST (44); we refer to here as FIRST-PG. This approach can determine flexible and rigid regions in a protein and quantify the internal conformational DOFs from a single protein conformational snapshot. FIRST generates a molecular constraint network (i.e., a graph) consisting of vertices (nodes) representing atoms and edges (interactions representing covalent bonds, hydrogen bonds, hydrophobic interactions, etc.). Each potential hydrogen bond is assigned with energy in kcal/mol which is dependent on donor-hydrogen acceptor geometry. FIRST is run with a selected hydrogen-bonding energy cutoff, where all bonds weaker than this cutoff are ignored in the network. On the resulting network, the well-developed mathematical and structural engineering concepts (45) of flexibility and rigidity of molecular frameworks and the PG algorithm are then used to identify rigid clusters, flexible regions, and overall available conformational DOFs. For a given antibody structure, DOFs for the protein backbone of the CDR-H3 loop were calculated at every hydrogen-bonding energy cutoff value between 0 and −7 kcal/mol in increment steps of 0.01 kcal/mol. This calculation was repeated for every member of that antibody ensemble (i.e., 10 lowest-energy models of the ensemble) and finally, at each energy cutoff, the DOF count was averaged over the entire ensemble.

For a given energy cutoff and a given member of the ensemble, the DOF count for the CDR-H3 loop (residues 95–102) was obtained using a special PG operation which calculates the maximum number of pebbles that can be gathered on the backbone atoms (Cα, C, N) of the CDR-H3 loop (40). The PG algorithm starts with the constrained molecular graph and generates a directed multigraph, where available free pebbles are absorbed one by one by independent edges (constraints). Each pebble represents one of six DOF associated with an atom. After PG completion, the remaining free pebbles can be collected on the CDR-H3 backbone (i.e., a subgraph in the constrained network) represent its conformational DOF count.

DOF Scaling

To compare flexibility across CDR-H3 loops of different lengths, the DOF metric computed above is scaled by a theoretical maximum DOF. We define $\mathrm{sDOF} = \frac{\mathrm{DOF}}{2L+6}$, where, $2L$ (the loop length in residues) represents the backbone DOFs (torsion angles: ϕ, ψ), and 6 represents the trivial, but ever-present rigid-body DOFs (i.e., combination of rotations and translations in 3D).

Area under Curve (AUC) Calculation

The AUC is approximated by simple numerical integral (akin to trapezoidal integration), where the first term defines a rectangle and the second term defines a triangle:

$$\mathrm{AUC} \equiv \sum (x_i - x_{i-1}) \cdot y_{i-1} + \frac{1}{2}(x_i - x_{i-1})(y_i - y_{i-1}).$$

Crystallographic Dataset

On June 27th, 2017, a summary file was generated from the Structural Antibody Database (SAbDab) (46), using the "non-redundant search" option to search for antibodies with maximum 99% sequence identity, paired heavy and light chains, and a resolution cutoff of 3.0 Å. The summary file, containing 1,021 antibodies, was used as input to a SAbDab download script which yielded corresponding sequences, Chothia-numbered PDBs, and IMGT data (on occasion this had to be updated to match the reported germline in the IMGT 3Dstructure-DB) (47). The structures were further pruned: structures were omitted if there were unresolved CDR-H3 residues, as this would preclude flexibility calculations, or if the antibody was neither human nor mouse, as this would prevent alignment to germline. Prior to analysis, structures were truncated to the Fv region (removing all residues, but light chain residues numbered 1–108 and heavy chain residues numbered 1–112, in Chothia numbering) and duplicate and non-antibody (for example, bound antigen) chains were removed. A total of 922 antibody crystal structures were analyzed. The following CDR definitions were used throughout this paper, in conjunction with the Chothia numbering scheme: L1 spans light chain residue numbers 24–34, L2 spans 50–56, L3 spans 89–97, H1 spans heavy chain residue numbers 26–35, H2 spans 50–56, and H3 spans 95–102.

Alignment to Germline

The germline of each antibody was determined by IMGT lookup (47). Then, BLASTP (version 2.2.29+) with the BLOSUM50 scoring matrix was used to align the antibody variable region heavy and light sequences to corresponding germline sequences (IGHV, IGKV, and IGLV loci only, downloaded from IMGT).

The number of mismatches according to BLAST was considered as the number of amino acid mutations from germline. Table S1 in Supplementary Material details the PDB ID, CDR-H3 length, number of heavy chain mutations, number of light chain mutations, heavy germline gene, and light germline gene data for each structure in the dataset.

B-Factor *z*-Score Calculation

Temperature factors (B-factors) were extracted for all Cα atoms in the variable region of the antibody heavy chain (V_H, Chothia numbering 1–112). The arithmetic mean and sample SD values were calculated for the B-factors. For each Cα atom in the CDR-H3 region, residue numbers spanning 95–102 under the Chothia numbering convention (11), the *z*-score was calculated as $\frac{(x-\mu)}{\sigma}$, where x is the B-factor of the current Cα atom and μ and σ are the mean and SD of B-factors for all Cα atoms in the V_H, respectively. PDB IDs 2NR6 and 3HAE were excluded from B-factor analysis because all reported B-factors were identical and so the *z*-scores were 0 by definition.

B-Factor *z*-Score Distribution Randomization Testing

To test whether two observed B-factor distributions arose from the same underlying distribution, we turned to randomization testing. First, we computed the difference of the observed distribution means. Next, we pooled the data from the two distributions (e.g., CDR-H3 loop B-factor *z*-scores) and randomly sampled the pooled data to create two simulated distributions (e.g., randomly assigning *z*-scores to either the naïve or mature category). Finally, we computed the simulated difference of the randomized distribution means. This process was repeated 10,000 times, so we could identify the fraction of random distributions with differences greater than the observed. Since, this process is stochastic and does not exhaustively sample all permutations of the data, it was further repeated 10 times to acquire a SD.

Rosetta Relaxation and Ensemble Generation

Antibody structural ensembles with 10 members were generated using either the Rosetta FastRelax (48, 49) or Rosetta KIC protocol (50), and Rosetta version 2017.26-dev59567 was used for all simulations (corresponding to weekly release version 2017.26). The Rosetta FastRelax protocol consists of five cycles of side-chain repacking and gradient-based energy minimization in the REF2015 version of the Rosetta energy function (51). Thus, FastRelax ensembles explore the local energy minimum of the crystal structure. KIC ensembles are more diverse and representative of RosettaAntibody homology models: each ensemble member was generated by running the CDR-H3 refinement step of the RosettaAntibody protocol, consisting of V_H–V_L docking, CDR-H3 loop remodeling, and all-CDR loop minimization (38, 39). Sample command lines are given in Supplementary Material. The structural ensembles produced by both FastRelax and KIC were used for rigidity analysis. For technical reasons, 6 targets could not be analyzed from the FastRelax ensemble, and 177 targets from the KIC ensemble were omitted due to non-trivial incompatibilities between the input structure numbering and Rosetta's internal antibody numbering scheme and a computing cluster time limitation. The excluded targets were randomly distributed and likely would not affect the conclusions.

MD Simulations

The Fv regions were retrieved from the original PDB files. The MD simulations were performed using the NAMD 2.12 package (52) with the CHARMM36m force field and the CMAP backbone energy correction (53). The truncated Fv structures were solvated with TIP3P water in a rectangular box such that the minimum distance to the edge of the box was 12 Å under periodic boundary conditions. Na or Cl ions were added to neutralize the protein charge, then further ions were added corresponding to a salt solution of concentration 0.14 M. The time step was set to two Fs throughout the simulations. A cutoff distance of 10 Å for Coulomb and van der Waals interactions was used. Long-range electrostatics was evaluated through the Particle Mesh Ewald method (54).

The initial structures were energy minimized by the conjugate gradient method (10,000 steps), and heated from 50 to 300 K during 100 ps, and the simulations were continued by 1 ns with NVT ensemble, where protein atoms were initially held fixed whereas non-protein atoms freely moved, gradually releasing the whole system to facilitate a stable simulation over the 1 ns. Further simulations were performed with NPT ensemble at 300 K for 200 ns without any restraints other than the SHAKE algorithm to constrain bonds involving hydrogen atoms. The last 180 ns of each trajectory were used for the subsequent clustering analyses. Similar to a previous work (55), a total of 2,000 evenly spaced frames from each trajectory were clustered based on RMSD of the Cα and Cβ atoms using the K-means clustering algorithm implemented in the KCLUST module in the MMTSB tool set (56). The cluster radius was adjusted to maintain 20 clusters in each trajectory. The structure closest to the center of each cluster was chosen as a representative structure of each cluster. The 10 representative structures were chosen from the top 10 largest clusters and these representative structures were energy minimized by the conjugate gradient method (10,000 steps) in a rectangular water box. The minimized antibody Fv structures were used as the inputs for the rigidity analysis.

Root-mean-square quantities of the MD trajectories were calculated based on the past 180 ns trajectories. After superposing Cα atoms of the FR of the heavy chain (FR_H) of each snapshot onto Cα atoms of FR_H of the reference structures (i.e., crystal structures), Cα-RMSD of the CDR-H3 loop was calculated as the time average. Similarly, after superposing Cα atoms of entire Fv domains of each snapshot onto those of the reference structures, the root-mean-square fluctuation (RMSF) of a residue i was defined as the time average:

$$\mathrm{RMSF}_i = \sqrt{\left\langle \left(x_i - \langle x_i \rangle \right)^2 \right\rangle}$$

where x_i is the distance between the Cα atom of the snapshots at a given time and the Cα atom of the ith residue of the reference structures (57).

RESULTS

Immunomic Repertoire Reveals No Difference in Flexibility between Naïve and Mature CDR-H3 Loops

We initially asked whether CDR-H3 loop rigidification, having been observed in many past studies, was present in a large set of antibodies derived from human peripheral blood cells. Previously, DeKosky et al. used RosettaAntibody to model the structures of 1,911 common antibodies found in the peripheral blood cells of two human donors (37). Paired V_H–V_L sequences were derived from either $CD3^-CD19^+CD20^+CD27^-$ naïve B cells or $CD3^-CD19^+CD20^+CD27^+$ antigen-experienced B cells (mature) isolated from peripheral mononuclear cells. RosettaAntibody structural models were created by identifying homologous templates for the CDRs, V_H–V_L orientation, and FRs; assembling the templates into one model; *de novo* modeling the CDR-H3 loop; rigid-body docking the V_H–V_L interface; side-chain packing; and minimizing in the Rosetta energy function (38). Since *de novo* modeling of long loops is challenging, DeKosky et al. limited their antibody set to the more tractable subset of antibodies with CDR-H3 loop lengths under 16 residues. They compared their models for seven human germline antibodies with solved crystal structures and found models had under 1.4 Å backbone RMSD for the FR and under 2.4 Å backbone RMSD for the CDR-H3 loop.

We used the FIRST-PG method (40, 44) to estimate flexibility from the RosettaAntibody homology models, determining the number of backbone DOFs for the CDR-H3 loop as each hydrogen bond is broken in order from weakest to strongest. FIRST models the antibody as a molecular graph where nodes represent atoms and edges represent atomic interactions. An extension of the PG algorithm uses this molecular graph to compute the DOFs of the CDR-H3 loop. To mitigate the effects of homology modeling, inaccuracies on the FIRST-PG analysis, we used an ensemble of 10 lowest-energy RosettaAntibody models. FIRST-PG analysis on structural ensembles has been shown to predict hydrogen–deuterium exchange and protein flexibility (51). To account for varying CDR-H3 loop lengths, we scaled the calculated DOFs by a theoretical maximum value (see Materials and Methods). **Figure 1A** shows a curve of the scaled DOFs averaged over all naïve or mature antibodies as a function of the hydrogen-bonding energy cutoff used in the FIRST-PG analysis. At a cutoff of 0 kcal/mol, all hydrogen bonds are intact and the average CDR-H3 loop-scaled DOFs are about 20% of the theoretical maximum. Moving from right to left on the plot increases the minimum energy cutoff for including interactions in the FIRST graph; effectively hydrogen bonds of increasing

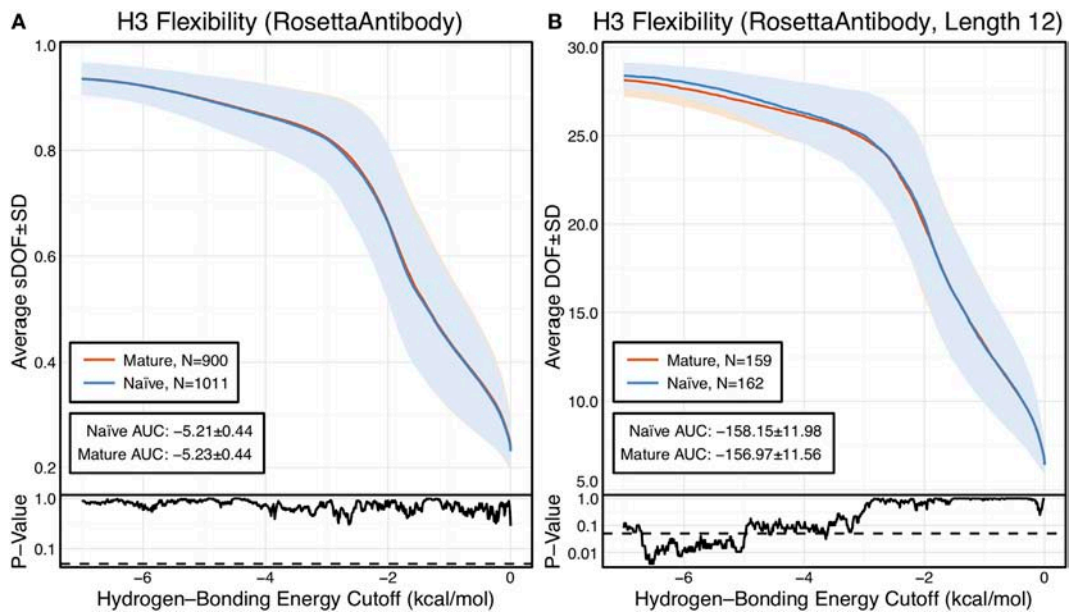


FIGURE 1 | CDR-H3 loop flexibility analysis of the immunomic antibody set reveals that no difference in naïve (blue) and mature (red) antibodies. Floppy inclusions and rigid substructure topography-Pebble Game was used to determine the degrees of freedoms (DOFs) as a function of hydrogen-bonding energy cutoff in RosettaAntibody models of the 1,911 most frequent public antibodies. Results were split, depending on whether the antibody was naïve or mature, as determined by B-cell surface receptors, and the mean DOFs were calculated along with the SD, shown in a lighter shade of the respective color. Subplots, below each main plot, show the *p*-value computed by a two-sample Kolmogorov–Smirnov (KS) test comparison of the naïve and mature DOFs distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same. A dashed line indicates a *p*-value of 0.05. **(A)** To permit comparison across loops of multiple lengths, the DOFs were scaled to a theoretical maximum for each length (a value of 1 indicates all DOFs are available, whereas a value of 0 indicates no DOFs are available). We found the scaled DOFs to be similar for both naïve and mature antibodies, quantified by the KS test *p*-values and area under the curve (AUC) ± SD: −5.21 ± 0.44 and −5.23 ± 0.44, respectively. **(B)** To exclude length effects on flexibility calculations, we compared DOFs for the most popular length (12 residues). We found the naïve AUC ± SD at −158.15 ± 11.98 and mature AUC ± SD at −156.97 ± 11.56 to be similar. The distributions appear similar at cutoffs between 0 and −5.0 kcal/mol, according to the KS test *p*-values.

strength are "broken" and the available DOFs rise from 20 to above 90% of the maximum theoretical flexibility, while the loop becomes unstructured (unfolded) in FIRST.

We compared the DOFs distributions for naïve and mature antibodies at every hydrogen-bonding energy cutoff by two-sample Kolmogorov–Smirnov (KS) testing, with null hypothesis being that the two distributions are identical (**Figure 1A**). There is no difference in the average, scaled DOFs. To further quantify this comparison, we computed the average AUC plus-or-minus one SD for both antibody sets. The average AUC values are identical for the naïve (-5.21 ± 0.44) and mature antibody repertoires (-5.23 ± 0.44). This lack of difference persists (AUC $-$ 158.15 $\pm$ 11.98 [naïve] vs. -156.97 ± 11.56 [mature]) when accounting for CDR-H3 loop length, by comparing loops of only length 12, the most popular length (**Figure 1B**), and so the observed similarity of DOFs in naïve and mature antibodies is not due to averaging over loops of different lengths. Thus, on the immunomic repertoire scale, we do not observe the difference in flexibility between naïve and mature antibodies predicted by the paratope rigidification hypothesis.

Before amending the rigidification hypothesis in light of these results, we considered several alternative explanations for our observations. First, we addressed whether the use of homology models for flexibility analysis introduced inaccuracies by analyzing a large set of antibody crystal structures and Rosetta-generated models from that set with varying quality, ranging from models with sub-angstrom backbone RSMD to models that may be several angstroms off (and more representative of an average homology model). Next, we addressed whether backbone DOFs, as calculated by FIRST-PG, were a good measure of flexibility, by assessing flexibility through two alternative measures: B-factors and MD simulations. Additionally, we addressed whether averaging flexibilities and comparing across many germlines affected results, by detailed flexibility analysis of previously studied naïve–mature antibody pairs and RosettaAntibody-modeled pairs.

Only Small Flexibility Differences Are Observed between Naïve and Mature Antibodies in the Crystallographic Set

Preparation of an Antibody Crystal Structure Dataset

Of course, the strongest critique of the immunomic antibody set is that these models are only approximating the actual antibody structure. Thus, we applied FIRST-PG analysis to a large set of antibody crystal structures. We curated the set of all non-redundant mouse and human antibody crystal structures from SAbDab (46). To be consistent with the models produced by RosettaAntibody, we truncated the structure of each antibody to only the Fv domain, excluding other antibody regions or antigen. Then, we used IMGT/3Dstructure-DB (58) to identify the variable domain genes and determined the number of somatic mutations by aligning the sequence derived from the crystal structure to the IMGT-determined V-gene. We defined mature antibodies as those possessing at least one somatic mutation in either V-gene. Our full dataset has 922 antibodies of which 23 are naïve. CDR-H3 loop lengths and germline assignments are summarized in Table S1 in Supplementary Material. Summary statistics are plotted in Figures S1–S3 in Supplementary Material.

FIRST-PG Analysis of Crystal Structures

From the crystal structures, we created two sets of structural ensembles and assessed flexibility by FIRST-PG. Flexibility analysis has previously been shown to be more accurate on ensembles in comparison to analysis using single (snapshot) conformers (41, 59). Ensembles of 10 representative structures were generated from the initial crystal structure by using either Rosetta FastRelax (48) or refinement step of RosettaAntibody (38, 39), which we term KIC ensembles after the loop modeling algorithm used in refinement (50). Rosetta FastRelax samples structures around the crystallographic, local energy minimum, with typically <1 Å backbone RMSD, whereas the refinement step of RosettaAntibody samples a more diverse set of low-energy CDR-H3 loop conformations and V_H–V_L orientations. Thus, FastRelax ensembles are representative of the crystal structures, whereas KIC ensembles are representative of RosettaAntibody homology models. By comparative FIRST-PG analysis of the two sets, we can assess the effects of modeling inaccuracies on flexibility analysis.

The scaled DOFs as calculated by FIRST-PG for FastRelax ensembles of antibody crystal structures are shown in **Figure 2A**. There are only minor differences between the naïve and mature flexibility curves, two-sample KS testing reveals insignificant p-values ($\rangle\rangle 0.05$) for all hydrogen-bonding energy cutoffs, and the AUC is similar for both sets (-4.70 ± 0.46 [naïve] vs. -4.70 ± 0.48 [mature]). Again, we considered the possibility that different distributions of loop lengths in the two sets obscures the affinity maturation contributions to flexibility. Therefore, we analyzed loops of length 10 (**Figure 2B**), the single most common length in the crystallographic set. When loops of a single length were compared, there was a separation between the naïve and mature sets, with the naïve antibody set average DOFs being consistently greater than the mature set, but not significantly so, except for some energy cutoffs below -5 kcal/mol, according to KS testing. As expected, the AUC values differ, but are within a SD (-128.2 ± 9.0 [naïve] vs. -121.9 ± 10.1 [mature]). We repeated FIRST-PG analysis for KIC ensembles of antibody crystal structures and observed similar results (Figure S4 in Supplementary Material): for scaled DOFs, the AUC was -5.91 ± 0.20 (naïve) vs. -5.81 ± 0.26 (mature) and, for loops of length 10 only, the AUC was -154.10 ± 4.80 (naïve) vs. -150.44 ± 7.73 (mature). Thus, there does not appear to be a large, consistent CDR-H3 loop flexibility difference across all antibody crystal structures analyzed.

B-Factor Analysis of Crystal Structures

However, we have not accounted for the possibility that backbone DOFs as calculated by FIRST-PG may not capture the effects of affinity maturation on CDR-H3 loop flexibility. Thus, we assessed loop flexibility as determined by atomic temperature factors or B-factors. In protein crystal structures, B-factors measure the heterogeneity of atoms in the crystal lattice. Thus, rigid regions have lower B-factors as they are more homogenous throughout the crystal, whereas flexible regions have higher B-factors as they

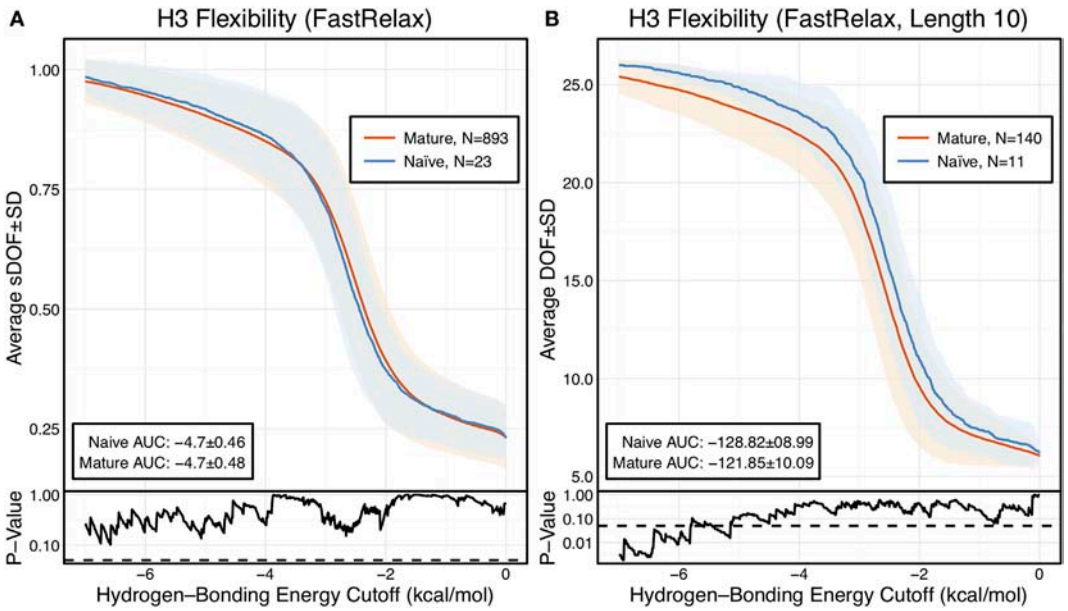


FIGURE 2 | When accounting for length, CDR-H3 loop flexibility analysis of the crystallographic antibody set reveals naïve (blue) antibodies to be slightly more flexible than mature (red). Floppy Inclusions and Rigid Substructure Topography-Pebble Game was used to determine the degrees of freedoms (DOFs) as a function of hydrogen-bonding energy cutoffs in crystal structure ensembles created by Rosetta FastRelax. Results were split, depending on whether the antibody was naïve or mature, as determined by BLAST alignment to its germline V-genes, and the mean DOFs were calculated along with the SD, shown in a lighter shade of the respective color. Subplots, below each main plot, show the *p*-value computed by a Kolmogorov–Smirnov (KS) test comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same. A dashed line indicates a *p*-value of 0.05. **(A)** To permit comparison across loops of multiple lengths, the DOFs were scaled to a theoretical maximum for each length (a value of 1 indicates all DOFs are available whereas a value of 0 indicates not DOFs are available). We found the scaled DOFs to be similar for both naïve and mature antibodies, quantified by KS test *p*-values and the areas under the curve (AUCs) ± SD: −4.70 ± 0.46 and −4.70 ± 0.48, respectively. **(B)** To exclude length effects on flexibility calculations, we compared DOFs for the most popular length (10 residues). We found the naïve AUC ± SD at −128.82 ± 8.99 was greater than the mature AUC ± SD at −121.85 ± 10.09, but still within a SD. The distributions appear similar at cutoffs between 0 and −6.0 kcal/mol, according to the KS test *p*-values.

are less homogenous throughout the crystal. B-factors are also affected by crystal resolution, so we cannot compare raw values across structures of varying resolution. Instead, we computed a normalized B-factor *z*-score, which has 0 mean and unit SD for each antibody chain. Finally, to account for different CDR-H3 loop lengths, we averaged the B-factor *z*-scores for the CDR-H3 loop residues.

Figure 3A shows the distributions of B-factor *z*-scores averaged over the CDR-H3 loop residues of naïve and mature antibodies. Both distributions span a similar range and overlap significantly, with the naïve curve peak shifted toward higher values than the mature. The majority of the naïve CDR-H3 loop B-factor *z*-score averages were positive (65%), whereas the majority of the mature CDR-H3 loop B-factor *z*-score averages were negative (64%). To address the question whether these distributions arose from the same underlying distribution we turned to randomization testing, as described in Section "Materials and Methods." The observed difference in distribution means is matched by only 0.066 ± 0.026% of simulated differences (**Figure 3B**), indicating that naïve and mature distributions are likely distinct. Furthermore, a two-sample KS test confirms the distributions to be distinct, with a maximum vertical deviation, *D*, of 0.36 and a *p*-value of 0.006.

However, we were concerned that the mixing of bound and unbound crystal structures would influence results, as we previously observed bound structures to have lower average B-factors (60). Furthermore, in the PDB-derived dataset, naïve antibodies were mostly crystallized in the unbound state (19 of 23), whereas mature antibodies were mostly co-crystallized with their cognate antigen (544 of 899). In conjunction, these two observations suggested that the high number of antigen-bound mature antibody crystal structures was the primary driver of the difference between naïve and mature B-factor *z*-scores. Thus, we compared the B-factor averages of unbound structures only and found that while the distributions appear to be distinct (**Figure 4A**), when the difference in distribution means is compared to a randomized set, 3.4 ± 0.2% of random differences are greater than or equal to the observed differences, and the distributions fail a two-sample KS test ($D = 0.27$, $p = 0.15$). Thus, the difference between naïve and mature antigen-free crystal structures does not appear significant.

As we conjectured, a significant difference was found between the bound and unbound distributions (**Figure 5**), with a two-sample KS test confirming the difference between the distributions ($D = 0.31$, $p < 2.16\text{E}{-16}$) and randomized testing never showing a difference in means as large as the observed difference. Additionally, we considered other possible origins of difference between the naïve and mature distributions that are not related to affinity maturation, including comparison across species, crystal structure resolutions, CDR-H3 loop lengths, and whether the

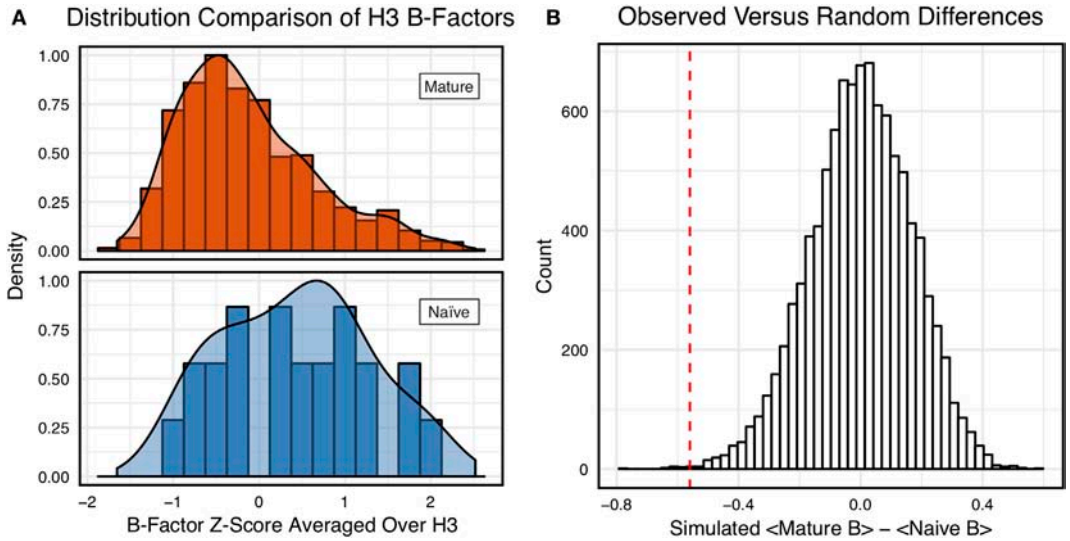


FIGURE 3 | Comparison of the distribution of average CDR-H3 loop B-factor *z*-scores in antibody crystal structures suggests that naïve are more flexible than mature. **(A)** Distributions of average CDR-H3 loop B-factors for the crystallographic set of antibodies are distinct for the mature (orange) and naïve (blue) sets. The mature antibody CDR-H3 loops have lower B-factors than the naïve, corresponding to more rigidity. Bars show binned counts in intervals of 0.25. Both the bars and smoothed densities are normalized so the maximum value is 1. A two-sample Kolmogorov–Smirnov test confirms different underlying distributions with a *p*-value of 0.006 and maximum vertical deviation, *D*, of 0.36. **(B)** The observed difference in distribution means is difficult to replicate by random chance, occurring only 6.6 ± 2.6 times out of 10,000 simulations. Comparing the observed difference in means (red line, dashed) to simulated differences (white bars) acquired by randomly assigning B-factor values from the original distributions to either a naïve or mature set, in the observed numbers (N_{mature} = 897 and N_{naive} = 23), before computing the difference in means.

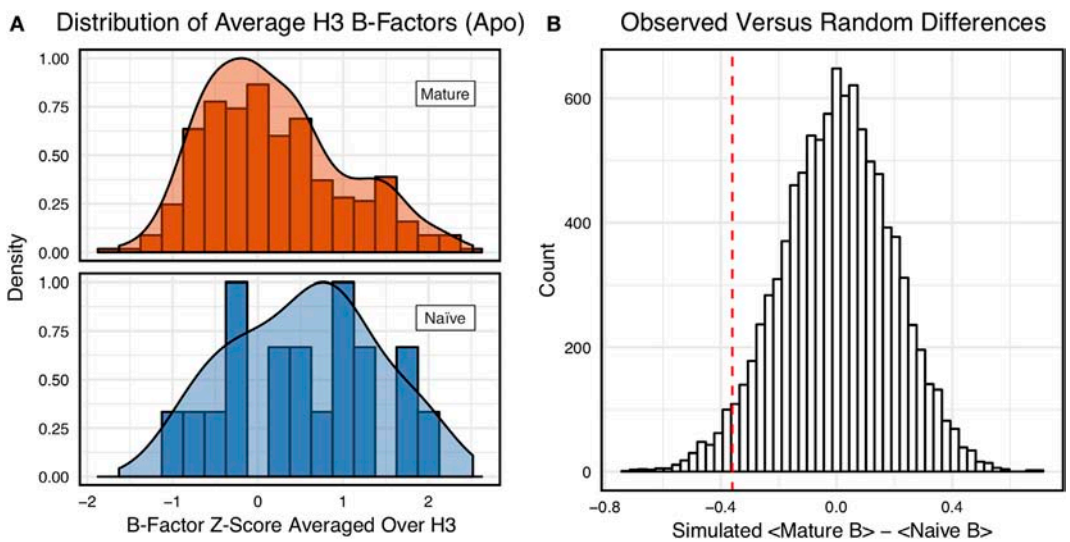


FIGURE 4 | When considering only antigen-free crystal structures (to control for rigidification upon antigen biding), the difference between naïve and mature average CDR-H3 loop B-factor *z*-score distributions is small. **(A)** The distributions of CDR-H3 loop average B-factors are less distinct between the mature (orange) and naïve (blue) sets. Bars show binned counts in intervals of 0.25. Both the bars and smoothed densities are normalized so the maximum value is 1. A two-sample Kolmogorov–Smirnov test results in a *p*-value of 0.15 and *D* of 0.27, indicating that the null hypothesis of indistinguishable underlying distributions cannot be discarded. **(B)** The observed difference in distribution means (red line, dashed) is occasionally replicated in random resampling (white bars). When average CDR-H3 loop B-factor *z*-scores are pooled and randomly assigned to either a naïve or mature set, in the observed numbers (N_{mature} = 355 and N_{naive} = 18), the observed difference in means is matched or surpassed in 340 ± 20 out of 10,000 simulated differences.

CDR-H3 loop was at a crystal contact or not. We found none of these to have as clear of an effect on the distribution of B-factor averages as whether or not antigen was bound (Figures S5 and S6 in Supplementary Material). In summary, the distributions of B-factor *z*-score averages (**Figures 3**–**5**) suggest that both the naïve and mature antibody sets possess CDR-H3 loops of varying

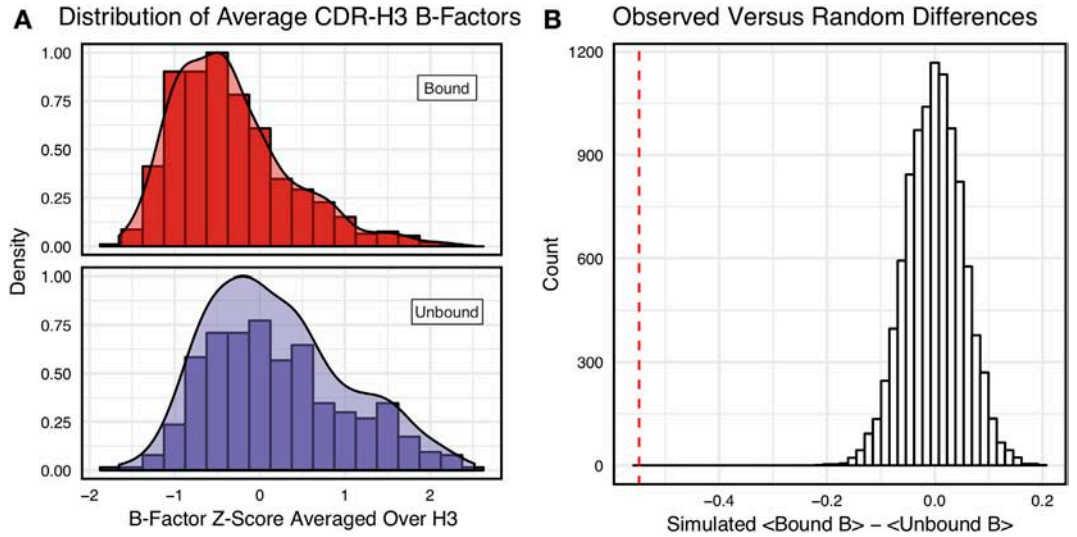


FIGURE 5 | Antigen-bound and antigen-free distributions of B-factor *z*-scores are distinct. **(A)** Distributions of CDR-H3 loop average B-factors for the crystallographic set of antibodies are distinct for the antigen-bound (red) and antigen-free (purple) sets. Bound antibody CDR-H3 loops have lower B-factors than unbound, corresponding to more rigidity. Bars show binned counts in intervals of 0.25. Both the bars and smoothed densities are normalized so the maximum value is 1. Distributions appear distinct according to a two-sample Kolmogorov–Smirnov test with a *p*-value of 2.2E-16 and *D* of 0.31. **(B)** The observed difference in distribution means (red line, dashed) is never replicated in 10,000 attempts at random resampling (white bars). Simulated differences were acquired by randomly assigning values from both sets to either a naïve or mature set, in the observed numbers (N_{bound} = 546 and N_{naive} = 374), before computing the difference means.

flexibility and that neither set is significantly more flexible or rigid than the other.

Comparison of Mature to Naïve-Reverted Models Reveals Varying Rigidification across Matched Pairs

Having not observed consistent rigidification of the CDR-H3 loop in two large sets of antibodies, we postulated that rigidification was not a repertoire-wide phenomenon (i.e., all mature antibodies are not more rigid than all naïve antibodies), but it could still be plausible that matched pairs of naïve and mature antibodies would reveal rigidification.

To investigate this hypothesis, we selected 10 mature antibodies from our SAbDab set with CDR-H3 loops of length 10, a length for which loop modeling performs well (50, 61). We identified antibodies that had at least 5 (~97% sequence identity), but no more than 25 (~85% sequence identity), mutations when compared to the germline V-genes. To control for species, half of the selected antibodies were human and half were mouse. We reverted the mature antibody sequences to naïve using the germline sequences from the aligned V-genes, as described in the methods, and using germline J-genes from sequence alignments from IMGT/DomainGapAlign (47). The reverted sequences are reported in Section "Sequences Used to Model Naïve-Reverted Antibodies" in Supplemental Material. We then used RosettaAntibody to generate homology models for the naïve-reverted sequences. We analyzed the ensembles of the 10 lowest-energy homology models using FIRST-PG. To ensure fair comparison, we also used FIRST-PG to analyze homology model ensembles of the mature sequences. To provide an estimate for the accuracy of RosettaAntibody homology models, we computed RMSDs for the mature models using the known crystal structures and found all had sub-2-Å CDR-H3 loop backbone RMSD, calculated after alignment of the heavy chain FR, with 7 of 10 antibodies having sub-Å RMSD (Figures S7–S9 in Supplementary Material).

Of the 10 naïve/mature antibody pairs we analyzed, 6 showed a decrease in flexibility and 4 showed an increase in flexibility upon affinity maturation (**Figure 6**). These 10 antibodies demonstrated the breadth of possible affinity maturation effects, from an expected flexibility decrease in antibody 2AGJ, with AUC decreasing by 9.34%, to the unexpected flexibility increase in antibody 1RZ7, with AUC increasing by 10.65%.

Analysis of 48G7 Antibody

Having analyzed 1,911 models, 922 crystal structures, and 10 paired-reverted models, we had yet to observe a consistent difference in CDR-H3 loop flexibility between naïve and mature antibodies, as previously reported in literature. Thus, we turned to three previously studied antibodies with known crystal structures and measured CDR-H3 loop flexibility. These are (1) the esterolytic antibody 48G7 (16, 32, 33, 35), (2) the anti-fluorescein antibody 4-4-20 (23, 26–28, 31, 33), and (3) a broadly neutralizing influenza virus antibody (22). For all three antibodies, the effects of affinity maturation on CDR-H3 loop flexibility have been previously studied by both experiment and simulation, allowing comparison with our results. For brevity, we presently discuss the 48G7 antibody here and full results for all antibodies are available in the Supplementary Material.

The 48G7 antibody was first studied through crystallography, with structures capturing the bound (holo) and unbound (apo) states of both the naïve and mature antibody (16). Comparison

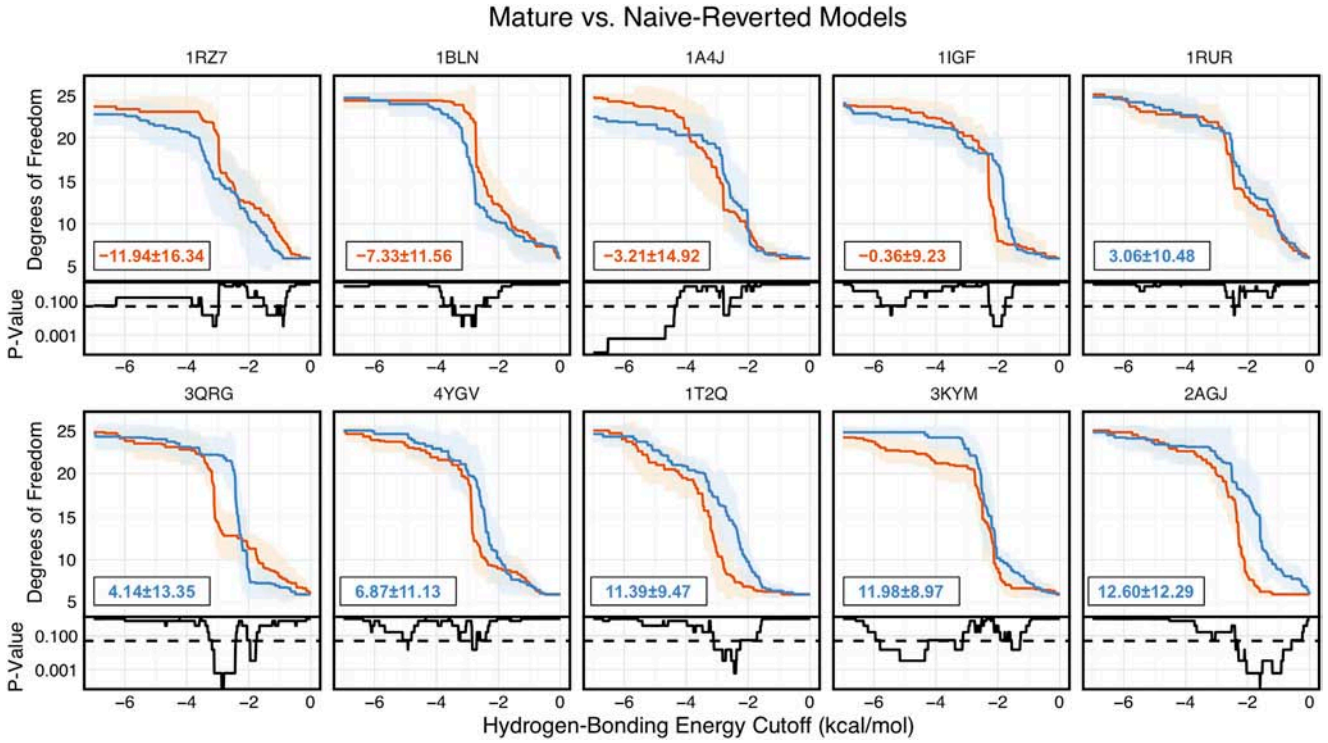


FIGURE 6 | Floppy Inclusions and Rigid Substructure Topography-Pebble Game analysis of 10 RosettaAntibody-modeled mature/naïve-reverted antibody pairs (CDR-H3 loop length of 10 residues) shows that affinity maturation does not always result in CDR-H3 loop rigidification. Naïve values are colored blue, while mature values are color red. The difference between mature and naïve area under the curves (AUCs) is reported in the bottom left of each sub-figure, with a positive value indicate a more flexible naïve antibody. 4 out of the 10 cases have mature antibodies with AUC greater than their naïve counterparts. Subplots, below each main plot, show the *p*-value computed by a Kolmogorov–Smirnov test comparison of the naïve and mature degrees of freedom distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same and a dashed line indicating a *p*-value of 0.05.

between the naïve and mature CDR loop motions from the free to the bound state revealed minor changes, with the mature CDR-H3 loop being slightly more rigid and moving an Angstrom less than the naïve upon antigen binding (Figures S10 and S11 in Supplementary Material). For each of the four crystal structures, we extracted B-factors and computed B-factor *z*-scores for the CDR-H3 loop, measuring the distance from the B-factor mean in SDs. B-factor *z*-scores for the CDR-H3 loop of apo-48G7 are shown in **Figure 7A**. The mature antibody has lower B-factors than the naïve antibody throughout the entire CDR-H3 loop. This observation also holds for the holo-48G7 antibody structures as well (Figure S12 in Supplementary Material). Table S2 in Supplementary Material summarizes B-factors, averaged over the whole CDR-H3 loop. These B-factor results agree with the prior crystallographic observations.

Follow-up studies on 48G7 used MD simulations to assess flexibility. Briefly, 500 ps short MD simulations of the naïve and mature antibodies in the presence of antigen with an explicit solvent model (TIP3P) found the CDR-H3 loop to be more flexible in the naïve than in the mature antibody by comparison of RMSFs (30), but 15 ns MD simulations of the naïve and mature antibodies in the absence of antigen with an implicit solvent model (GB/SA) found no difference between the two, again by comparison of RMSFs (32). Another study based on an elastic network model also suggested that, in the absence of antigen, the fluctuations of the naïve and mature 48G7 were similar, but their binding mechanisms could differ depending on response to antigen binding; the naïve antibody shows a discrete conformational change induced by antigen, whereas the mature antibody shows lock-and-key binding (62). Due to the contentious nature of these results, we ran 200 ns MD simulations for the 48G7 naïve and mature antibodies in the absence of antigen with an explicit solvent model (TIP3P). We measured both RMSDs and RMSFs for the Cα atoms along the CDR-H3 loop and computed the difference between the naïve and mature antibodies (Table S2 in Supplementary Material). **Figure 7B** shows that the CDR-H3 loop RMSFs are consistently greater for the mature than the naïve 48G7 antibody.

Finally, as we have done through this study, we used FIRST-PG to measure CDR-H3 loop flexibility. To limit the effects of crystal structure artifacts on FIRST-PG analysis, we used an ensemble of 10 representative structures, derived by clustering trajectory frames and selecting 10 structurally distinct cluster medians from the MD simulations, similar to a previous flexibility study for this antibody (33). The CDR-H3 loop flexibility of apo-48G7, as determined by FIRST-PG analysis of MD ensembles is shown in **Figure 8**. The FIRST-PG analysis showed no significant difference between the mature and naïve antibodies.

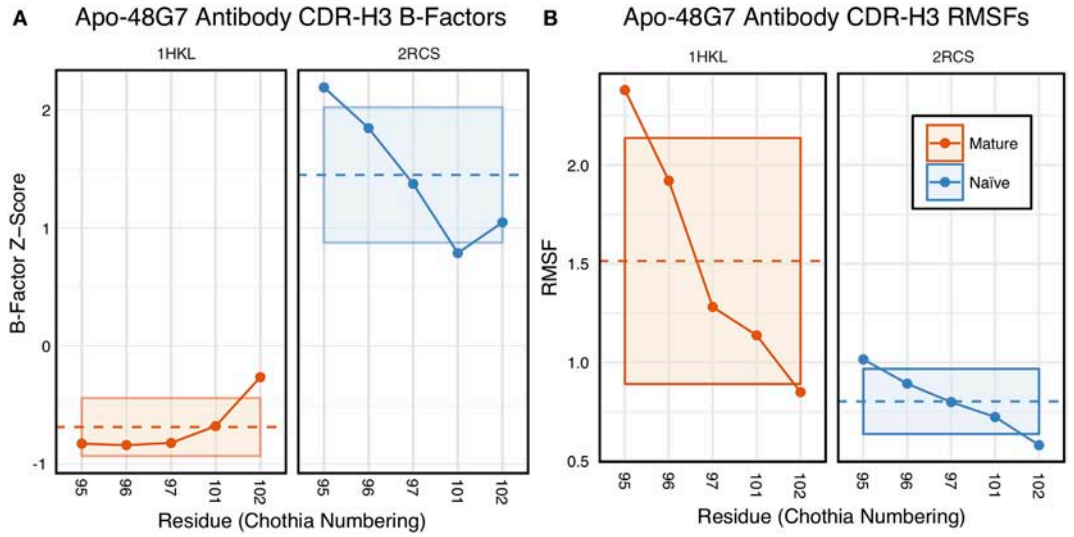


FIGURE 7 | Analysis of catalytic antibody 48G7 by CDR-H3 loop B-factors and root-mean-square fluctuations (RMSFs) shows conflicting results. **(A)** Comparison of normalized B-factor values for the CDR-H3 loop of the 48G7 antibody in crystal structures of the unbound naïve (dark blue) and mature (dark orange) antibodies reveals a more rigidity in the mature antibody. The dashed line indicates the average value and is outlined by a box defined by the average plus-or-minus the SD. **(B)** Comparison of CDR-H3 loop RMSFs for the molecular dynamics simulations of the naïve and mature 48G7 antibodies shows the opposite.

In addition to using MD simulations to generate ensembles, we used ensembles generated by RosettaAntibody and Rosetta FastRelax, permitting direct comparison. The CDR-H3 loop flexibility of apo-48G7, determined by FIRST-PG analysis of FastRelax and RosettaAntibody ensembles, is shown in **Figure 8**. The curves from FastRelax and the MD simulation are similar for low-energy cutoffs (e.g., in the range of 0.0 to −3.0 kcal/mol), with the naïve and mature DOFs being the same. These curves diverge at higher energy cutoffs, where the FastRelax curve shows a more flexible naïve antibody and the MD curve does not. The curve from RosettaAntibody ensembles differs from the two and shows a more flexible mature antibody at low-energy cutoffs and a more flexible naïve at high-energy cutoffs. For less visual and more quantitative comparisons, we computed the AUC of the DOF vs. hydrogen-bonding energy cutoff plots (Table S2 in Supplementary Material). We find the AUC is only slightly greater for naïve than mature antibodies in the FastRelax and RosettaAntibody ensembles, with the naïve AUC reducing by only 3.9 and 0.2%, respectively, upon maturation. MD ensembles show the opposite outcome, with the mature antibody having 1.3% greater AUC than the naïve.

Further validation was carried out on two other previously studied antibodies and reported in the Table S2 and Figures S12 and S13 in Supplementary Material. For the 4-4-20 antibody, antigen-bound structures were compared and the average mature B-factors were within a SD of the naïve. For the influenza antibody, average B-factors were compared between an unbound naïve and a bound mature crystal structure, showing significant rigidification. However, results are conflated due to the lack of unbound crystal structures, as in bound structures antibody–antigen contacts artificially increase rigidity of the CDR-H3 loop. In contrast to B-factor analyses, FIRST-PG analyses yielded mixed results for these two antibodies. The 4-4-20 antibody was found to become more flexible upon maturation by FIRST-PG analysis of all, but Rosetta KIC ensembles. The influenza antibody was found to become more rigid upon mature by FIRST-PG analysis of all, but Rosetta FastRelax ensembles. Finally, we analyzed RMSDs and RMSFs from MD simulations and found that the mature 4-4-20 antibody has higher CDR-H3 loop RMSD, but lower RMSF, values than the naïve while the mature influenza antibody was found to have lower values for both (Table S2 in Supplementary Material). As with our repertoire analysis, we do not see consistent rigidification in previously studied antibodies. We consider the significance of this result and compare our analysis in detail to past analyses of flexibility in Section "Discussion."

DISCUSSION

The Varying Effects of Affinity Maturation on CDR-H3 Flexibility

Affinity maturation, through a series of somatic hypermutation events and selection processes, can evolve a low-affinity, naïve antibody to bind an antigen with both high affinity and specificity (63). Elucidating the affinity maturation process is desirable to understand molecular evolution, develop antibody engineering methods, and guide vaccine development (64). Past studies have suggested that, with few exceptions (29, 65, 66), naïve antibodies are highly flexible and maturation leads to improved affinity and specificity through the optimization and rigidification of the antibody paratope, and in particular the CDR-H3 loop (22, 27, 28, 31–33). However, these studies have been limited, often focusing on a single antibody and assessing flexibility indirectly. We sought to test the generalizability of the rigidification-upon-maturation hypothesis. We were enabled by the large number of antibody structures in the PDB, homology models generated from

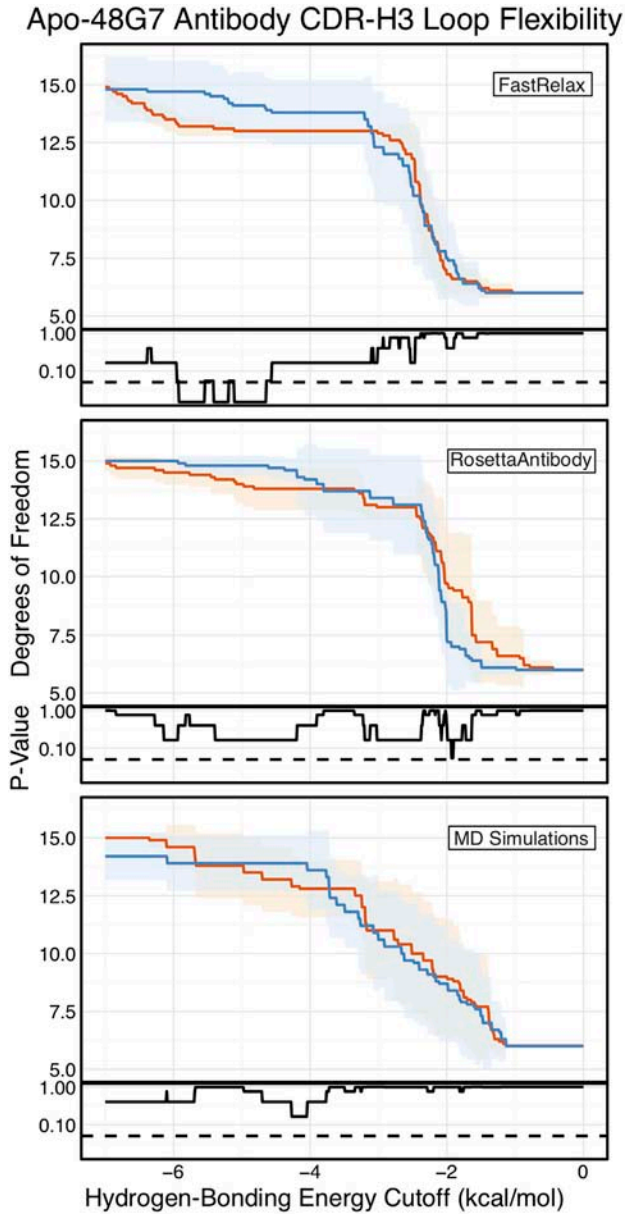


FIGURE 8 | Floppy inclusions and rigid substructure topography-Pebble Game (FIRST-PG) analysis of naïve (dark blue) and mature (dark orange) 48G7 antibodies using either Rosetta FastRelax-, RosettaAntibody-, or molecular dynamics (MD)-generated 10-member ensembles does not show a difference between the naïve and mature antibodies. FIRST-PG analysis calculates the degrees of freedom (DOFs) of CDR-H3 loop as a function of hydrogen-bonding energy cutoff. Subplots, below each main plot, show the *p*-value computed by a Kolmogorov–Smirnov test comparison of the naïve and mature DOF distributions for each hydrogen-bonding energy cutoff, with null hypothesis being that the distributions are the same and a dashed line indicating a *p*-value of 0.05. While the FastRelax ensembles appear distinct in the range of −6 to −3 kcal/mol, the naïve and mature are indistinguishable for both the RosettaAntibody and MD ensembles.

high-throughput repertoire sequencing data, and the FIRST-PG method for rapid structural flexibility calculation to ask whether affinity maturation leads to CDR-H3 loop rigidification.

Unexpectedly, in a comparison of flexibility of repertoires, our data show little difference between naïve and mature antibodies: FIRST-PG calculations showed no difference for RosettaAntibody homology model ensembles of the most common naïve and mature antibodies in human peripheral blood cells. The same calculations showed no difference in CDR-H3 loop DOFs of crystal structures under two different refinement schemes (FastRelax and KIC). After accounting for the presence/absence of antigen, CDR-H3 loop B-factor distributions were similar for both mature and naïve antibody crystal structures. These results indicate that rigidification of the CDR-H3 loop does not always occur upon affinity maturation.

Since our observations did not indicate clear rigidification over two sets of antibodies, we considered the following possibilities: (1) comparison of different length CDR-H3 loops was unfair because longer loops are inherently more flexible, (2) comparison of different antibodies was unfair because different combinations of gene segments and V_H–V_L pairs will result in different flexibilities, (3) mutations within CDR-H3 loop, which we could not identify for the PDB set because of the difficulty in D/J-gene alignments, may have modulated flexibilities of CDR-H3, (4) inaccuracies in the computational methods could preclude observation of rigidification, and (5) FIRST-PG-measured backbone DOFs are not a good measure of flexibility. To address the first concern, we analyzed loops of consistent length *via* B-factor and FIRST-PG (**Figures 1B** and **2B**; Figures S4 and S5 in Supplementary Material). We found that, according to KS testing and when accounting for the presence/absence of antigen, B-factor distributions were not distinct for naïve and mature sets of antibodies with same length CDR-H3 loops (length 10 for the crystallographic set and 12 for the repertoire model set). We also found that FIRST-PG DOFs AUC values of the naïve and mature sets of antibodies with the same length CDR-H3 loops were within a SD for RosettaAntibody, FastRelax, and KIC ensembles. So, even when accounting for length, mature antibodies are not significantly more rigid than naïve ones.

To address the concern that comparison of sets of antibodies originating from different V_H and V_L genes is unfair, we analyzed mature/naïve antibody pairs that had been previously studied and mature/naïve-reverted pairs that we generated with RosettaAntibody and analyzed by FIRST-PG (**Figures 6–8**; Table S2 in Supplementary Material). We found that CDR-H3 loop B-factors did not always indicate rigidification upon maturation and for the 7G12 antibody we observed the reverse effect (Figure S14 in Supplementary Material). We also found that mature antibodies did not always become more flexible upon naïve reversion, but instead displayed a breadth of behaviors (**Figure 6**). So, when analyzing matched naïve/mature pairs, we do not see consistent rigidification upon maturation.

Our analysis of previously studied naïve/mature antibody pairs coupled with the earlier repertoire analysis should alleviate concerns that our flexibility results for the PDB set were strongly affected by our inability to align D/J-gene segments, and thus consider mutations in the CDR-H3 loop. The previously studied pairs included CDR-H3 mutations and the repertoire set had antibody sequences determined by Illumina MiSeq sequencing with naïve/mature status assigned by the absence/presence of the

CD27 cell-surface receptor. In both cases, the naïve and mature sequences were determined through the entire Fv, and flexibility analysis still revealed mixed results.

Finally, to address the concern that RosettaAntibody models may not be accurate enough to be useful for FIRST-PG calculations, we tested FIRST-PG on a range of structural ensembles with varying deviation from the crystal structure. We found no difference in the naïve vs. mature antibody CDR-H3 loop AUC of the FIRST-PG results, regardless of the ensemble generation method used (compare **Figure 2**; Figure S4 in Supplementary Material). We also determined flexibility through alternative measures, such as crystal structure B-factors and RMSFs in MD simulations. For both, affinity maturation was not found to have a consistent, rigidifying effect. Thus, even if model inaccuracies confound analysis, other data support the same hypothesis.

Comparison with Prior Results

Our analysis included several antibodies that have been the subject of previous flexibility studies, permitting a direct comparison (Table S4 in Supplementary Material summarizes past studies). One of the most studied antibodies is the anti-fluorescein antibody, 4-4-20. Spectroscopic experiments measuring the response of a fluorescent probe (fluorescein) and MD simulations measuring Cα atom fluctuations suggested that somatic mutations restrict conformational fluctuations in the mature antibody (26, 28, 31). Our analysis of 4-4-20 was not as clear: we observed no significant difference in naïve vs. mature CDR-H3 loop crystallographic B-factors (Figure S12 in Supplementary Material) and found the mature antibody to be more rigid in FIRST-PG calculations only in the −2.0 to −0.0 kcal/mol range of hydrogen-bonding energy cutoffs (Figure S13 in Supplementary Material). Similar mixed results were observed by Li et al. (33) who used a Distance Constraint Model (DCM) to analyze flexibility in an ensemble of 4-4-20 conformations drawn from MD simulations. They found increases in structural rigidity of the CDR-H3 loop, as determined by the DCM, occurred upon affinity maturation, but these increases did not correspond to decreases in dynamic conformational fluctuations, as determined by RMSFs from MD simulations. Further studies artificially matured 4-4-20 by directed evolution, resulting in a femtomolar-affinity antibody, 4M5.3 (67), but the crystal structures of 4M5.3 and 4-4-20 were almost identical (the reported backbone RMSD is 0.60 Å) and thermodynamic measurements suggested that the affinity improvement was achieved primarily through the enthalpic interactions with subtle conformational changes (68). This observation was contradicted by Fukunishi et al. (69), who performed steered MD simulations to analyze the effects of the mutations on the flexibility of 4-4-20 and 4M5.3. By applying external pulling forces between the antibodies and the antigen along a reaction coordinate, they quantified the interactions and showed that, during the simulations, fluctuations of the antibody, especially the CDR-H3 loop, and of the antigen were indeed larger in 4-4-20 than in the more matured antibody, 4M5.3 (69). Thus, there is some variation not only in our results, but also in the literature as to the effects of affinity maturation on 4-4-20.

Another set of well-studied antibodies are the four catalytic antibodies: 48G7, 7G12, 28B4, and AZ-28. In fact, the first crystallography studies to suggest rigidification of the CDR-H3 loop as a consequence of affinity maturation were performed on 48G7. Wedemayer et al. observed larger structural rearrangements upon antigen binding in the CDR-H3 loop for the naïve antibody than the mature antibody (Figures S10 and S11 in Supplementary Material) (16). Crystallization of the naïve unbound, naïve bound, mature unbound, and mature bound states for 7G12, 28B4, and AZ-28 revealed similar results (18, 19). Additionally, MD simulations of the four catalytic antibodies in implicit solvent were used to calculate CDR Cα atom B-factors (32). Wong et al. showed a decrease in mature CDR-H3 loop B-factors in three cases (7G12, 28B4, and AZ-28), whereas no significant difference was observed for 48G7 (see Figure 2 in Wong et al.). Furthermore, for 48G7, Li et al. used MD simulation to generate structural ensembles and DCM analysis to determine flexibility. They found that the mature CDR-H3 loop is more rigid than the naïve, according to DCM, but used an unusual loop definition that included five additional flanking residues (see Figure 1 in Li et al.), making comparison challenging (longer loops will be inherently more flexible), and they observed increases in the mature CDR-H3 loop RMSFs (see Figure 8 in Li et al.) (33). Our analysis of CDR-H3 loop B-factors showed rigidification upon maturation for some of the 48G7 and 28B4 crystal structures (**Figure 7**; Figure S14 in Supplementary Material), but not for 7G12 and AZ-28 structures (Figures S14 and S15 in Supplementary Material). FIRST-PG analysis of FastRelax, RosettaAntibody, and MD ensembles for 48G7 showed slight to no rigidification (**Figure 8**). Additionally, RMSFs from MD simulations for 48G7 showed higher values for the mature loop, contrary to the expectation that it is more rigid. Our mixed results for the effects of affinity maturation on 48G7 are consistent with literature, but there is variation between our results and the literature as to the effects of affinity maturation on the other catalytic antibodies.

Finally, Schmidt et al. used X-ray crystallography, MD simulations, and thermodynamics measurements to investigate how somatic mutations affected the binding mechanism of anti-influenza antibodies (22). They identified three mature antibodies, their unmutated common ancestor (UCA), and a common intermediate, all derived from a subject immunized with an influenza vaccine. The affinities of the mature antibodies were about 200-fold better than the UCA. MD simulations of the UCA and the mature antibodies showed that CDR-H3 loop of the UCA could sample more diverse conformations than the mature antibodies, whose CDR-H3 loop sampled only conformations optimal for antigen binding, supporting the hypothesis that somatic mutations rigidify antibody structures. In another study by the same group (70), further MD simulations were performed on the same systems, showing that, although many somatic mutations typically accumulate in broadly neutralizing antibodies during maturation, only a handful of mutations substantially stabilize CDR-H3 loop and hence enhance the affinity of the antibodies for antigen. In our studies, all the results (Figures S12 and S13 and Table S2 in Supplementary Material) for the anti-influenza antibody, except FIRST-PG flexibility calculations for the Rosetta

FastRelax ensemble, show rigidification of the CDR-H3 loop as an effect of affinity maturation and agree with the detailed analysis of Schmidt et al.

For the three antibody families we analyzed in detail, we observed mixed effects of affinity maturation on two (catalytic antibodies and 4-4-20) and clear rigidification in one (anti-influenza antibody). For the two with mixed results, we note that past work has also shown conflicting results. We interpret these results as supportive of our repertoire-wide analysis that affinity maturation does not always rigidify the CDR-H3 loop.

Biophysical Properties Underlying Antibody Binding

Why is antibody CDR-H3 loop rigidification not a consistent result of affinity maturation? Consider the process of affinity maturation, which selects for antibody–antigen binding and against interactions with self or damaged antibodies (i.e., when deleterious mutations are introduced by activation-induced cytidine deaminase) (71). Under these selection pressures, what is the benefit of CDR-H3 loop rigidification? Loop rigidification can only decrease the protein-entropy cost for antibody–antigen binding, having ostensibly no effect on enthalpy and solvent entropy of binding, and self-interactions. If CDR-H3 loop rigidification is just one of many biophysical mechanisms that can be selected for during affinity maturation, then we do not expect to observe it consistently, in line with our results.

What are the other possible mechanisms then? Collectively, studies have shown that improved antibody affinity and specificity for antigen can be achieved by introducing additional interfacial interactions, including hydrogen bonds, salt bridges, and van der Waals contacts (16, 72–74); increasing the buried surface area, either polar or apolar, depending on the antigen (20); and improving interface shape complementarity (20, 75), in addition to rigidification of the paratope (22). A detailed review on the structural basis of antibody affinity maturation, by Mishra and Mariuzza, can be found in this research topic (76).

An interesting consequence of the biological antibody selection process is the anti-hapten antibody, SPE7 (77). For SPE7, mutations leading to multi-specificity or promiscuity were beneficial—antibodies are multivalent, so an antibody capable of binding multiple antigens with intermediate affinity can gain an effective advantage through cooperative binding over an antibody capable of binding only one antigen. Crystal structures of SPE7 with different antigens and in its apo-state demonstrated that SPE7 can assume different conformations. Motivated by these observations, Wang et al. exploited MD simulations to investigate the binding mechanisms of SPE7 (78). The MD simulations and subsequent analyses suggested that multi-specific antigen binding is mediated by a combined mechanism of conformer selection and induced fit. Similar behavior, where the mature antibody is more flexible than the naïve has been observed for an antibody that recognizes the tumor-associated ganglioside GD2 (79). Such antibodies could not have arisen if CDR-H3 loop rigidification were a consistent result of affinity maturation.

CONCLUSION

We have conducted the largest-scale flexibility study of antibody CDR-H3 loops, analyzing 922 crystal structures and 1,911 homology models. We used B-factors and FIRST-PG to assess flexibility. We sought to identify the effects of affinity maturation on CDR-H3 loop flexibility, expecting the CDR-H3 loop to rigidify. We found that there were no differences in the CDR-H3 loop B-factor distributions or FIRST-PG DOFs for naïve vs. mature antibody crystal structures and in the CDR-H3 FIRST-PG DOFs for homology models of repertoires of naïve and mature antibodies. These findings suggest that there is no general difference between naïve and mature antibody CDR-H3 loop flexibility in repertoires of naïve and mature antibodies. However, we observed rigidification of the CDR-H3 loop for some, but not all, antibodies when the mature antibodies were compared directly to their germline predecessors. So, we conclude that increased rigidity occurs alongside other affinity increasing changes, such as improved interfacial interactions, increased buried surface area, and improved shape complementarity.

Further work must be done to address the issues observed here, i.e., inconsistent results across the different methods are used to measure flexibility. One possible route is to explore experimental methods that directly measure protein dynamics across several timescales, and use them to study a relatively large (more than one or two antibodies) and diverse (e.g., from different source organisms or capable of binding different antigens) set of antibodies. For example, HDX-MS is capable of identifying protein regions with dynamics on timescales from milliseconds to days, has been previously used to study antibody dynamics, and has been correlated to FIRST-PG (29, 41, 80).

Finally, we note the need for more rapid and accurate antibody modeling methods. With the advent of high-throughput sequencing, there now exits a plethora of antibody sequence data, but little structural data. Accurate modeling can overcome the lack of high-throughput structure determination method and provide crucial structural data. These structures can then be used to address scientific questions on a larger scale than before, on the scale of the human antibody repertoire.

AUTHOR CONTRIBUTIONS

JJ, AS, DK, and JG designed the research. JJ, AS, DK, and NT performed the research. JJ, AS, and DK analyzed the data. JJ, AS, DK, NT, NK, KT, and JG wrote the paper.

ACKNOWLEDGMENTS

The authors would like to acknowledge Oana I. Lungu and Erik L. Johnson for sharing the antibody repertoire homology models. The super computing resources in this study have been provided in part by the Maryland Advanced Research Computing Center, the ROIS National Institute of Genetics, and the Human Genome Center at the Institute of Medical Science, The University of Tokyo, Japan.

FUNDING

JJ was funded by NIGMS grants F31-GM123616 and T32-GM008403. AS was supported by JST CREST Grant Number JPMJCR1402 (Japan) and NSERC (Canada). NT and NK were supported by JST CREST Grant Number JPMJCR1402 (Japan). DK was funded by Japan Society for the Promotion of Science Grant Number 17K18113 and Japanese Initiative for Progress of Research on Infectious Disease for Global Epidemic (J-PRIDE) Grant Number JP17fm0208022h. JJ and JG were funded by NIGMS grant R01-GM078221.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/fimmu.2018.00413/full#supplementary-material.

REFERENCES

1. Tonegawa S. Somatic generation of antibody diversity. *Nature* (1983) 302(5909):575–81. doi:10.1038/302575a0
2. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* (2007) 76:1–22. doi:10.1146/annurev.biochem.76.061705.090740
3. de los Rios M, Criscitiello MF, Smider VV. Structural and genetic diversity in antibody repertoires from diverse species. *Curr Opin Struct Biol* (2015) 33:27–41. doi:10.1016/j.sbi.2015.06.002
4. Clark LA, Ganesan S, Papp S, van Vlijmen HW. Trends in antibody sequence changes during the somatic hypermutation process. *J Immunol* (2006) 177(1):333–40. doi:10.4049/jimmunol.177.1.333
5. Burkovitz A, Sela-Culang I, Ofran Y. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J* (2014) 281(1):306–19. doi:10.1111/febs.12597
6. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) 196(4):901–17. doi:10.1016/0022-2836(87)90412-8
7. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature* (1989) 342(6252):877–83. doi:10.1038/342877a0
8. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* (1997) 273(4):927–48. doi:10.1006/jmbi.1997.1354
9. Kuroda D, Shirai H, Kobori M, Nakamura H. Systematic classification of CDR-L3 in antibodies: implications of the light chain subtypes and the VL-VH interface. *Proteins* (2009) 75(1):139–46. doi:10.1002/prot.22230
10. North B, Lehmann A, Dunbrack RL Jr. A new clustering of antibody CDR loop conformations. *J Mol Biol* (2011) 406(2):228–56. doi:10.1016/j.jmb.2010.10.030
11. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* (1998) 275(2):269–94. doi:10.1006/jmbi.1997.1442
12. Kuroda D, Shirai H, Kobori M, Nakamura H. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins* (2008) 73(3):608–20. doi:10.1002/prot.22087
13. Weitzner BD, Dunbrack RL Jr, Gray JJ. The origin of CDR H3 structural diversity. *Structure* (2015) 23(2):302–11. doi:10.1016/j.str.2014.11.010
14. Tsuchiya Y, Mizuguchi K. The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Sci* (2016) 25(4):815–25. doi:10.1002/pro.2874
15. Regep C, Georges G, Shi J, Popovic B, Deane CM. The H3 loop of antibodies shows unique structural characteristics. *Proteins* (2017) 85(7):1311–8. doi:10.1002/prot.25291
16. Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC. Structural insights into the evolution of an antibody combining site. *Science* (1997) 276(5319):1665–9. doi:10.1126/science.276.5319.1665
17. Mundorff EC, Hanson MA, Varvak A, Ulrich H, Schultz PG, Stevens RC. Conformational effects in biological catalysis: an antibody-catalyzed oxy-cope rearrangement. *Biochemistry* (2000) 39(4):627–32. doi:10.1021/bi9924314
18. Yin J, Mundorff EC, Yang PL, Wendt KU, Hanway D, Stevens RC, et al. A comparative analysis of the immunological evolution of antibody 28B4. *Biochemistry* (2001) 40(36):10764–73. doi:10.1021/bi010536c
19. Yin J, Beuscher AEt, Andryski SE, Stevens RC, Schultz PG. Structural plasticity and the evolution of antibody affinity and specificity. *J Mol Biol* (2003) 330(4):651–6. doi:10.1016/S0022-2836(03)00631-4
20. Li Y, Li H, Yang F, Smith-Gill SJ, Mariuzza RA. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Biol* (2003) 10(6):482–8. doi:10.1038/nsb930
21. Manivel V, Sahoo NC, Salunke DM, Rao KV. Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity* (2000) 13(5):611–20. doi:10.1016/S1074-7613(00)00061-3
22. Schmidt AG, Xu H, Khan AR, O'Donnell T, Khurana S, King LR, et al. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proc Natl Acad Sci U S A* (2013) 110(1):264–9. doi:10.1073/pnas.1218256109
23. Thielges MC, Zimmermann J, Yu W, Oda M, Romesberg FE. Exploring the energy landscape of antibody-antigen complexes: protein dynamics, flexibility, and molecular recognition. *Biochemistry* (2008) 47(27):7237–47. doi:10.1021/bi800374q
24. Adhikary R, Yu W, Oda M, Zimmermann J, Romesberg FE. Protein dynamics and the diversity of an antibody response. *J Biol Chem* (2012) 287(32):27139–47. doi:10.1074/jbc.M112.372698
25. Adhikary R, Yu W, Oda M, Walker RC, Chen T, Stanfield RL, et al. Adaptive mutations alter antibody structure and dynamics during affinity maturation. *Biochemistry* (2015) 54(11):2085–93. doi:10.1021/bi501417q
26. Jimenez R, Salazar G, Baldridge KK, Romesberg FE. Flexibility and molecular recognition in the immune system. *Proc Natl Acad Sci U S A* (2003) 100(1):92–7. doi:10.1073/pnas.262411399
27. Jimenez R, Salazar G, Yin J, Joo T, Romesberg FE. Protein dynamics and the immunological evolution of molecular recognition. *Proc Natl Acad Sci U S A* (2004) 101(11):3803–8. doi:10.1073/pnas.0305745101
28. Zimmermann J, Oakman EL, Thorpe IF, Shi X, Abbyad P, Brooks CL III, et al. Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics. *Proc Natl Acad Sci U S A* (2006) 103(37):13722–7. doi:10.1073/pnas.0603282103
29. Davenport TM, Gorman J, Joyce MG, Zhou T, Soto C, Guttman M, et al. Somatic hypermutation-induced changes in the structure and dynamics of HIV-1 broadly neutralizing antibodies. *Structure* (2016) 24(8):1346–57. doi:10.1016/j.str.2016.06.012
30. Chong LT, Duan Y, Wang L, Massova I, Kollman PA. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proc Natl Acad Sci U S A* (1999) 96(25):14330–5. doi:10.1073/pnas.96.25.14330
31. Thorpe IF, Brooks CL III. Molecular evolution of affinity and flexibility in the immune system. *Proc Natl Acad Sci U S A* (2007) 104(21):8821–6. doi:10.1073/pnas.0610064104
32. Wong SE, Sellers BD, Jacobson MP. Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins* (2011) 79(3):821–9. doi:10.1002/prot.22920
33. Li T, Tracka MB, Uddin S, Casas-Finet J, Jacobs DJ, Livesay DR. Rigidity emerges during antibody evolution in three distinct antibody systems: evidence from QSFR analysis of Fab fragments. *PLoS Comput Biol* (2015) 11(7):e1004327. doi:10.1371/journal.pcbi.1004327
34. Di Palma F, Tramontano A. Dynamics behind affinity maturation of an anti-HCMV antibody family influencing antigen binding. *FEBS Lett* (2017) 591(18):2936–50. doi:10.1002/1873-3468.12774
35. Babor M, Kortemme T. Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal

for conformational flexibility. *Proteins* (2009) 75(4):846–58. doi:10.1002/prot.22293
36. Willis JR, Briney BS, DeLuca SL, Crowe JE Jr, Meiler J. Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput Biol* (2013) 9(4):e1003045. doi:10.1371/journal.pcbi.1003045
37. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113(19):E2636–45. doi:10.1073/pnas.1525510113
38. Weitzner BD, Jeliazkov JR, Lyskov S, Marze N, Kuroda D, Frick R, et al. Modeling and docking of antibody structures with Rosetta. *Nat Protoc* (2017) 12(2):401–16. doi:10.1038/nprot.2016.180
39. Weitzner BD, Kuroda D, Marze N, Xu J, Gray JJ. Blind prediction performance of Rosetta antibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins* (2014) 82(8):1611–23. doi:10.1002/prot.24534
40. Sljoka A. *Algorithms in Rigidity Theory with Applications to Protein Flexibility and Mechanical Linkages [Thesis Ph.D.]*. Toronto, Canada: Graduate Programme in Mathematics and Statistics, York University (2012).
41. Sljoka A, Wilson D. Probing protein ensemble rigidity and hydrogen-deuterium exchange. *Phys Biol* (2013) 10(5):056013. doi:10.1088/1478-3975/10/5/056013
42. Kim TH, Mehrabi P, Ren Z, Sljoka A, Ing C, Bezginov A, et al. The role of dimer asymmetry and protomer dynamics in enzyme catalysis. *Science* (2017) 355:6322. doi:10.1126/science.aag2355
43. Deng B, Zhu S, Macklin AM, Xu J, Lento C, Sljoka A, et al. Suppressing allostery in epitope mapping experiments using millisecond hydrogen/deuterium exchange mass spectrometry. *MAbs* (2017) 9(8):1327–36. doi:10.1080/19420862.2017.1379641
44. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. *Proteins* (2001) 44(2):150–65. doi:10.1002/prot.1081
45. Whiteley W. Counting out to the flexibility of molecules. *Phys Biol* (2005) 2(4):S116–26. doi:10.1088/1478-3975/2/4/S06
46. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. *Nucleic Acids Res* (2014) 42(Database issue):D1140–6. doi:10.1093/nar/gkt1043
47. Ehrenmann F, Kaas Q, Lefranc MP. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* (2010) 38(Database issue):D301–7. doi:10.1093/nar/gkp946
48. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popovic Z, et al. Algorithm discovery by protein folding game players. *Proc Natl Acad Sci U S A* (2011) 108(47):18949–53. doi:10.1073/pnas.1115898108
49. Nivon LG, Moretti R, Baker D. A Pareto-optimal refinement method for protein design scaffolds. *PLoS One* (2013) 8(4):e59004. doi:10.1371/journal.pone.0059004
50. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Methods* (2009) 6(8):551–2. doi:10.1038/nmeth0809-551
51. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* (2017) 13(6):3031–48. doi:10.1021/acs.jctc.7b00125
52. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. *J Comput Chem* (2005) 26(16):1781–802. doi:10.1002/jcc.20289
53. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* (2017) 14(1):71–3. doi:10.1038/nmeth.4067
54. Darden T, York D, Pedersen L. Particle Mesh Ewald – an N.Log(N) method for Ewald sums in large systems. *J Chem Phys* (1993) 98(12):10089–92. doi:10.1063/1.464397
55. Li T, Tracka MB, Uddin S, Casas-Finet J, Jacobs DJ, Livesay DR. Redistribution of flexibility in stabilizing antibody fragment mutants follows Le Chatelier's principle. *PLoS One* (2014) 9(3):e92870. doi:10.1371/journal.pone.0092870
56. Feig M, Karanicolas J, Brooks CL III. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J Mol Graph Model* (2004) 22(5):377–95. doi:10.1016/j.jmgm.2003.12.005
57. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* (2011) 32(10):2319–27. doi:10.1002/jcc.21787
58. Kaas Q, Ruiz M, Lefranc MP. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res* (2004) 32(Database issue):D208–10. doi:10.1093/nar/gkh042
59. Mamonova T, Hespenheide B, Straub R, Thorpe MF, Kurnikova M. Protein flexibility using constraints from molecular dynamics simulations. *Phys Biol* (2005) 2(4):S137–47. doi:10.1088/1478-3975/2/4/S08
60. Kuroda D, Gray JJ. Pushing the backbone in protein-protein docking. *Structure* (2016) 24(10):1821–9. doi:10.1016/j.str.2016.06.025
61. ÓConchúir S, Barlow KA, Pache RA, Ollikainen N, Kundert K, O'Meara MJ, et al. A web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design. *PLoS One* (2015) 10(9):e0130433. doi:10.1371/journal.pone.0130433
62. Demirel MC, Lesk AM. Molecular forces in antibody maturation. *Phys Rev Lett* (2005) 95(20):208106. doi:10.1103/PhysRevLett.95.208106
63. Murphy K, Weaver C. *Janeway's Immunobiology*. 9th ed. New York, NY: Garland Science (2017). p. 1–904.
64. Kuroda D, Shirai H, Jacobson MP, Nakamura H. Computer-aided antibody design. *Protein Eng Des Sel* (2012) 25(10):507–21. doi:10.1093/protein/gzs024
65. Furukawa K, Akasako-Furukawa A, Shirai H, Nakamura H, Azuma T. Junctional amino acids determine the maturation pathway of an antibody. *Immunity* (1999) 11(3):329–38. doi:10.1016/S1074-7613(00)80108-9
66. Furukawa K, Shirai H, Azuma T, Nakamura H. A role of the third complementarity-determining region in the affinity maturation of an antibody. *J Biol Chem* (2001) 276(29):27622–8. doi:10.1074/jbc.M102714200
67. Boder ET, Midelfort KS, Wittrup KD. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc Natl Acad Sci U S A* (2000) 97(20):10701–5. doi:10.1073/pnas.170297297
68. Midelfort KS, Hernandez HH, Lippow SM, Tidor B, Drennan CL, Wittrup KD. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J Mol Biol* (2004) 343(3):685–701. doi:10.1016/j.jmb.2004.08.019
69. Fukunishi H, Shimada J, Shiraishi K. Antigen-antibody interactions and structural flexibility of a femtomolar-affinity antibody. *Biochemistry* (2012) 51(12):2597–605. doi:10.1021/bi3000319
70. Xu H, Schmidt AG, O'Donnell T, Therkelsen MD, Kepler TB, Moody MA, et al. Key mutations stabilize antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. *Proteins* (2015) 83(4):771–80. doi:10.1002/prot.24745
71. Eisen HN, Chakraborty AK. Evolving concepts of specificity in immune reactions. *Proc Natl Acad Sci U S A* (2010) 107(52):22373–80. doi:10.1073/pnas.1012051108
72. Alzari PM, Spinelli S, Mariuzza RA, Boulot G, Poljak RJ, Jarvis JM, et al. Three-dimensional structure determination of an anti-2-phenyloxazolone antibody: the role of somatic mutation and heavy/light chain pairing in the maturation of an immune response. *EMBO J* (1990) 9(12):3807–14.
73. Mizutani R, Miura K, Nakayama T, Shimada I, Arata Y, Satow Y. Three-dimensional structures of the Fab fragment of murine N1G9 antibody from the primary immune response and of its complex with (4-hydroxy-3-nitrophenyl)acetate. *J Mol Biol* (1995) 254(2):208–22. doi:10.1006/jmbi.1995.0612
74. Yuhasz SC, Parry C, Strand M, Amzel LM. Structural analysis of affinity maturation: the three-dimensional structures of complexes of an anti-nitrophenol antibody. *Mol Immunol* (1995) 32(14–15):1143–55. doi:10.1016/0161-5890(95)00063-1
75. Kuroda D, Gray JJ. Shape complementarity and hydrogen bond preferences in protein-protein interfaces: implications for antibody modeling and protein-protein docking. *Bioinformatics* (2016) 32(16):2451–6. doi:10.1093/bioinformatics/btw197
76. Mishra AK, Mariuzza RA. Insights into the structural basis of antibody affinity maturation from next-generation sequencing. *Front Immunol* (2018) 9:117. doi:10.3389/fimmu.2018.00117

77. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. *Science* (2003) 299(5611):1362–7. doi:10.1126/science.1079731
78. Wang W, Ye W, Yu Q, Jiang C, Zhang J, Luo R, et al. Conformational selection and induced fit in specific antibody and antigen recognition: SPE7 as a case study. *J Phys Chem B* (2013) 117(17):4912–23. doi:10.1021/jp4010967
79. Sterner E, Peach ML, Nicklaus MC, Gildersleeve JC. Therapeutic antibodies to ganglioside GD2 evolved from highly selective germline antibodies. *Cell Rep* (2017) 20(7):1681–91. doi:10.1016/j.celrep.2017.07.050
80. Weis DD. *Hydrogen Exchange Mass Spectrometry of Proteins: Fundamentals, Methods, and Applications*. Chichester, West Sussex: John Wiley & Sons, Inc (2016). xxiii, 350, 46 p.

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.

The reviewer OL declared a past co-authorship with several of the authors (DK JG) to the handling editor.

Copyright © 2018 Jeliazkov, Sljoka, Kuroda, Tsuchimura, Katoh, Tsumoto and Gray. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

ORIGINAL RESEARCH
published: 08 March 2018
doi: 10.3389/fimmu.2018.00395

Many Routes to an Antibody Heavy-Chain CDR3: Necessary, Yet Insufficient, for Specific Binding

Sara D'Angelo[1†], Fortunato Ferrara[1†], Leslie Naranjo[1], M. Frank Erasmus[1], Peter Hraber[2] and Andrew R. M. Bradbury[1]*

[1] Specifica Inc., Santa Fe, NM, United States, [2] Los Alamos National Laboratory, Los Alamos, NM, United States

Because of its great potential for diversity, the immunoglobulin heavy-chain complementarity-determining region 3 (HCDR3) is taken as an antibody molecule's most important component in conferring binding activity and specificity. For this reason, HCDR3s have been used as unique identifiers to investigate adaptive immune responses *in vivo* and to characterize *in vitro* selection outputs where display systems were employed. Here, we show that many different HCDR3s can be identified within a target-specific antibody population after *in vitro* selection. For each identified HCDR3, a number of different antibodies bearing differences elsewhere can be found. In such *selected* populations, all antibodies with the same HCDR3 recognize the target, albeit at different affinities. In contrast, within *unselected* populations, the majority of antibodies with the same HCDR3 sequence do not bind the target. In one HCDR3 examined in depth, all target-specific antibodies were derived from the same VDJ rearrangement, while non-binding antibodies with the same HCDR3 were derived from many different V and D gene rearrangements. Careful examination of previously published *in vivo* datasets reveals that HCDR3s shared between, and within, different individuals can also originate from rearrangements of different V and D genes, with up to 26 different rearrangements yielding the same identical HCDR3 sequence. On the basis of these observations, we conclude that the same HCDR3 can be generated by many different rearrangements, but that specific target binding is an outcome of unique rearrangements and VL pairing: the HCDR3 is necessary, albeit insufficient, for specific antibody binding.

Keywords: heavy-chain complementarity-determining region 3, single-chain Fv display, binding specificity, rearrangement, inverse PCR

OPEN ACCESS

Edited by:
Jacob Glanville, Distributed Bio, United States

Reviewed by:
Mepur Hanumantha-Rao Ravindranath, Terasaki Foundation, United States
Ramya Yarlagadda, Intrexon, United States

***Correspondence:**
Andrew R. M. Bradbury
abradbury@specifica.bio

†These authors have contributed equally to this work.

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *01 December 2017*
Accepted: *13 February 2018*
Published: *08 March 2018*

Citation:
D'Angelo S, Ferrara F, Naranjo L, Erasmus MF, Hraber P and Bradbury ARM (2018) Many Routes to an Antibody Heavy-Chain CDR3: Necessary, Yet Insufficient, for Specific Binding. Front. Immunol. 9:395. doi: 10.3389/fimmu.2018.00395

INTRODUCTION

Antibodies bind their targets using diversified loops, termed complementarity-determining regions (CDRs), with three in each rearranged VH and VL gene. CDRs 1 and 2 are encoded by germline V genes, while CDR3s in both VH and VL are the product of gene recombination. Compared to other CDRs, the varied length and biochemical properties of heavy-chain complementarity-determining region 3 (HCDR3) contribute to enhanced sequence diversity (1). It has been estimated (2) that the theoretical HCDR3 diversity exceeds 10^{15} variants, generated from fixed genomic sequences by combinatorial and junctional diversification mechanisms. This underlies the vast diversity of the human antibody repertoire. The fully assembled V(D)J gene and its incorporated HCDR3 are derived from the sequential random assembly of 56 VH, 23 DH, and 6 JH genes (3–6). While both

VH and JH contribute to the HCDR3, the DH forms the central core. Although DH genes are predominantly read in one frame, all three frames can be used (7, 8), further increasing potential diversity. It was initially thought that D genes could also be inverted and duplicated (9, 10); however, recent deep sequencing results indicate that this is unlikely (7). Diversity is further increased by P-nucleotide-mediated (11) or N-nucleotide-mediated (12, 13) addition, or exonuclease-mediated loss (11, 14), of nucleotides between the VH/DH and DH/JH segments. Recombination between VH genes after rearrangement provides further diversification although it remains unclear how much this contributes overall (15).

There is much evidence that the HCDR3 is the major determinant of antibody-binding specificity. Specific antibodies have been selected from synthetic antibody libraries where diversity is restricted to the HCDR3 (16–18). It has been shown that a greater number of antibodies were selected from a synthetic library containing only HCDR3 diversity than when the same library was combined with LCDR3 diversity (18). However, due to the random diversity in both CDR3s, this lower performance may have been due to a reduced fitness profile, caused by either the higher mutational load or potential inter-CDR structural clashes. In addition to display libraries, transgenic mice with antibody diversity restricted to the HCDR3 were able to generate high-affinity responses *in vivo* (19). HCDR3's themselves have also been harvested as diversity elements (20–22), and low-affinity binders have been selected from fluorescent scaffold libraries in which they provide the only diversity (23). Further evidence that HCDR3s are the major determinants of antibody-binding specificity arises from the observation that peptides derived from HCDR3 structures can show biological activity similar to the antibodies from which they were derived (24, 25), in one case even demonstrating *in vivo* viral neutralization (25). Furthermore, peptide libraries generated from naïve IgM HCDR3s produce specific binders against targets (21, 26), often more efficiently than synthetic peptide libraries. HCDR3s have also been transplanted from antibodies to other proteins, conferring the expected binding activity upon those non-antibody scaffolds. These include the HCDR3 from different antibodies transplanted into neocarzinostatin (27), sfGFP (23), or an epidermal growth factor-like module of human tissue-type plasminogen activator (28). In each of these cases, the grafted HCDR3 recapitulated the antibody-binding activity. Although it is evident that the HCDR3 is critical in antigen binding, diversity confined to the LCDR3 can still generate specific antibodies (29), and it is known from affinity maturation experiments that the affinities of antibodies with identical HCDR3s may differ by up to 100-fold (30, 31).

Although the other antibody-binding loops have defined canonical structures (32–35), the prediction of the HCDR3 conformation is not trivial and has been found to have a wide variety of different possible configurations (35). In addition to the structural variability of HCDR3s with different sequences, the same HCDR3 can adopt different conformations within the same antibody when bound to different targets (36) or in uncomplexed antibodies with different VH/VL frameworks (37). This reflects the important role that HCDR3 plays in target recognition by antibodies (10, 19, 38) and likely shows that HCDR3 conformational flexibility is an additional diversity mechanism employed by the immune system.

The diverse nature of the HCDR3 has led to its use as a fingerprint both *in vivo* (39–43) and *in vitro* (44–50). In this article, we have assessed the diversity of HCDR3 sequences in an *in vitro* selected antibody population. We found that *in vitro* selection elicits hundreds of different target-specific HCDR3s, but that only within the context of a target-specific antibody population, antibodies with the same HCDR3 recognize the target. In an unselected population, we were unable to identify any sequenced antibodies with the same HCDR3 that was target specific. We conclude that the HCDR3 is necessary, but insufficient, for specific antibody binding.

RESULTS

Selection of Anti-CDK2 Antibodies from a Naïve Human Recombinant Library

We selected antibodies against CDK2, a human cyclin-dependent kinase, provided by the Structural Genomic Consortium (SGC; Toronto), from a well-validated (44, 45, 51–59) large naïve phage antibody library in the single-chain Fv (scFv) format, created by site-specific recombination (59, 60). This library was previously used to develop a combined phage and yeast display approach (45, 53), which has the advantage that many more antibodies can be identified than by regular phage display, particularly when combined with next-generation sequencing (NGS) (44). After two rounds of phage selection (using biotinylated CDK2 antigen and streptavidin-coated magnetic beads) and two rounds of yeast sorting (at 100 nM antigen concentration), almost all yeast displaying antibodies recognized the target, as shown in **Figure 1A**.

The HCDR3s of the final sorted population were sequenced using IonTorrent. 32,138 total HCDR3 sequences were obtained and analyzed with the Antibody Mining Toolbox (61). 535 different HCDR3s aa sequences made up 98% of all the sequences analyzed. The remaining 2% of sequences mainly comprised HCDR3 represented by one or two sequences and were ignored for the purposes of this study as the result of possible sequencing errors. The 535 HCDR3 sequences were ranked by abundance, and their distribution is presented in **Figure 1B**. The majority of HCDRs were represented by limited numbers of clones.

In a previous publication (44), we described the identification and affinity determination of representative clones of 8 of the 10 most abundant HCDR3 clones. All isolated clones bound CDK2, and for each HCDR3 sequence identified, the affinity of one representative clone was assessed directly on the yeast surface (62). The affinities of these eight most abundant identified clones were found to range from 2 to 75 nM, as previously described (44).

HCDR3-Based Rescue of Anti-CDK2 Antibodies and Characterization

To assess whether an identified HCDR3 corresponded to a single clone, or a sublibrary of clones, we used an inverse PCR approach anchored within the most abundant unique HCDR3

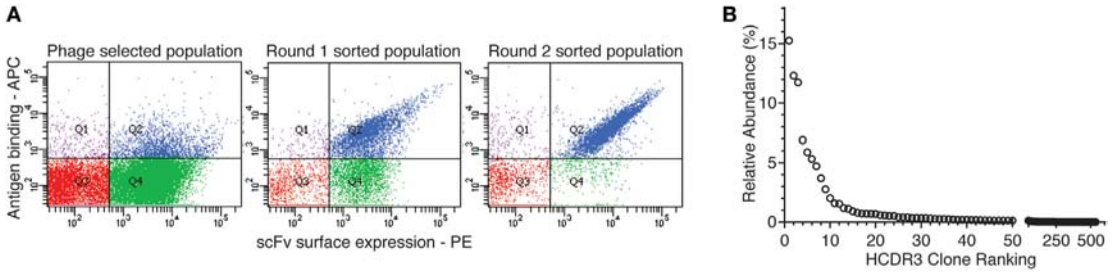


FIGURE 1 | Yeast display sorting and analysis of CDK2-specific antibodies. **(A)** After two rounds of phage selection, the antibody population was displayed on yeast cells and enriched for CDK2-specific binders by two rounds of flow cytometry-assisted sorting performed at 100 nM. **(B)** Abundance distribution of the selected anti-CDK2 clones as identified by their heavy-chain complementarity-determining region 3 (HCDR3s) by next-generation sequencing.

sequence (50) to isolate all clones containing this HCDR3 within the sorting-enriched population.

When 24 random clones from this sublibrary were Sanger sequenced, eight different antibodies, all containing the same HCDR3, were identified. Additional mutations were identified in the rest of the VH, and significantly more in the VL. In fact, for the VL, four different CDR1, two different CDR2 and two different CDR3 sequences, were identified (**Figure 2A**). However, the different clones shared 91.6–97.8% homology, and the same rearrangements were responsible for VH (5-51, D2-08, and J3) and VL (IGLV3-21 and IGLJ1). Finally, when the affinities of the eight expressing and binding clones were calculated, they spanned a 10-fold range (from 30.1 to 352.5 nM) (**Figure 2B**), reflecting the 100 nM target concentration used for sorting.

HCDR3-Based Rescue of Non-Selected Antibodies and Characterization

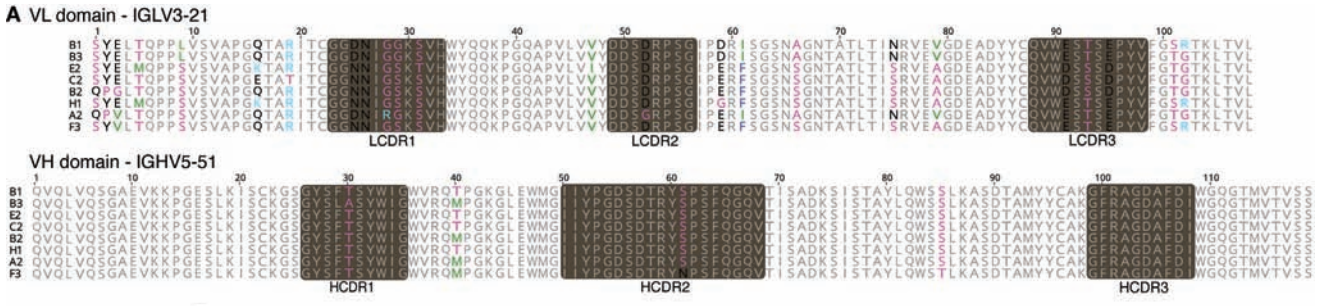
Given this broad variation in affinity, and the known importance of the HCDR3 in antibody binding specificity, we applied the inverse PCR technique to the original naïve, unselected library to assess the relative abundance of clones containing the HCDR3 sequence of interest. The inverse PCR reaction, using primers specific for the top-ranked clone, was performed using a plasmid preparation of the naïve library as a template. The obtained mini-library was transformed into yeast-competent cells, and, upon induction, the cells were sorted for well-displayed antibodies (**Figure 3**, left and middle panels). Of note, after sorting for expression and analyzing such population for binding to CDK2 (**Figure 3**, third panel), only 0.15% of the clones showed binding for the cognate antigen. Ninety six clones of the population sorted for expression were sequenced. We identified 55 different scFvs containing the same HCDR3 with a far greater variation in both VH and VL than seen in the clones isolated from the selected population: 37.7–73.2% homology to the selected clones. **Figure 4A** shows the alignment of the VH regions of the sequenced clones. The VL families, not being under any particular selective pressure, were very diverse, derived from 12 VL-kappa and 10 VL-lambda germline genes, with 4 JL-kappa genes and 3 JL-lambda genes represented. The VH genes, on the other hand, having all been selected to contain the same HCDR3 sequence, were, not surprisingly, all found to have the same JH and DH genes (with one exception). More surprising was the diversity of the VH germline genes, which comprised 5 VH families derived from 19 different germline VH genes (**Figure 4B**; **Table 1**).

Testing of the scFvs by flow cytometry from the selected and naïve populations revealed that none of the scFvs containing this HCDR3 isolated from the naïve library were able to bind CDK2, while all those from the selected population bound CDK2 (**Figure 4C**).

We were surprised that the same HCDR3 could be assembled from so many different germline VH genes in the naïve unselected library. As the library we used was originally created by cloning the rearranged VH and VL genes of peripheral blood lymphocytes from 40 donors, this convergent HCDR3 assembly may be a normal consequence of the generation of antibody diversity, or it could be a result of the various PCR reactions we performed to create the library, as well as the final inverse PCR anchored within the HCDR3.

In Vivo HCDR3 Generation

In order to assess the prevalence of identical HCDR3s derived from different germline genes *in vivo*, we analyzed two publicly available datasets of naïve B cell sequences (63, 64), referred to respectively as the "DeKosky" and the "DeWitt" datasets. In the first, ~55,000 naïve VH sequences from three donors were obtained by paired end MiSeq reads, and 23 HCDR3 sequence pairs were found to be shared between two of the three donors in the naïve repertoires. This represents a frequency of 0.083% shared HCDR3s. No HCDR3 was found to be shared among all three donors. Interestingly, all of these 23 HCDR3 pairs were discordant for identified VH germline genes, and seven were also discordant for the identified DH gene (**Table 2A**). However, all pairs share the same JH gene. In the second analysis (64), based on a dataset of 8,596,145 productive MiSeq reads comprising 7,984,053 unique HCDR3s from the naïve B cells of three donors, we identified 568 identical HCDR3s (0.007% of the total unique HCDR3s) generated by different VDJ recombinations (as determined by IMGT). These were generated using from two to 26 different VDJ combinations (see Table S1 in Supplementary Material), and 176 of these rearrangements were found in all three donors. Two of the HCDR3s with the



B

Clone ID	VL-region identity (%)	VL gene	JL gene	VH-region identity (%)	VH gene	DH gene	JH gene	Kd (nM)	R^2
B1	94.98	IGLV3-21	IGLJ1	98.96	IGHV5-51	IGHD2-8	IGHJ3	30.1	0.98
B3	95.34	IGLV3-21	IGLJ1	98.61	IGHV5-51	IGHD2-8	IGHJ3	61.6	0.98
E2	97.13	IGLV3-21	IGLJ1	98.61	IGHV5-51	IGHD2-8	IGHJ3	64.5	0.98
C2	97.85	IGLV3-21	IGLJ1	98.26	IGHV5-51	IGHD2-8	IGHJ3	84.4	0.97
B2	97.49	IGLV3-21	IGLJ1	98.61	IGHV5-51	IGHD2-8	IGHJ3	135.8	0.97
H1	97.49	IGLV3-21	IGLJ1	98.61	IGHV5-51	IGHD2-8	IGHJ3	173.6	0.98
A2	93.91	IGLV3-21	IGLJ1	98.26	IGHV5-51	IGHD2-8	IGHJ3	203.9	0.97
F3	98.57	IGLV3-21	IGLJ1	99.31	IGHV5-51	IGHD2-8	IGHJ3	352.5	0.99

FIGURE 2 | Sequences and affinities of single-chain Fvs with an identical heavy-chain complementarity-determining region 3 (HCDR3). Inverse PCR of the most abundant clone of the selected population was analyzed by Sanger sequencing. **(A)** Sanger sequence analysis of the eight unique clones with identical HCDR3 obtained by inverse PCR. **(B)** VDJ gene usage and affinity values of the eight different clones.

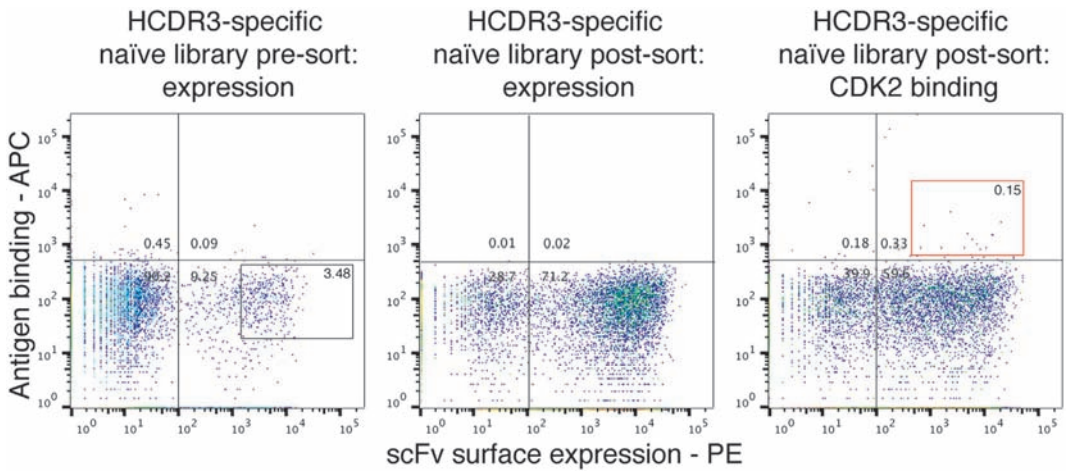


FIGURE 3 | Analysis of the naïve library for single-chain Fvs (scFvs) with identical heavy-chain complementarity-determining region 3 (HCDR3s). Yeast clones displaying scFvs from the HCDR3-specific mini-library obtained from the naïve library *via* inverse PCR were analyzed by flow cytometry. Cells expressing scFvs on the surface and detected by anti-SV5-PE antibody were gated and sorted (left panel). Once grown and analyzed, they showed an increase in the expressing population (middle panel) with a limited, but detectable, binding for biotinylated CDK2 detected by streptavidin-APC conjugation (right panel).

greatest number of rearrangements are illustrated in **Table 2B**. The first, CARDRGDYW, was generated from 14 different VH genes (from five different VH gene families), five different DH genes and one JH gene in 26 different combinations. Five of these combinations were found in two donors, and two were found in all three donors, the remaining were unique combinations found in individual donors. The second, CARDSSGWYYFDYW, was a longer HCDR3 and was generated from 20 different VH genes (from all seven different VH gene families) and only one DH and one JH gene. Six of the combinations were found in two donors and four in all three donors.

DISCUSSION

Next-generation sequencing has been widely applied to many areas of human immunology, helping, for instance, to increase understanding of antibody repertoires (64–71), VH/VL pairing (39, 67), humoral responses to pathogens (72–74), vaccination (41, 73, 75, 76), and the role of antibodies in autoimmune conditions (77) and cancer (78). In addition to its role in understanding natural *in vivo* humoral responses, NGS has also been used in the practice of *in vitro* antibody selection, including in the sequencing of antibodies selected by phage (47–49) and yeast

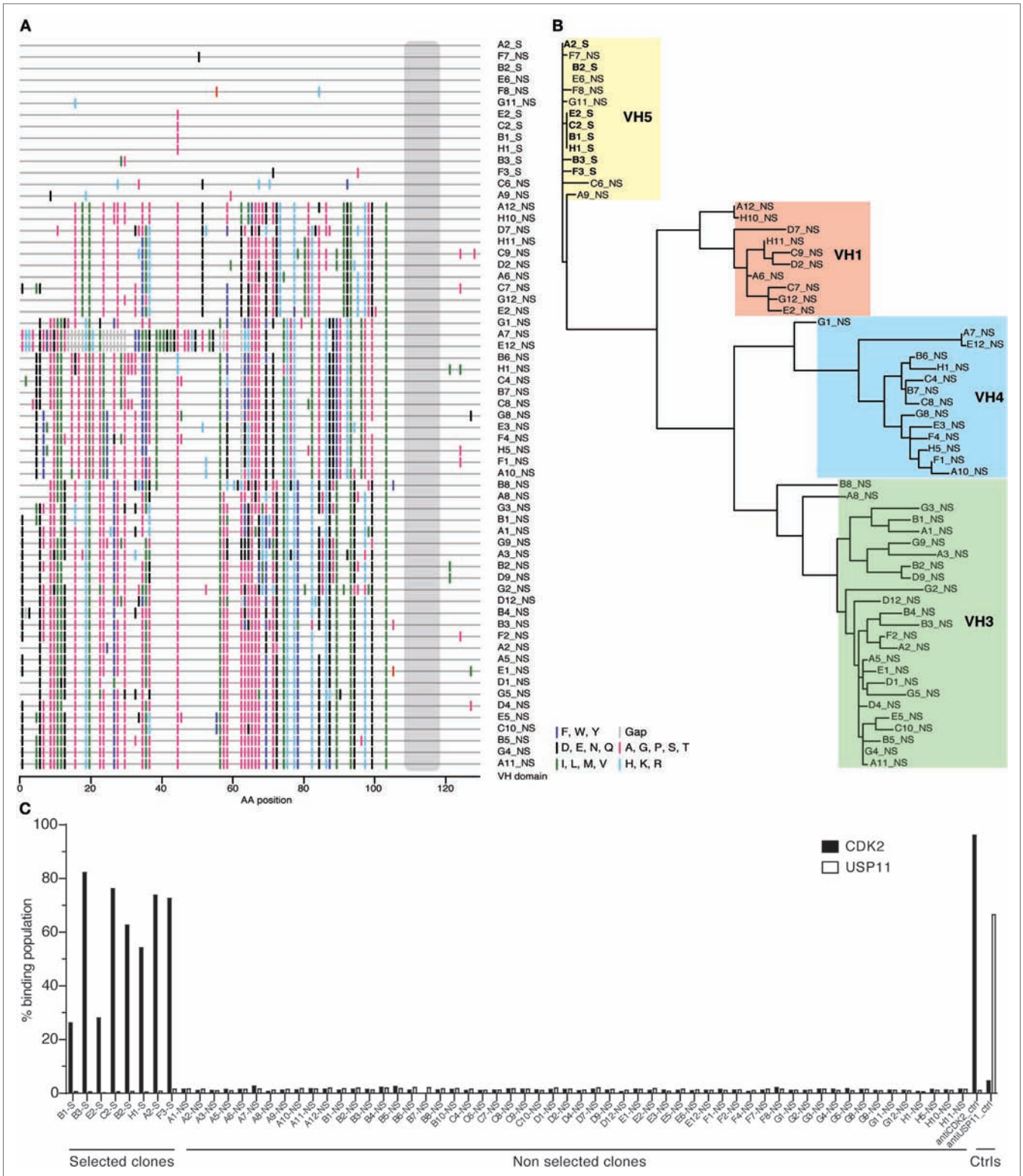


FIGURE 4 | Analysis of antibody sequences obtained by inverse PCR. **(A)** The alignment of antibody sequences with the same heavy-chain complementarity-determining region 3 (HCDR3) from naïve and selected libraries shows differences in VH gene sequences. This representation, a so-called highlighter plot, indicates amino acid mutations in the sites that differ from the first sequence. The alignment was generated with the Highlighter Tool (hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html). **(B)** The sequences of the antibodies sharing the same HCDR3 obtained by inverse PCR from the population enriched for specific antigen binders were phylogenetically clustered to show relations among variants that share VH alleles. **(C)** Single-chain Fvs (scFvs) sharing the same HCDR3 but obtained from either the selected or the naïve population were displayed on yeast and their binding activity was measured by flow cytometry and reported in the histogram as percentage of binding. None of the scFvs isolated from the naïve library bound CDK2, as opposed to the ones from the selected population.

TABLE 1 | Analyses of the gene usage from the heavy-chain complementarity-determining region 3 (HCDR3)-specific non-selected population.

Clone ID	VH gene	DH gene	JH gene	HCDR3	VL gene	JL gene	LCDR3
C7	IGHV1-2	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2-28	IGKJ3	CMQTLQTPFTF
G12	IGHV1-2	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV2-14	IGLJ1	CSSYTSVSTYVF
E2	IGHV1-2	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV7-46	IGLJ2	CLLDYTDARVF
D7	IGHV1-3	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-39	IGKJ1	CQQSYSTPWTF
A6	IGHV1-46	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2-28	IGKJ4	CMQSLQTPLTF
C9	IGHV1-46	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV2-8	IGLJ2	CSSYAGSNNVVF
D2	IGHV1-46	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV2-14	IGLJ1	CCSYGGPYVF
H11	IGHV1-46	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-5	IGKJ4	CQQYYSTPLTF
A12	IGHV1-69	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV1-40	IGLJ1	CQSYDSSLSGYVF
H10	IGHV1-69	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV2-18	IGLJ1	CSSYTSSSTYVF
B2	IGHV3-21	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-12	IGKJ3	CQQTNSFPFTF
D9	IGHV3-21	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-9	IGKJ5	CHQTDTLPITF
A11	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-5	IGKJ2	CQQYDTLPRTF
B5	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-39	IGKJ4	CQQSYSTPPTF
G4	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV4-1	IGKJ2	CQQYHSTPYTF
A2	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-5	IGKJ2	CQQYVECSF
A5	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2-28	IGKJ2	CMQALQSPRTF
A8	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3-11	IGKJ2	CQQYNNWPPYTF
B4	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-17	IGKJ4	CLQHNLYPRTF
D1	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3D-15	IGKJ4	CQQYNNWPPLTF
D4	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3D-20	IGKJ2	CQQFGGSPKCSF
E1	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-17	IGKJ4	CLQHNTYPLTF
F2	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2-28	IGKJ1	CMQALQTPWTF
G2	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV3-1	IGLJ2	CQVWDSNSHVVF
G5	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3D-20	IGKJ5	CQQRSNWPLTF
B3	IGHV3-23	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV2-14	IGLJ2	CAAWDSSLSAVVF
A1	IGHV3-30	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-33	IGKJ5	CQQYDKLPLTF
B1	IGHV3-30	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-12	IGKJ2	CQQGYSFPRTF
D12	IGHV3-53	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV3-21	IGLJ2	CQAWDTHDDPWGVF
E5	IGHV3-64	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-33	IGKJ2	CVQHRGYPRYTF
C10	IGHV3-64D	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3-11	IGKJ3	CQQRINRVTF
A3	IGHV3-7	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3D-20	IGKJ3	CQQYSLYPLSF
G9	IGHV3-7	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV1-44	IGLJ1	CQAWDSRTAVF
B8	IGHV3-72	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-9	IGKJ4	CQQLNSYPLAF
G3	IGHV3-9	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3-11	IGKJ5	CQQRGNWPPGATF
G1	IGHV4-31	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV3-1	IGLJ2	CQAWDSGTVVF
B6	IGHV4-31	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2-28	IGKJ2	CMQALQSPRTF
H1	IGHV4-31	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV3-1	IGLJ2	CQAWDSSTAVF
F1	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-17	IGKJ1	CLQHNNYPRTF
G8	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2D-29	IGKJ2	CMQGTHWPRTF
H5	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV4-1	IGKJ4	CQQYYSTPLTF
A10	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-17	IGKJ1	CLQHNNYPRTF
E3	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV2-28	IGKJ1	CMQALQAPWTL
F4	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV3-21	IGLJ2	CQVWDSRDQHVAF
A7	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-33	IGKJ2	CQQYDNLRYSF
E12	IGHV4-34	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV3-11	IGKJ2	CQQRSNSPPFTF
C4	IGHV4-4	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV1-51	IGLJ3	CGTWDSSLSAGVF
B7	IGHV4-59	IGHD3-10	IGHJ3	CAKGFRAGDAFDIW	IGKV4-1	IGKJ3	CQQFYSTPPLFTF
C8	IGHV4-61	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV2-8	IGLJ2	CSSYTGSNNWRVVF
A9	IGHV5-51	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-5	IGKJ3	CQQSYSTPLFTF
C6	IGHV5-51	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-5	IGKJ4	CLQHDEYPLTF
E6	IGHV5-51	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-5	IGKJ4	CQQADSVPLTF
F7	IGHV5-51	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV4-1	IGKJ3	CQQYSSIPFTF
F8	IGHV5-51	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGLV7-43	IGLJ3	CLLYYGGAQLGVF
G11	IGHV5-51	IGHD2-8	IGHJ3	CAKGFRAGDAFDIW	IGKV1-39	IGKJ1	SQQSYDSPPMTF

(44, 50) display, as well as in the analysis of naïve antibody libraries (61, 79, 80). We have previously shown (61) that the number of specific antibody HCDR3s that can be identified using NGS after a combined phage/yeast selection protocol far exceeds the number that can be isolated using standard low-throughput analysis and sequencing methods. After *in vitro* selections, we routinely use the HCDR3s as unique identifiers to rank antibody abundance. Identified clones can then be isolated by inverse PCR (50).

In the work presented here, we show that 535 different HCDR3s are identified by NGS of a yeast displayed population that is positive for binding to CDK2 (**Figure 1A**). This mirrors previous work in which we showed that hundreds of different HCDR3s were able to mediate specific binding against a number

TABLE 2 | Analyses of (A) DeKosky and (B) DeWitt data for identical HCDR3s.

(A)

HCDR3 aa sequence	Length	VH family	VH gene	DH gene	JH gene
CAKDGYW	5	IGHV01	IGHV1-46		IGHJ4
CAKDGYW	5	IGHV03	IGHV3-23		IGHJ4
CARADDAFDIW	9	IGHV01	IGHV1-69		IGHJ3
CARADDAFDIW	9	IGHV04	IGHV4-34	IGHD1-1	IGHJ3
CARALYYFDYW	9	IGHV01	IGHV1-46	IGHD1-7	IGHJ4
CARALYYFDYW	9	IGHV01	IGHV1-2	IGHD3-16	IGHJ4
CARDKYYFDYW	9	IGHV01	IGHV1-69	IGHD1-14	IGHJ4
CARDKYYFDYW	9	IGHV04	IGHV4-59		IGHJ4
CARDLDYW	6	IGHV03	IGHV3-11		IGHJ4
CARDLDYW	6	IGHV03	IGHV3-33		IGHJ4
CARDPFDYW	7	IGHV01	IGHV1-69		IGHJ4
CARDPFDYW	7	IGHV01	IGHV1-46		IGHJ4
CARDPGPW	6	IGHV01	IGHV1-69	IGHD1-14	IGHJ5
CARDPGPW	6	IGHV03	IGHV3-33	IGHD1-14	IGHJ5
CARDRSSSFDYW	10	IGHV03	IGHV3-33	IGHD6-13	IGHJ4
CARDRSSSFDYW	10	IGHV03	IGHV3-74	IGHD6-13	IGHJ4
CARDSGNDYW	8	IGHV07	IGHV7-4	IGHD6-13	IGHJ4
CARDSGNDYW	8	IGHV01	IGHV1-8	IGHD4-23	IGHJ4
CARDSSGYFDYW	10	IGHV01	IGHV1-46	IGHD3-22	IGHJ4
CARDSSGYFDYW	10	IGHV01	IGHV1-18	IGHD3-22	IGHJ4
CARDYCSGGSCYFDYW	14	IGHV04	IGHV4-31	IGHD2-15	IGHJ4
CARDYCSGGSCYFDYW	14	IGHV03	IGHV3-21	IGHD2-15	IGHJ4
CARGAAPDYW	8	IGHV01	IGHV1-46	IGHD5-12	IGHJ4
CARGAAPDYW	8	IGHV03	IGHV3-53	IGHD2-15	IGHJ4
CARGAYYFDYW	9	IGHV04	IGHV4-59	IGHD3-16	IGHJ4
CARGAYYFDYW	9	IGHV03	IGHV3-33		IGHJ4
CARGGNWFDPW	9	IGHV04	IGHV4-34	IGHD3-10	IGHJ5
CARGGNWFDPW	9	IGHV04	IGHV4-30	IGHD2-15	IGHJ5
CARGGYGDYVDYW	11	IGHV01	IGHV1-46	IGHD4-17	IGHJ4
CARGGYGDYVDYW	11	IGHV01	IGHV1-18	IGHD4-17	IGHJ4
CARGIAAADYW	9	IGHV03	IGHV3-48	IGHD6-13	IGHJ4
CARGIAAADYW	9	IGHV01	IGHV1-69	IGHD6-13	IGHJ4
CARGRVFDYW	8	IGHV04	IGHV4-34	IGHD3-16	IGHJ4
CARGRVFDYW	8	IGHV02	IGHV2-5	IGHD1-26	IGHJ4
CARGSSFDYW	8	IGHV04	IGHV4-59	IGHD3-10	IGHJ4
CARGSSFDYW	8	IGHV03	IGHV3-53	IGHD6-6	IGHJ4
CARGVAARDYW	9	IGHV04	IGHV4-59	IGHD6-6	IGHJ4
CARGVAARDYW	9	IGHV03	IGHV3-48	IGHD6-6	IGHJ4
CARRFDPW	6	IGHV03	IGHV3-21		IGHJ5
CARRFDPW	6	IGHV04	IGHV4-34		IGHJ5
CARRLGNWYFDLW	11	IGHV03	IGHV3-11	IGHD3-10	IGHJ2
CARRLGNWYFDLW	11	IGHV04	IGHV4-61	IGHD7-27	IGHJ2
CARVGSGWYFDYW	11	IGHV03	IGHV3-66	IGHD6-19	IGHJ4
CARVGSGWYFDYW	11	IGHV03	IGHV3-7	IGHD6-19	IGHJ4
CASNDAFDIW	8	IGHV01	IGHV1-46		IGHJ3
CASNDAFDIW	8	IGHV05	IGHV5-10		IGHJ3

(B)

HCDR3 aa Sequence	# rearrangements	VH family	VH gene	DH gene	JH gene	Donor representation
CARDRGDYW	26	IGHV01	IGHV01-02	IGHD03-10	IGHJ04-01	donor1
CARDRGDYW	26	IGHV01	IGHV01-02	IGHD05-24	IGHJ04-01	donor3
CARDRGDYW	26	IGHV01	IGHV01-03	IGHD01-26	IGHJ04-01	donor3
CARDRGDYW	26	IGHV01	IGHV01-03	IGHD03-10	IGHJ04-01	donor3
CARDRGDYW	26	IGHV01	IGHV01-03	IGHD06-25	IGHJ04-01	donor1
CARDRGDYW	26	IGHV01	IGHV01-18	IGHD03-10	IGHJ04-01	donor3
CARDRGDYW	26	IGHV01	IGHV01-18	IGHD03-16	IGHJ04-01	donor1, donor3
CARDRGDYW	26	IGHV01	IGHV01-18	IGHD05-24	IGHJ04-01	donor2
CARDRGDYW	26	IGHV01	IGHV01-18	IGHD06-25	IGHJ04-01	donor3
CARDRGDYW	26	IGHV01	IGHV01-46	IGHD03-10	IGHJ04-01	donor1, donor2, donor3
CARDRGDYW	26	IGHV01	IGHV01-46	IGHD03-16	IGHJ04-01	donor2
CARDRGDYW	26	IGHV01	IGHV01-69	IGHD03-10	IGHJ04-01	donor1, donor2, donor3

(*Continued*)

TABLE 2 | Continued

(B)

HCDR3 aa Sequence	# rearrangements	VH family	VH gene	DH gene	JH gene	Donor representation
CARDRGDYW	26	IGHV01	IGHV01-69	IGHD03-16	IGHJ04-01	donor2
CARDRGDYW	26	IGHV03	IGHV03-11	IGHD03-10	IGHJ04-01	donor1
CARDRGDYW	26	IGHV03	IGHV03-11	IGHD03-16	IGHJ04-01	donor1
CARDRGDYW	26	IGHV03	IGHV03-13	IGHD03-10	IGHJ04-01	donor1
CARDRGDYW	26	IGHV03	IGHV03-48	IGHD03-10	IGHJ04-01	donor3
CARDRGDYW	26	IGHV03	IGHV03-53	IGHD03-10	IGHJ04-01	donor1, donor3
CARDRGDYW	26	IGHV03	IGHV03-53	IGHD03-16	IGHJ04-01	donor3
CARDRGDYW	26	IGHV03	IGHV03-53	IGHD05-24	IGHJ04-01	donor1
CARDRGDYW	26	IGHV03	IGHV03-64	IGHD03-10	IGHJ04-01	donor1, donor3
CARDRGDYW	26	IGHV03	IGHV03-66	IGHD03-10	IGHJ04-01	donor2, donor3
CARDRGDYW	26	IGHV04	IGHV04-39	IGHD03-10	IGHJ04-01	donor1, donor3
CARDRGDYW	26	IGHV04	IGHV04-39	IGHD03-16	IGHJ04-01	donor3
CARDRGDYW	26	IGHV05	IGHV05-51	IGHD03-16	IGHJ04-01	donor2
CARDRGDYW	26	IGHV07	IGHV07-04	IGHD03-10	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV01	IGHV01-02	IGHD06-19	IGHJ04-01	donor1, donor2, donor3
CARDSSGWYYFDYW	20	IGHV01	IGHV01-03	IGHD06-19	IGHJ04-01	donor1, donor2, donor3
CARDSSGWYYFDYW	20	IGHV01	IGHV01-08	IGHD06-19	IGHJ04-01	donor1
CARDSSGWYYFDYW	20	IGHV01	IGHV01-18	IGHD06-19	IGHJ04-01	donor1, donor2, donor3
CARDSSGWYYFDYW	20	IGHV01	IGHV01-46	IGHD06-19	IGHJ04-01	donor1, donor3
CARDSSGWYYFDYW	20	IGHV01	IGHV01-69	IGHD06-19	IGHJ04-01	donor2, donor3
CARDSSGWYYFDYW	20	IGHV02	IGHV02-70	IGHD06-19	IGHJ04-01	donor1, donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-11	IGHD06-19	IGHJ04-01	donor1
CARDSSGWYYFDYW	20	IGHV03	IGHV03-20	IGHD06-19	IGHJ04-01	donor1
CARDSSGWYYFDYW	20	IGHV03	IGHV03-23	IGHD06-19	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-48	IGHD06-19	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-53	IGHD06-19	IGHJ04-01	donor1, donor2, donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-64	IGHD06-19	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-66	IGHD06-19	IGHJ04-01	donor2, donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-72	IGHD06-19	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV03	IGHV03-74	IGHD06-19	IGHJ04-01	donor1, donor2
CARDSSGWYYFDYW	20	IGHV04	IGHV04-39	IGHD06-19	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV05	IGHV05-51	IGHD06-19	IGHJ04-01	donor3
CARDSSGWYYFDYW	20	IGHV06	IGHV06-01	IGHD06-19	IGHJ04-01	donor1, donor2
CARDSSGWYYFDYW	20	IGHV07	IGHV07-04	IGHD06-19	IGHJ04-01	donor3

of different targets (44). When antibodies containing the most abundant HCDR3 were isolated from the *selected* pool using specific inverse primers, a single scFV gene was not obtained, but an "oligoclonal" population of specific binders, comprising at least eight different antibody sequences. These are all very similar to one another (91.6–97.8% homology), with most of the variation occurring in the VL but with the germline VH, DH, JH, VL, and JL genes identified as being identical. Analysis of these clones revealed a 10-fold difference in affinity (Kd), confirming the importance of additional antibody structure beyond the HCDR3 in modulating binding activity (30, 31) and indicating that the true diversity of anti-CDK2 antibodies could be significantly higher than 535, when variability in VL and HCDR1 and HCDR2 is also taken into account.

Given the identification of different antibodies with identical HCDR3s, all of which bound the target in the selected population, we turned to the naïve library to assess whether the same HCDR3 within the context of different antibodies would also be able to bind the target. By using the same inverse PCR approach, a far more diverse collection of antibodies, all of which contained the same HCDR3, was isolated. However, none of these were able to bind the target, and analysis of the aligned sequences revealed that apart from the identical HCDR3's, these antibodies comprised very different VL genes. This was not surprising since VLs were randomly recombined and not under selective pressure. More surprising was the finding that the frameworks and CDR1 and CDR2 of the VHs were largely diverse, corresponding to 19 different germline VH genes. When this population was tested for binding to CDK2 by flow cytometry, only 0.15% of displayed antibodies with the identical HCDR3 bound the target. On the basis of these findings, we conclude that a specific HCDR3 will only define a particular binding specificity within a very narrow structurally appropriate context: i.e., HCDR3 is necessary, but is insufficient to define specific antibody-binding properties unless combined with appropriate VL and VH germline genes. This is perhaps not surprising given a recent report in which structural analysis of the same HCDR3 sequence placed within the context of different VH and VL genes shows significant conformational diversity (37). Those results, along with those presented here, suggest that the conformations of HCDR3 conformations are modified not only by their sequences but also by the structural environment in which they are found: in particular their VH and VL pairing.

It is remarkable that so many different rearranged VH genes, derived from 19 germline genes, were found to contain the same HCDR3. This begs the question as to whether the generation of

identical HCDR3s from different germline genes is biological in nature, or a result of the molecular biological manipulations we had undertaken in these experiments. In a couple of published NGS analyses of *in vivo* naïve B cell HCDR3 repertoires (63, 64), 0.04–0.08% of HCDR3s were found to be shared between any pairs of donors. Further analysis of the sequences described in the study by DeKosky et al. (63) (**Table 2A**) reveals that *all* these so-called public HCDR3s were derived from different germline VH (and in some cases DH) genes, suggesting that the generation of identical HCDR3 sequences is stochastic and usually occurs using different germline VH and DH combinations. This conclusion was confirmed and extended by a much larger second dataset (64), which was generated by sequencing the naïve B cell repertoire of three individuals at far greater depth (>8.5 M productive reads total). Different rearrangements encoding identical HCDR3s were found both within and between donors. Altogether, 568 different HCDR3s generated with from 2 to 26 different rearrangements were identified (see Table S1 in Supplementary Material). Of these, 176 rearrangements, comprising 155 different HCDR3s, were found in all 3 donors. In a particularly notable example, the same four rearrangements (using four different VH genes and the same DH and JH genes) were found in all three donors for two of these HCDR3s (CARGYSSGWYYFDYW and CARDSSGWYYFDYW) (see Table S1 in Supplementary Material; **Table 2B**). These results demonstrate that the creation of identical HCDR3s from different VH or DH germline genes is a regular, albeit rare, occurrence *in vivo* and that the sequences of the HCDR3s, as well as the rearrangements used to create them, are shared among different individuals. The observed *in vitro* HCDR3 rearrangement diversity, therefore, more likely reflects the original *in vivo* recombination, rather than the consequence of molecular biological manipulation. This is further confirmed by *in vitro* selection experiments from natural naïve libraries (81), in which it was found that antibodies with the same HCDR3 sequence were derived using different VH genes.

The library used here (59) was created from the rearranged V genes of 40 donors and is estimated to comprise approximately 3.3×10^6 different HCDR3s (61). The *in vivo* data described above suggest that most, if not all, of the identical HCDR3s identified in the naïve library were stochastically derived from different germline VH, DH, and JH gene rearrangements in the original donors. However, it cannot be excluded that this natural diversity was supplemented by some of the *in vitro* molecular biological manipulations we have carried out. In particular, the inverse PCR primers we used to isolate all identical HCDR3s may be "correcting" slightly different sequences to the desired HCDR3, even if, given the primer lengths (18–23 bases), they would have to be extremely similar to be able to do this. Furthermore, inadvertent PCR errors may have increased the apparent diversity of the surrounding VH gene despite the use of a proofreading polymerase. One surprising result was the low percentage (0.15%) of CDK2-binding clones containing the identified HCDR3 in the naïve library.

The earliest naïve *in vitro* antibody libraries (16, 82, 83) had claimed diversities of ~5×10^7, and an average of ~4 antibodies were selected per target. A smaller subset (10^7) of a much larger library yielded ~1 antibody per target (84). Assuming the diversity estimates for the sizes of these (sub)libraries is correct, these results suggest that one should expect one positive antibody per ~10^7 different antibodies, consistent with theoretical analyses of library size (85). However, as library size has scaled upward (to claimed diversities of >10^{11}), the number of antibodies selected against individual targets has generally remained below 100 in the absence of heroic efforts (81). The use of deep sequencing described here, and elsewhere (44, 47–50), indicates that the gap between the potential diversity of selectable antibodies, and the significantly lower number usually analyzed is predominantly a sampling problem, which can be overcome with ongoing improvements in sequencing technology. This will allow the calculation, rather than the estimation, of the true diversity of antibody repertoires and antigen-specific populations selected from them.

MATERIALS AND METHODS

Bacterial and Yeast Strains

DH5aF′: F′/endA1 hsdR17(rKmK+) supE44 thi-1 recA1 gyrA (Na1r) relA1 D(lacZYAargF) U169 (m80lacZDM15)
Omnimax (Life Technologies): F′ {proAB lacIq lacZM15 Tn10(TetR) (ccdAB)} mcrA (mrr hsdRMS-mcrBC) 80(lacZ) M15 (lacZYAargF)U169 endA1 recA1 supE44 thi-1 gyrA96 relA1 tonA panD
EBY100 (kindly provided by Prof. Dane Wittrup): MATa AGA::GAL1-AGA1::URA3 ura3-52 trp1 leu2-delta200 his3-delta200 pep4::HIS3 prb11.6R can1 GAL

scFv Antibody Selections

In vivo biotinylated His-tagged CDK2 protein (NP_001789.2), produced by the SGC (Toronto) was used for the scFv phage display selections. The naïve scFv library described in the study by Sblattero and Bradbury (59) was used for two rounds of phage display against the antigen with streptavidin magnetic beads. Two additional rounds of yeast display sorting were performed using 100 nM of antigen. The detailed protocol for antibody selections against biotinylated proteins is described in the study by Ferrara et al. (45).

For the selection of clones sharing same HCDR3 derived from the naïve library and cloned into yeast display vector (see below), cells were induced and labeled with anti-SV5-PE to assess the scFv display level.

All the flow cytometry-assisted sorting experiments were performed using the FACSAria (Becton Dickinson) sorter and analyzed using FlowJo software (FlowJo LLC).

When single clones were analyzed for their specificity, they were stained with CDK2, unrelated antigen, and conjugated streptavidin, as negative controls. All experiments with single clones were performed in a 96-well format using the LSRII (Becton Dickinson) flow cytometer.

Next-Generation Sequencing

The plasmid DNA of the anti-CDK2 second sort output was used as a template for the PCR targeting the HCDR3 region of

the scFvs. A set of forward primers mapping to the framework region upstream of the HCDR3 and carrying one of the Ion Torrent sequencing adaptors were used in combination with a barcoded reverse primer mapping to the common SV5 tag region of the yeast display vector and carrying the second adaptor required for sequencing. The primer sequences and method are described in detail by D'Angelo et al. (61). Once amplified with the proofreading Phusion polymerase (NEB), gel extracted, and quantified (Q-bit, HS-DNA kit, Invitrogen), the amplicon libraries were processed using the Ion Xpress Amplicon library protocol and then prepared for sequencing on the Ion 316 Chip (Life Technologies). The sequences analysis was performed using the AbMining Toolbox as described by D'Angelo et al. (61).

Primer Design and Inverse PCR

The inverse PCR strategy is described in the study by D'Angelo et al. (50). Briefly, primers were designed on the DNA consensus sequence for the HCDR3 of the top-ranked clone as back to back primers directed outward from the middle of the HCDR3, with a 5′ phosphorylated forward primer. The inverse PCR was carried out using a high-fidelity polymerase with proofreading activity (Phusion High Fidelity Polymerase, NEB) and either 0.03 fmol of plasmid DNA obtained from the yeast sorted population (1,000–10,000 times the diversity of the sorting output) or 0.3 fmol of the original phage naïve library were used as a template.

After amplification, the PCR product was gel extracted and purified (Qiaquick Gel extraction kit, Qiagen) to avoid contamination from the original plasmid template. The purified products were ligated with T4 ligase and transformed into DH5aF′ bacterial cells.

For the clones obtained from the second yeast sort enriched for CDK2 binders, single clones were analyzed by Sanger sequencing to confirm the presence of the correct HCDR3 and obtain the sequence of the full-length scFv, before carrying out binding assays. The sequenced plasmid clones were then retransformed into the EBY100 yeast display strain (Yeast transformation kit, Sigma) for testing by flow cytometry.

When the entire unselected naïve library was used as a template to isolate clones sharing the same HCDR3 by inverse PCR, ligation and bacteria transformation were performed. A plasmid preparation was obtained and transformed into the EBY100 yeast display cells. The final product was a sublibrary of scFvs, with 10^6 clones sharing the same HCDR3.

Affinity Measurement

The affinity of the selected clones was determined by yeast display using the equilibrium binding titration curve to extrapolate the equilibrium dissociation constant (K_D), as described by Boder and Wittrup (86). Briefly, the induced, monoclonal populations of yeast displaying scFvs were incubated with eight concentrations of biotinylated antigen, ranging from 1 to 500 nM for 30 min at room temperature to allow the binding reaction to reach equilibrium followed by 5 min on ice for 5 min to reduce the off-rate. After washing, the yeast cells were incubated with the secondary reagents (streptavidin-Alexa633, to detect antigen binding to the displayed scFvs, and anti-SV5-PE, to detect the yeast displayed scFvs) on ice for 30 min. After the final washes, the samples were analyzed by flow cytometry on the BDAriaIII (BD Biosciences). The mean fluorescence intensities of the gated binding/displaying populations were plotted against the antigen concentration, and a non-linear least-squares analysis was used to fit the curve and obtain the K_D values for each scFv.

In Vivo Database Analysis

In the first database (63), the identity of VH, DH, and JH genes making up each individual HCDR3 were determined on the basis of the nucleotide sequences using IMGT (87) and NCBI IgBlast software (88), along with a CDR3 motif identification algorithm (89). The final data set comprising HCDR3 sequences found in three individuals was kindly provided by DeKosky and Georgiou. The second database (64) used a scored alignment across a definition list of all the known VH, DH, and JH genes in IMGT (90). The naïve data sets for three individual donors were downloaded from the public repository found at http://adaptivebiotech.com/pub/robins-bcell-2016.[1] Data were filtered through RStudio software package [RStudio Team (2015) RStudio: Integrated Development for R. RStudio, Inc. (Boston, MA, USA)],[2] whereby 7.4×10^6, 6.0×10^6, and 8.4×10^6 individual sequences were processed for the three separate donors, respectively. In some cases, the families could be identified, but not the individual germline VH, DH, or JH genes. Therefore, this initial data set was processed further to include only complete data sets (e.g., HCDR3, VH, DH, JH designations). Any sequence that contained multiple gene or family designations and/or stop codons within HCDR3 was excluded. The final curated set consisted of 2.6×10^6, 2.4×10^6, and 3.6×10^6 in the three respective donors, of which the unique HCDR3 sequences and VH, DH, and JH gene recombination were tabulated.

AUTHOR CONTRIBUTIONS

SD and FF equally contributed to this work. SD, FF, and AB contributed to research design. SD, FF, and LN conducted experiments; SD, FF, PH, and ME performed data analysis. SD, FF, ME, PH, and AB wrote the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Brandon DeKosky and Dr. George Georgiou for kindly sharing their antibody data set and Marissa Vignali for help accessing the DeWitt database.

FUNDING

This work was supported by the National Institutes of Health (1-U54-DK093500-01 to AB).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/articles/10.3389/fimmu.2018.00395/full#supplementary-material.

[1] http://doi.org/10.21417/B71018.

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

REFERENCES

1. Zemlin M, Klinger M, Link J, Zemlin C, Bauer K, Engler JA, et al. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol* (2003) 334(4):733–49. doi:10.1016/j.jmb.2003.10.007
2. Elhanati Y, Sethna Z, Marcou Q, Callan CG Jr, Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci* (2015) 370(1676):20140243. doi:10.1098/rstb.2014.0243
3. Jung D, Giallourakis C, Mostoslavsky R, Alt FW. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* (2006) 24:541–70. doi:10.1146/annurev.immunol.23.021704.115830
4. Bassing CH, Swat W, Alt FW. The mechanism and regulation of chromosomal V(D)J recombination. *Cell* (2002) 109(Suppl):S45–55. doi:10.1016/S0092-8674(02)00675-X
5. Matsuda F, Shin EK, Hirabayashi Y, Nagaoka H, Yoshida MC, Zong SQ, et al. Organization of variable region segments of the human immunoglobulin heavy chain: duplication of the D5 cluster within the locus and interchromosomal translocation of variable region segments. *EMBO J* (1990) 9(8):2501–6.
6. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* (2009) 37(Database issue):D1006–12. doi:10.1093/nar/gkn838
7. Benichou J, Glanville J, Prak ET, Azran R, Kuo TC, Pons J, et al. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol* (2013) 190(11):5567–77. doi:10.4049/jimmunol.1201929
8. Corbett SJ, Tomlinson IM, Sonnhammer EL, Buck D, Winter G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J Mol Biol* (1997) 270(4):587–97. doi:10.1006/jmbi.1997.1141
9. Sanz I. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J Immunol* (1991) 147(5):1720–9.
10. VanDyk L, Meek K. Assembly of IgH CDR3: mechanism, regulation, and influence on antibody diversity. *Int Rev Immunol* (1992) 8(2–3):123–33. doi:10.3109/08830189209055568
11. Feeney AJ, Victor KD, Vu K, Nadel B, Chukwuocha RU. Influence of the V(D)J recombination mechanism on the formation of the primary T and B cell repertoires. *Semin Immunol* (1994) 6(3):155–63. doi:10.1006/smim.1994.1021
12. Desiderio SV, Yancopoulos GD, Paskind M, Thomas E, Boss MA, Landau N, et al. Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells. *Nature* (1984) 311(5988):752–5. doi:10.1038/311752a0
13. Benedict CL, Gilfillan S, Thai TH, Kearney JF. Terminal deoxynucleotidyl transferase and repertoire development. *Immunol Rev* (2000) 175:150–7. doi:10.1111/j.1600-065X.2000.imr017518.x
14. Gauss GH, Lieber MR. Mechanistic constraints on diversity in human V(D)J recombination. *Mol Cell Biol* (1996) 16(1):258–69. doi:10.1128/MCB.16.1.258
15. Wilson PC, Wilson K, Liu YJ, Banchereau J, Pascual V, Capra JD. Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes. *J Exp Med* (2000) 191(11):1881–94. doi:10.1084/jem.191.11.1881
16. Barbas CF III, Bain JD, Hoekstra DM, Lerner RA. Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc Natl Acad Sci U S A* (1992) 89(10):4457–61. doi:10.1073/pnas.89.10.4457
17. Braunagel M, Little M. Construction of a semisynthetic antibody library using trinucleotide oligos. *Nucleic Acids Res* (1997) 25(22):4690–1. doi:10.1093/nar/25.22.4690
18. Mahon CM, Lambert MA, Glanville J, Wade JM, Fennell BJ, Krebs MR, et al. Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. *J Mol Biol* (2013) 425(10):1712–30. doi:10.1016/j.jmb.2013.02.015
19. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* (2000) 13(1):37–45. doi:10.1016/S1074-7613(00)00006-6
20. Kiss C, Fisher H, Pesavento E, Dai M, Valero R, Ovecka M, et al. Antibody binding loop insertions as diversity elements. *Nucleic Acids Res* (2006) 34(19):e132. doi:10.1093/nar/gkl681
21. Deroo S, Fischer A, Beaupain N, Counson M, Boutonnet N, Pletinckx J, et al. Non-immunized natural human heavy chain CDR3 repertoires allow the isolation of high affinity peptides mimicking a human influenza hemagglutinin epitope. *Mol Immunol* (2008) 45(5):1366–73. doi:10.1016/j.molimm.2007.09.001
22. Venet S, Ravn U, Buatois V, Gueneau F, Calloud S, Kosco-Vilbois M, et al. Transferring the characteristics of naturally occurring and biased antibody repertoires to human antibody libraries by trapping CDRH3 sequences. *PLoS One* (2012) 7(8):e43471. doi:10.1371/journal.pone.0043471
23. Dai M, Temirov J, Pesavento E, Kiss C, Velappan N, Pavlik P, et al. Using T7 phage display to select GFP-based binders. *Protein Eng Des Sel* (2008) 21(7):413–24. doi:10.1093/protein/gzn016
24. Levi M, Sallberg M, Ruden U, Herlyn D, Maruyama H, Wigzell H, et al. A complementarity-determining region synthetic peptide acts as a miniantibody and neutralizes human immunodeficiency virus type 1 in vitro. *Proc Natl Acad Sci U S A* (1993) 90(10):4374–8. doi:10.1073/pnas.90.10.4374
25. Bourgeois C, Bour JB, Aho LS, Pothier P. Prophylactic administration of a complementarity-determining region derived from a neutralizing monoclonal antibody is effective against respiratory syncytial virus infection in BALB/c mice. *J Virol* (1998) 72(1):807–10.
26. Chevigne A, Fischer A, Mathu J, Counson M, Beaupain N, Plesseria JM, et al. Selection of a CXCR4 antagonist from a human heavy chain CDR3-derived phage library. *FEBS J* (2011) 278(16):2867–78. doi:10.1111/j.1742-4658.2011.08208.x
27. Nicaise M, Valerio-Lepiniec M, Minard P, Desmadril M. Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. *Protein Sci* (2004) 13(7):1882–91. doi:10.1110/ps.03540504
28. Smith JW, Tachias K, Madison EL. Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin alpha IIb beta 3. *J Biol Chem* (1995) 270(51):30486–90. doi:10.1074/jbc.270.51.30486
29. Persson H, Ye W, Wernimont A, Adams JJ, Koide A, Koide S, et al. CDR-H3 diversity is not required for antigen recognition by synthetic antibodies. *J Mol Biol* (2013) 425(4):803–11. doi:10.1016/j.jmb.2012.11.037
30. Yang WP, Green K, Pinz-Sweeney S, Briones AT, Burton DR, Barbas CF III. CDR walking mutagenesis for the affinity maturation of a potent human anti-HIV-1 antibody into the picomolar range. *J Mol Biol* (1995) 254(3):392–403. doi:10.1006/jmbi.1995.0626
31. Schier R, McCall A, Adams GP, Marshall KW, Merritt H, Yim M, et al. Isolation of picomolar affinity anti-c-erbB-2 single-chain Fv by molecular evolution of the complementarity determining regions in the center of the antibody binding site. *J Mol Biol* (1996) 263(4):551–67. doi:10.1006/jmbi.1996.0598
32. Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) 196(4):901–17. doi:10.1016/0022-2836(87)90412-8
33. Chothia C, Lesk AM, Gherardi E, Tomlinson IM, Walter G, Marks JD, et al. Structural repertoire of the human VH segments. *J Mol Biol* (1992) 227(3):799–817. doi:10.1016/0022-2836(92)90224-8
34. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature* (1989) 342(6252):877–83. doi:10.1038/342877a0
35. North B, Lehmann A, Dunbrack RL Jr. A new clustering of antibody CDR loop conformations. *J Mol Biol* (2011) 406(2):228–56. doi:10.1016/j.jmb.2010.10.030
36. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. *Science* (2003) 299(5611):1362–7. doi:10.1126/science.1079731
37. Teplyakov A, Obmolova G, Malia TJ, Luo J, Muzammil S, Sweet R, et al. Structural diversity in a human antibody germline library. *MAbs* (2016) 8(6):1045–63. doi:10.1080/19420862.2016.1190060
38. Kabat EA, Wu TT. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J Immunol* (1991) 147(5):1709–19.
39. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2014) 21(1):86–91. doi:10.1038/nm.3743
40. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, et al. Convergent antibody signatures in human dengue. *Cell Host Microbe* (2013) 13(6):691–700. doi:10.1016/j.chom.2013.05.008

41. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci U S A* (2014) 111(13):4928–33. doi:10.1073/pnas.1323862111
42. Lu DR, Tan YC, Kongpachith S, Cai X, Stein EA, Lindstrom TM, et al. Identifying functional anti-*Staphylococcus aureus* antibodies by sequencing antibody repertoires of patient plasmablasts. *Clin Immunol* (2014) 152 (1–2):77–89. doi:10.1016/j.clim.2014.02.010
43. Tan YC, Blum LK, Kongpachith S, Ju CH, Cai X, Lindstrom TM, et al. High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin Immunol* (2014) 151(1):55–65. doi:10.1016/j.clim.2013.12.008
44. Glanville J, D'Angelo S, Khan TA, Reddy ST, Naranjo L, Ferrara F, et al. Deep sequencing in library selection projects: what insight does it bring? *Curr Opin Struct Biol* (2015) 33:146–60. doi:10.1016/j.sbi.2015.09.001
45. Ferrara F, D'Angelo S, Gaiotto T, Naranjo L, Tian H, Graslund S, et al. Recombinant renewable polyclonal antibodies. *MAbs* (2015) 7(1):32–41. doi:10.4161/19420862.2015.989047
46. Ferrara F, Naranjo LA, D'Angelo S, Kiss C, Bradbury AR. Specific binder for lightning-link(R) biotinylated proteins from an antibody phage library. *J Immunol Methods* (2013) 395(1–2):83–7. doi:10.1016/j.jim.2013.06.010
47. Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods* (2013) 60(1):99–110. doi:10.1016/j.ymeth.2013.03.001
48. Ravn U, Gueneau F, Baerlocher L, Osteras M, Desmurs M, Malinge P, et al. By-passing in vitro screening – next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* (2010) 38(21):e193. doi:10.1093/nar/gkq789
49. Lovgren J, Pursiheimo JP, Pyykko M, Salmi J, Lamminmaki U. Next generation sequencing of all variable loops of synthetic single framework scFv-application in anti-HDL antibody selections. *N Biotechnol* (2016) 33(6):790–6. doi:10.1016/j.nbt.2016.07.009
50. D'Angelo S, Kumar S, Naranjo L, Ferrara F, Kiss C, Bradbury AR. From deep sequencing to actual clones. *Protein Eng Des Sel* (2014) 27(10):301–7. doi:10.1093/protein/gzu032
51. Ferrara F, Kim CY, Naranjo LA, Bradbury AR. Large scale production of phage antibody libraries using a bioreactor. *MAbs* (2015) 7(1):26–31. doi:10.4161/19420862.2015.989034
52. Close DW, Ferrara F, Dichosa AE, Kumar S, Daughton AR, Daligault HE, et al. Using phage display selected antibodies to dissect microbiomes for complete de novo genome sequencing of low abundance microbes. *BMC Microbiol* (2013) 13:270. doi:10.1186/1471-2180-13-270
53. Ferrara F, Naranjo LA, Kumar S, Gaiotto T, Mukundan H, Swanson B, et al. Using phage and yeast display to select hundreds of monoclonal antibodies: application to antigen 85, a tuberculosis biomarker. *PLoS One* (2012) 7(11):e49535. doi:10.1371/journal.pone.0049535
54. Lillo AM, Ayriss JE, Shou Y, Graves SW, Bradbury AR, Pavlik P. Development of phage-based single chain Fv antibody reagents for detection of *Yersinia pestis*. *PLoS One* (2011) 6(12):e27756. doi:10.1371/journal.pone.0027756
55. Velappan N, Martinez JS, Valero R, Chasteen L, Ponce L, Bondu-Hawkins V, et al. Selection and characterization of scFv antibodies against the Sin Nombre hantavirus nucleocapsid protein. *J Immunol Methods* (2007) 321(1–2):60–9. doi:10.1016/j.jim.2007.01.011
56. Ayriss J, Woods T, Bradbury A, Pavlik P. High-throughput screening of single-chain antibodies using multiplexed flow cytometry. *J Proteome Res* (2007) 6(3):1072–82. doi:10.1021/pr0604108
57. Kehoe JW, Velappan N, Walbolt M, Rasmussen J, King D, Lou J, et al. Using phage display to select antibodies recognizing post-translational modifications independently of sequence context. *Mol Cell Proteomics* (2006) 5(12):2350–63. doi:10.1074/mcp.M600314-MCP200
58. Lou J, Marzari R, Verzillo V, Ferrero F, Pak D, Sheng M, et al. Antibodies in haystacks: how selection strategy influences the outcome of selection from molecular diversity libraries. *J Immunol Methods* (2001) 253(1–2):233–42. doi:10.1016/S0022-1759(01)00385-4
59. Sblattero D, Bradbury A. Exploiting recombination in single bacteria to make large phage antibody libraries. *Nat Biotechnol* (2000) 18(1):75–80. doi:10.1038/71958
60. Sblattero D, Lou J, Marzari R, Bradbury A. In vivo recombination as a tool to generate molecular diversity in phage antibody libraries. *J Biotechnol* (2001) 74(4):303–15. doi:10.1016/S1389-0352(01)00022-8
61. D'Angelo S, Glanville J, Ferrara F, Naranjo L, Gleasner CD, Shen X, et al. The antibody mining toolbox: an open source tool for the rapid analysis of antibody repertoires. *MAbs* (2014) 6(1):160–72. doi:10.4161/mabs.27105
62. Colby DW, Kellogg BA, Graff CP, Yeung YA, Swers JS, Wittrup KD. Engineering antibody affinity by yeast surface display. *Methods Enzymol* (2004) 388:348–58. doi:10.1016/S0076-6879(04)88027-3
63. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113(19):E2636–45. doi:10.1073/pnas.1525510113
64. DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. A public database of memory and naive B-cell receptor sequences. *PLoS One* (2016) 11(8):e0160853. doi:10.1371/journal.pone.0160853
65. Khan TA, Friedensohn S, de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) 2(3):e1501371. doi:10.1126/sciadv.1501371
66. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC. Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires. *Curr Opin Chem Biol* (2015) 24:112–20. doi:10.1016/j.cbpa.2014.11.007
67. Busse CE, Czogiel I, Braun P, Arndt PF, Wardemann H. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur J Immunol* (2014) 44(2):597–603. doi:10.1002/eji.201343917
68. Mathonet P, Ullman CG. The application of next generation sequencing to the understanding of antibody repertoires. *Front Immunol* (2013) 4:265. doi:10.3389/fimmu.2013.00265
69. Kaplinsky J, Li A, Sun A, Coffre M, Koralov SB, Arnaout R. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc Natl Acad Sci U S A* (2014) 111(25):E2622–9. doi:10.1073/pnas.1403278111
70. Kaplinsky J, Arnaout R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* (2016) 7:11881. doi:10.1038/ncomms11881
71. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* (2011) 6(8):e22365. doi:10.1371/journal.pone.0022365
72. Tsioris K, Gupta NT, Ogunniyi AO, Zimnisky RM, Qian F, Yao Y, et al. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integr Biol (Camb)* (2015) 7(12):1587–97. doi:10.1039/c5ib00169b
73. Jackson KJ, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* (2014) 16(1):105–14. doi:10.1016/j.chom.2014.05.013
74. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* (2013) 5(171):171ra19. doi:10.1126/scitranslmed.3004794
75. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci U S A* (2014) 111(6):2259–64. doi:10.1073/pnas.1317793111
76. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA, et al. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol* (2014) 15:40. doi:10.1186/s12865-014-0040-5
77. Snir O, Mesin L, Gidoni M, Lundin KE, Yaari G, Sollid LM. Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J Immunol* (2015) 194(12):5703–12. doi:10.4049/jimmunol.1402611
78. DeFalco J, Harbell M, Manning-Bog A, Baia G, Scholz A, Millare B, et al. Non-progressing cancer patients have persistent B cell responses expressing shared antibody paratopes that target public tumor antigens. *Clin Immunol* (2017). doi:10.1016/j.clim.2017.10.002
79. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight

into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106(48):20216–21. doi:10.1073/pnas.0909775106
80. Fantini M, Pandolfini L, Lisi S, Chirichella M, Arisi I, Terrigno M, et al. Assessment of antibody library diversity through next generation sequencing and technical error compensation. *PLoS One* (2017) 12(5):e0177574. doi:10.1371/journal.pone.0177574
81. Edwards BM, Barash SC, Main SH, Choi GH, Minter R, Ullrich S, et al. The remarkable flexibility of the human antibody repertoire; isolation of over one thousand different antibodies to a single protein, BLyS. *J Mol Biol* (2003) 334(1):103–18. doi:10.1016/j.jmb.2003.09.054
82. Marks JD, Hoogenboom HR, Bonnert TP, McCafferty J, Griffiths AD, Winter G. By-passing immunization. Human antibodies from V-gene libraries displayed on phage. *J Mol Biol* (1991) 222(3):581–97. doi:10.1016/0022-2836(91)90498-U
83. Gram H, Marconi LA, Barbas CF III, Collet TA, Lerner RA, Kang AS. In vitro selection and affinity maturation of antibodies from a naive combinatorial immunoglobulin library. *Proc Natl Acad Sci U S A* (1992) 89(8):3576–80. doi:10.1073/pnas.89.8.3576
84. Griffiths AD, Williams SC, Hartley O, Tomlinson IM, Waterhouse P, Crosby WL, et al. Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J* (1994) 13(14):3245–60.
85. Perelson AS, Oster GF. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* (1979) 81(4):645–70. doi:10.1016/0022-5193(79)90275-3
86. Boder ET, Wittrup KD. Yeast surface display for directed evolution of protein expression, affinity, and stability. *Methods Enzymol* (2000) 328:430–44. doi:10.1016/S0076-6879(00)28410-3
87. Brochet X, Lefranc MP, Giudicelli V. 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* (2008) 36(Web Server issue):W503–8. doi:10.1093/nar/gkn316
88. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41(Web Server issue):W34–40. doi:10.1093/nar/gkt382
89. Ippolito GC, Hoi KH, Reddy ST, Carroll SM, Ge X, Rogosch T, et al. Antibody repertoires in humanized NOD-scid-IL2Rgamma(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS One* (2012) 7(4):e35497. doi:10.1371/journal.pone.0035497
90. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* (2006) 34(Database issue):D781–4. doi:10.1093/nar/gkj088

Conflict of Interest Statement: SD, FF, LN, FE, and AB are employees of Specifica Inc. The remaining 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.

Copyright © 2018 D'Angelo, Ferrara, Naranjo, Erasmus, Hraber and Bradbury. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

REVIEW
published: 14 March 2018
doi: 10.3389/fimmu.2018.00462

Analyzing Immunoglobulin Repertoires

*Neha Chaudhary and Duane R. Wesemann**

Division of Rheumatology, Department of Medicine, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

Somatic assembly of T cell receptor and B cell receptor (BCR) genes produces a vast diversity of lymphocyte antigen recognition capacity. The advent of efficient high-throughput sequencing of lymphocyte antigen receptor genes has recently generated unprecedented opportunities for exploration of adaptive immune responses. With these opportunities have come significant challenges in understanding the analysis techniques that most accurately reflect underlying biological phenomena. In this regard, sample preparation and sequence analysis techniques, which have largely been borrowed and adapted from other fields, continue to evolve. Here, we review current methods and challenges of library preparation, sequencing and statistical analysis of lymphocyte receptor repertoire studies. We discuss the general steps in the process of immune repertoire generation including sample preparation, platforms available for sequencing, processing of sequencing data, measurable features of the immune repertoire, and the statistical tools that can be used for analysis and interpretation of the data. Because BCR analysis harbors additional complexities, such as immunoglobulin (Ig) (i.e., antibody) gene somatic hypermutation and class switch recombination, the emphasis of this review is on Ig/BCR sequence analysis.

Keywords: B cell repertoire, next-generation sequencing, statistical analysis, immunoglobulin, repertoire

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Rachael Bashford-Rogers, University of Cambridge, United Kingdom
To-Ha Thai, Harvard Medical School, United States

***Correspondence:**
Duane R. Wesemann
dwesemann@bwh.harvard.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *07 December 2017*
Accepted: *21 February 2018*
Published: *14 March 2018*

Citation:
Chaudhary N and Wesemann DR (2018) Analyzing Immunoglobulin Repertoires. Front. Immunol. 9:462. doi: 10.3389/fimmu.2018.00462

INTRODUCTION

Analysis and interpretation of antibody repertoire data require an understanding of the complex processes of somatic antigen receptor gene dynamics. Antibodies are composed of a combination of two identical heavy (H) and two identical light (L) immunoglobulin (Ig) chains, each with variable (V) and constant (C) regions. The IgH V-region is encoded by an exon that is generated somatically from assembly of three gene segments, named variable (also abbreviated as V, not to be confused with the V segment-containing V exon), diversity (D), and joining (J) gene segments. The IgH locus contains many related, but distinct V_H, D_H, and J_H gene segments, which are genomically organized in tandem and selected in a semi-random process for somatic V(D)J assembly in bone marrow progenitor (pro-) B cells. There are two IgL loci—namely, Igκ and Igλ—which have their own pools of tandemly arranged V_L and J_L gene segments that are assembled by VJ recombination in precursor (pre-) B cells after productive IgH assembly (1, 2). Non-templated (N) and palindromic (P) nucleotides are added to inter-segment junctions, further adding to the diversity. V(D)J recombination is

Abbreviations: Ig, immunoglobulin; BCR, B cell receptor; TCR, T cell receptor; AID, Activation-induced cytidine deaminase; CSR, class switch recombination; SHM, somatic hypermutation; GC, germinal center; UMIs, unique molecular identifiers; JSD, Jensen–Shannon divergence; KLD, Kullback–Leibler divergence; SK, Storer–Kim (KMS); KMS, Kulinskaya–Morgenthaler–Staudte; RDI, repertoire dissimilarity index; PCA, principal component analysis.

dependent upon Rag1 and Rag2, occurs at the IgH locus before the IgL loci, and Igκ is usually attempted before Igλ assembly. V(D)J recombination usually occurs in an allelically ordered way. In this regard, if a V exon assembly attempt does not result in a productive reading frame, a subsequent attempt occurs on the sister allele. This process results in B cells monoallelically expressing one B cell receptor (BCR) specificity, although rare cells expressing IgH from two alleles, as well as both Igκ and Igλ, have been observed as well (3, 4). Although the IgH and IgL alleles that assemble non-productively do not produce protein, they are transcribed to contribute to the mRNA pool of the cell. Non-productive Ig sequences that appear in sequence data sets can be identified as such in the data processing stage.

Productive assembly of both IgH and IgL chains results in IgM expression on the surface of immature B cells, forming the antigen-binding part of the BCR. Mature naïve B cells express both IgM and IgD due to alternative C_H splicing of C_μ and C_δ. Upon activation, B cells can undergo two other forms of diversification, both initiated by activation-induced cytidine deaminase (AID). DNA cleavage and repair events can result in IgH class switch recombination (CSR), where removal of C_H region DNA positions alternative C_Hs (e.g., Cγ, Cε, Cα) downstream of the V exon. AID is also required for V exon somatic hypermutation (SHM), which typically occurs in activated germinal center (GC) B cells (5, 6). B cells can further differentiate into BCR-expressing memory B cells, or antibody-secreted plasma cells (7).

While the actual BCR diversity is not completely defined, estimates of the theoretical diversity enabled by V(D)J recombination number more than 10^{13} different potential specificities (8). In addition, only 2% of the BCR repertoire is accessible in circulation at any given time (9). The high diversity and the accessibility limitations constrain our ability to measure and analyze the human immune repertoire. Moreover, what can be learned from deep Ig sequencing is highly dependent upon sample preparation and statistical analysis utilized. In this context, various methods have been described for Ig library preparation and sequencing, and there are numerous statistical tools that have been applied to data analysis (**Figure 1**). Here, we will briefly review Ig library preparation and sequencing platforms and provide a more in-depth treatment of available analysis tools.

LIBRARY PREPARATION

Sample library preparation involves the isolation and amplification of the target nucleic acid fragments for sequencing. There are two starting materials that can serve as the initial template to sequence Ig repertoires—genomic DNA (gDNA) and mRNA. Use of gDNA as a template has the advantages of the superior stability of DNA over RNA and the fact that the initial Ig gene copy number is constant between cells. The use of mRNA as an initial template requires an additional step to convert RNA to DNA *via* reverse transcription (RT). Unique Molecular Identifiers (UMIs) can be added to cDNA molecules at this step. UMIs are randomly generated sequences of specific length (usually between 8 and 22 nt) designed to mark individual molecules. These help identify PCR repeats in the analysis, as all repeats from single mRNA will have same UMI. Using mRNA as a template also has the advantage of being intronless, enabling the sequencing of both V and C regions in the same sequence read fragment. Because the number of mRNAs per cell is much higher than DNA copies, the copy number per cell overestimates the number of cellular clones. Despite these disadvantages, the greater mRNA copy number per cell enhances sequence coverage and allows variable and constant region information to be captured on the same length of read (10).

A key objective of techniques designed so far in deep sequencing of Ig repertoires has been to exhaustively amplify the Ig repertoire with minimum error and bias. Primer selection, especially at the 5′ V-region end, is a crucial step to this process as there are many dozens of V gene segments. Some approaches use a mixture of degenerate V_H family primers (frame work region 1) as forward primers and a mix of J segment or C region reverse primers. Using a mixture of primers may lead to biases in priming and amplification. Furthermore, SHM-mediated sequence differences may also contribute to unwanted bias (11). The use of synthetic repertoires as control templates to identify and remove potential bias at the analysis stages have been used as an approach to address the problem of primer bias for T cell receptor (TCR) sequencing (12). Another way to reduce primer bias is with the use of 5′ adaptor sequences. This can be done by attaching an oligonucleotide to the 5′ of Ig mRNA molecules by RNA ligation, or by 5′ rapid amplification of cDNA ends (5′ RACE). This enables the attachment of a known sequence to the 5′ end, for use in subsequent PCR amplification steps (13). This approach requires only one set of gene-specific primers targeting the less variable J or C region sequences at the 3′ end. However, 5′ RACE is less able to represent the richness of the sample due to lower efficiency of sequence capture compared to direct priming. The bait capture method uses polyA and part of the sequence of interest attached to streptavidin magnetic beads to isolate the Ig mRNA. The beads are then washed, and the hybridized fragments eluted for sequencing (10). A more recent method called linear amplification-mediated high-throughput genome-wide translocation sequencing (LAM-HTGTS) uses translocation specific sequence at the 3′ end of J region to capture and isolate the complete V(D)J sequence from the gDNA after DNA fragmentation *via* sonication (14). Random fragmentation used with LAM-HTGTS risks losing rare clones. Direct comparison of multiplex PCR, RACE, and bait capture methods for Ig repertoire sequencing showed that these methods were generally concurrent (10).

Errors may be introduced into the sequence at several steps, including RT, PCR amplification, or during sequencing due to incorrect base call (15, 16). To control for errors that occur during PCR amplification, the UMI can be used to create a consensus sequence of PCR repeats (**Figure 2A**). A number of UMI-based methods have been devised to improve sequence quality (**Figures 2B–D**) or identify PCR bias (**Figure 2E**)—discussed here.

The Molecular Identifier Group based Error Correction (MIGEC) groups similar sequences with same UMI and uses a set of rules to predict errors (17). One rule is to identify a consensus sequence based on the most common variant within a UMI group. However, if the porportions of mismatches are such to evade consensus, the sequence is dropped. A problem with this is that an early error during library preparation could

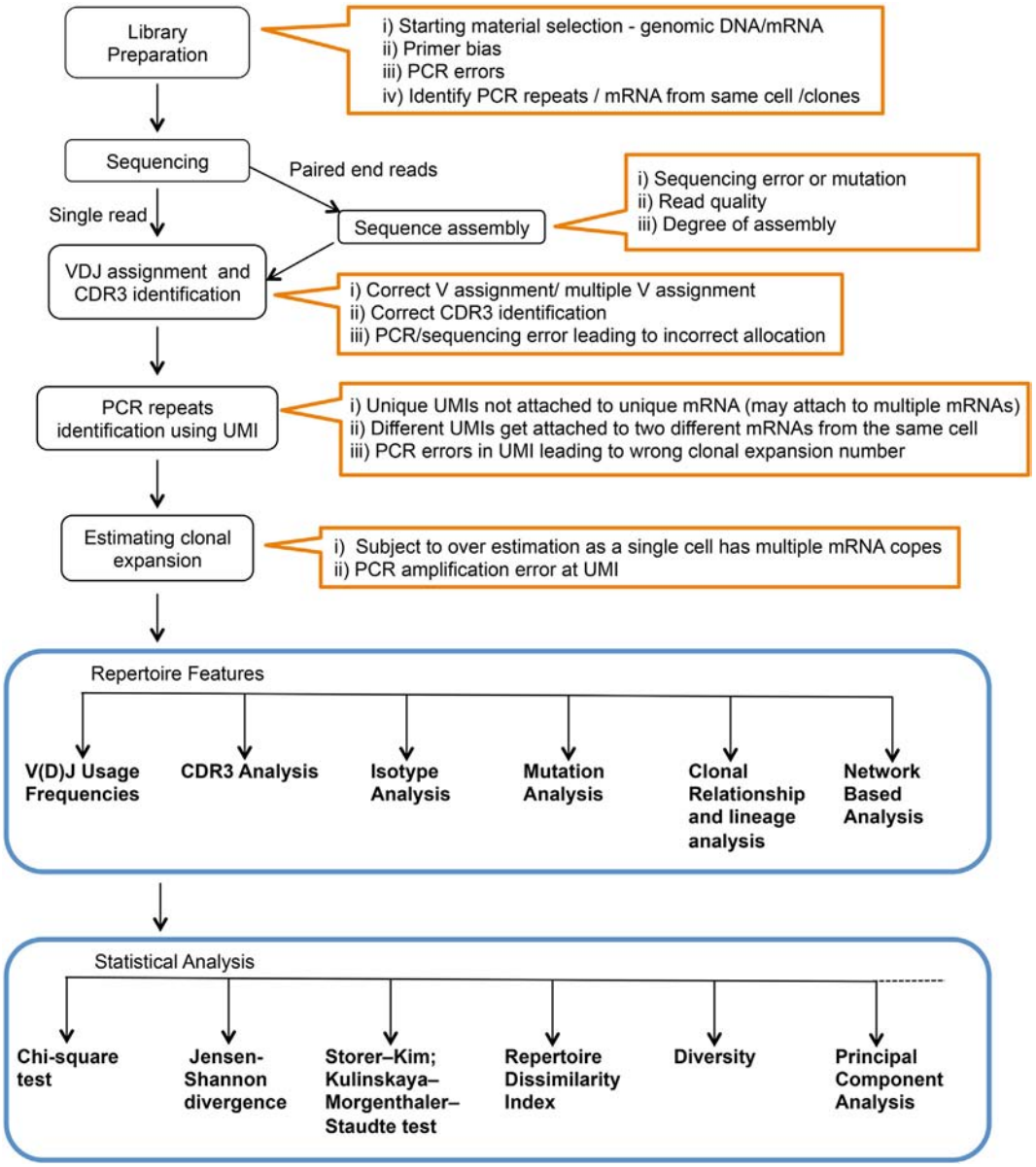


FIGURE 1 | Complete workflow for high-throughput sequencing and analysis of the immunoglobulin repertoire. Text within orange outlines the complications at each step.

provide a consensus that does not reflect the original template. To solve this, discarded sequences are assessed for PCR error hot spot locations. Sequences with changes within identified error hotspots can then be reevaluated (**Figure 2A**).

Duplex Sequencing adds UMI to both ends of the sequence and then sequences both strands separately (18). A mismatch has to be present in both the strands to be considered a true mutation (**Figure 2B**). Another method uses paired-end sequencing wherein both the forward and reverse strands are sequenced after adding a single UMI (19). Errors are removed for both the strands separately and they are overlapped to get the complete sequence (**Figure 2C**).

Another system uses a sequence target for Tn5 transposase attached to the forward or reverse primer. This allows random insertion into the UMI-containing sequence library (20). The complete sequence and the Tn transposase-foreshortened sequences can be overlapped to get the consensus sequence with less chances of error (**Figure 2D**). In molecular amplification fingerprinting (MAF), a reverse UMI (RUMI) is added at the RT step and a forward UMI (FUMI) is added with each PCR cycle keeping a track of the number of PCR cycles and PCR bias toward different sequences (**Figure 2E**) (21). The utility of each of these methods depends on the question under study. The most commonly applied methods of the five are MIGEC and paired-end

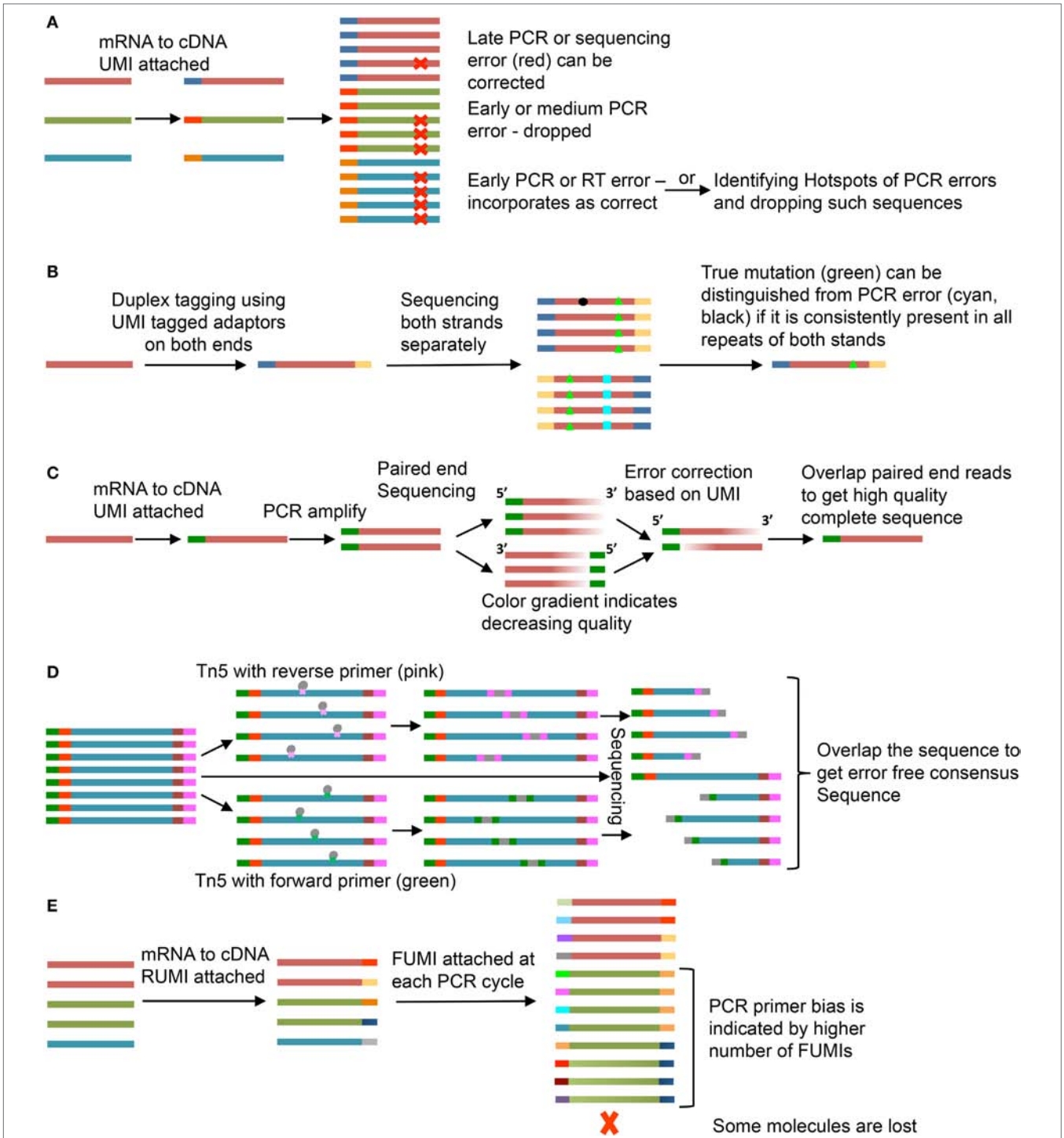


FIGURE 2 | Use of unique molecular identifiers (UMIs). Each strand is an mRNA or a cDNA and smaller bars are UMIs. Same color of the strand and bar represents copies of same mRNA and UMI, respectively. **(A)** Molecular Identifier Group based Error Correction (MIGEC) (17). Among all sequences with same UMI, only few have error (late PCR error) (red), the error is identified and removed; if near 50% of the sequences have the same error, the sequence is dropped; an early error (present in most sequences) would be unidentifiable but it is dropped if it falls on a PCR hotspot. **(B)** Duplex Sequencing (18). UMIs are added to both ends of the sequence and both strands are sequenced. If a mutation (green, black, or cyan) is present in only one of the two stands, it is an error. **(C)** Paired-end sequencing is done after UMI tagging. Error corrections are done for individual reads and then they are merged to get the full good quality sequence (19). **(D)** Tn5-enabled molecular identifier-guided amplicon sequencing (TMIseq) (20). The PCR amplified libraries are tagmented using Tn5 transposase where either forward (green) or reverse (pink) primer is inserted. Thus, only part of the sequence containing both forward and reverse primers gets amplified for sequencing. Both, the smaller libraries and the complete sequence library are sequenced and used to generate a consensus error-free sequence. **(E)** Molecular amplification fingerprinting (MAF) (21). A reverse UMI (RUMI) is added at the reverse transcription (RT) step and a forward UMI (FUMI) is added at each subsequent PCR amplification step. FUMIs keep track of PCR bias for different sequences. Some sequences are over amplified while some may be lost in the process.

sequencing. These are the simplest in terms of sequencing and preprocessing steps. If a more stringent analysis of SHM has to be done, Duplex Sequencing and Tn5 transposase method would be expected to offer increased accuracy. In case of MAF, addition of a FUMI at each PCR step would lead to gradual increase in length accompanied by reduced quality at the RUMI sequence site but can be used to understand PCR bias and loss due to random subsampling during sequencing.

Unique molecular identifier length affects the analysis results. Shorter UMIs lead to more non-unique attachment, where the same UMI sequence gets attached to different template molecules. Longer UMIs increase the risk of primer dimer formation and have higher chances of error during amplification and sequencing, which may lead to inflation, misinterpretation, and/or mismatch (22, 23). A UMI length of 8–12 nucleotides is most recommended. Assumptions usually held in the analysis are that UMIs are uniformly represented and all templates uniformly tagged. In practice, however, different target templates have been observed to attach to identical UMI sequences (24). Even with different methods being applied to overcome these issues (23, 25), the impact of erroneous barcodes (**Figure 3**) may not be trivial (26). We favor an approach of identification of PCR repeats by using both UMI and sequence information (with 1–2 nucleotide error).

PCR/primer bias for certain templates can complicate assignment of repeat sequences (27). In addition, different B-lineage cell populations can produce widely different amounts of Ig mRNA molecules per cell. In this regard, an activated B cell or plasma cell has a much higher copy number of mRNA than a naïve or memory B cell (28). Assigning identical Ig sequences to clonal expansion versus copies per cell typically requires single-cell sequencing. In addition, IgH and IgL can be paired accurately in single-cell sequencing. A growing number of single-cell sequencing techniques for Ig and TCR repertoire analysis are becoming available. These usually entail an initial barcoding step before amplification and sequencing. Summaries of high-throughput single-cell sequencing approaches are shown in the **Figure 4**.

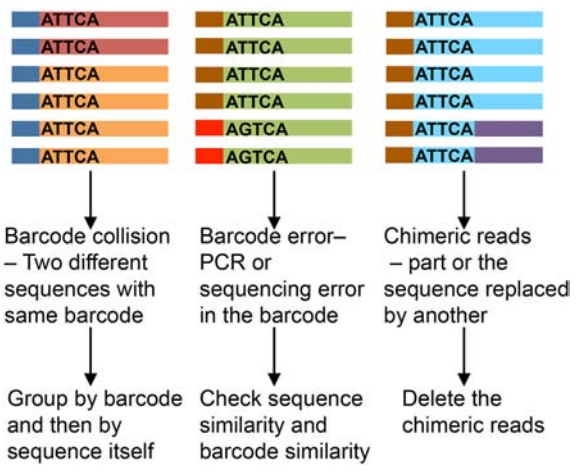


FIGURE 3 | Impact of erroneous barcodes (25). Each strand represents a mRNA. The bar at the end represents a unique molecular identifier (UMI). Same color of the strand and bar represents copies of same mRNA and UMI, respectively. The sequence of the UMI is mentioned within each strand.

SEQUENCING PLATFORMS

A number of sequencing platforms are available that differ in features like read length or the coverage of Ig gene, sequencing depth, cost, and run time (**Table 1**). The PacBio platform, due to its long read length, enables the amplification of H and L chains physically linked together, but is limited due to high error rate, high cost, and low reads per run. Illumina HiSeq offers the highest read depth, but at a cost of read length. **Table 1** illustrates the most commonly used platforms along with some of the important features. Larger read number provides higher coverage of a particular sequence giving greater chances of error correction in sequence. Some platforms also provide the feature of paired-end sequencing, in which sequencing is done from both ends of the DNA amplicon, and the final sequence is obtained by merging the two paired-end reads. This ensures superior read quality compared to single end sequencing. Illumina and Ion torrent provide paired-end sequencing. Choice of sequencing platform depends upon the research goals and experimental questions.

INITIAL PROCESSING AND ANNOTATION

The output for each of these platforms is a binary file format: standard flowgram format (.sff—Roche's 454 GS FLX), base call (.bcl—illumina), and Binary Alignment Map (BAM—PacBio). Ion torrent gives output in three formats—BAM, FASTQ, or VCF. Each of these has to be converted to Fasta or Fastq format either by running scripts that are part of the software platform (sffinfo-Roche; bcl2fastq-Illumina) or by using one of the many freely available scripts (bamtoFastq, sff_extract). Fasta and Fastq are the two common input formats for most analysis programs. Fasta format consists of a list of sequences with a unique identification tag preceding each sequence. Fastq files (34) also include the information regarding the quality of each residue in the sequence in the form of a Phred score (*Q* score). The *Q* score gives an estimated probability of error for each nucleotide position. They are encoded in the form of ASCII characters, which can be transformed into integers.

Once the data are available from the sequencing reaction, initial processing (often termed "preprocessing") of the sequences is necessary prior to annotation. Preprocessing includes filtering out low quality sequences, sequence trimming to remove continuous low quality nucleotides, merging paired-end sequences and, if possible, identifying and filtering out PCR repeats. The quality of the output sequences from various platforms is such that with increase in length from the 5′ toward the 3′ end, the quality of residues deteriorates. With Ig sequences, it is important to identify the mutations from sequencing errors. Thus, low quality residues, usually those with a *Q* score <30, at the 3′ end are excluded. In the case of paired-end sequencing, regions of sequence that are included in both reads (i.e., overlapping regions) can be used to form a consensus based on *Q* scores derived from both reads. Sequences with very long stretches of poor quality and paired-end

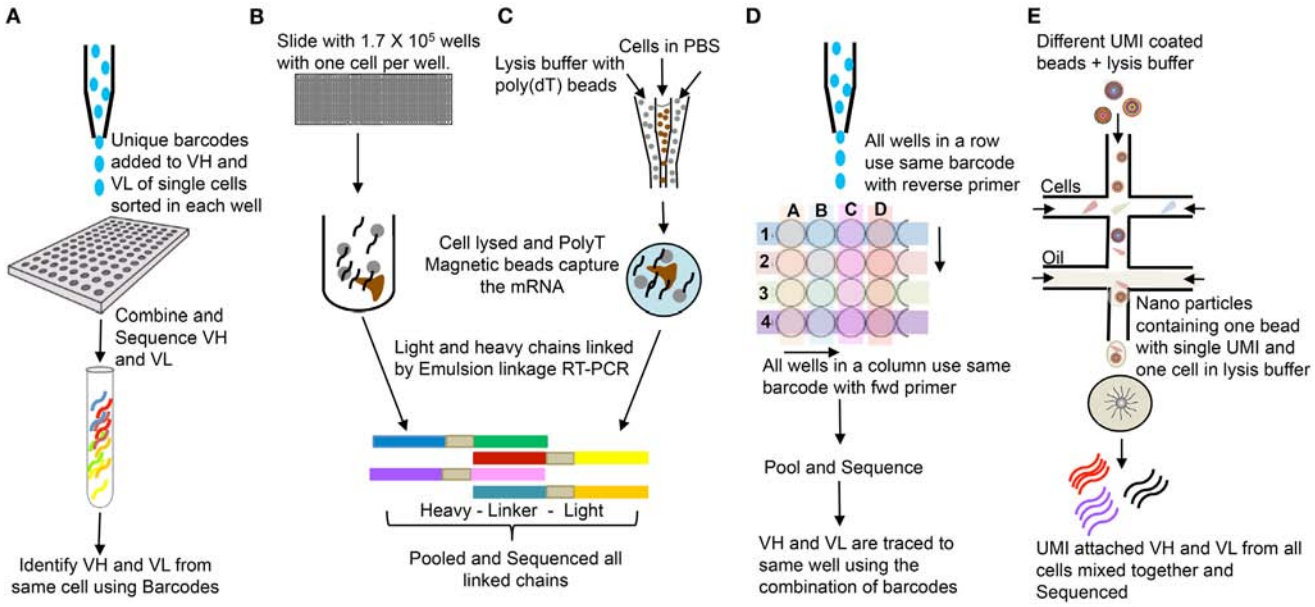


FIGURE 4 | Single cells bulk sequencing: **(A)** Single cells are sorted in 96-well plates, and VH and VL are tagged with cell specific unique molecular identifier (UMI). Sequences from all cells are pooled together and sequenced (29). **(B)** Single cells are isolated in polydimethylsiloxane slides (1.7×10^5 wells/slide-56-μm diameter wells); poly(dT) microbeads are added; wells are sealed with dialysis membrane and equilibrated with lysis buffer; VH and VL mRNAs get attached to poly(dT) beads; beads are emulsified for cDNA synthesis; linkage PCR generates paired VH:VL products which are pooled together and sequenced (30). **(C)** Single cells and poly(dT) magnetic beads are trapped into emulsions along with lysis buffer. VH and VL mRNAs annealed to poly(dT) beads and sequenced as in **(B)** (31). **(D)** Single cells are sorted in 384-well PCR plates. Instead of unique UMI for each cell, each row and column has unique UMIs attached to respective forward and reverse primers, which help trace back to the wells (32). The DNA is pooled and sequenced. **(E)** Microfluidic device joins two aqueous flows into distinct droplets: one with cells and other with barcoded primer beads in lysis buffer. The cell is lysed and its mRNAs hybridizes to the primers on the microparticle surface. The microparticles are collected, washed, and the mRNAs are reverse transcribed, each with unique UMI from the beads. They are pooled and bulk sequenced together (33).

TABLE 1 | Common platforms used for immunoglobulin repertoire sequencing.

Platforms	Roche's 454 GS FLX	Illumina MiSeq	Illumina HiSeq	PacBio	Ion torrent
Mechanism	Pyrosequencing	Dye terminator sequencing	Synthesis (fluoresces attached to nt is excited and detected after each run)	Synthesis (florescence tag attached to phosphate chain)	Synthesis (detect H^+)
Read length	700 bp	300 × 2	250 × 2	860–1,100	>100
Run time	18–20 h	26 h	8 days	0.5–2 h	2 h
Reads/run	1M	3.5M	2B	0.01M	60–80M
Error rate (%)	1	~0.1	~0.1	~13	~1
Type of errors	Indel	Substitution	Substitution	Indel	Indel
Cost/mbp ($)	12.40	0.74	0.10	11–180	<7.5
Region of antibody covered	FWR1-CR	FWR1-CR	FWR1-CR	Amplification of linked H and L chains	FWR3 to CR

sequences with no overlapping regions are excluded. High-quality filtered and merged sequences can be grouped based on common UMIs (if available from the library preparation), as discussed above, to filter out PCR repeats. In addition, appropriate steps have to be taken to remove sequences with barcode error and remove chimeric reads (25).

Most analysis methods use alignment of the sequence with the germline to assign the respective V, D, and J segments. IMGT database (35) is the most extensively used database for germline Ig sequences. IMGT (36) and IgBlast (37) are the most common annotation software packages, and both use the IMGT database to align sequences. Though alignment with germline seems straightforward, the presence of SHM can make identification problematic as some V gene segments are very homologous and differentiating between allelic differences in the germline and somatically generated mutations may not be straight forward. Also, Ds and Js are small and have insertions and deletions as a result of V(D)J recombination. In many cases, the D segment remains unidentified due to its small size or several can multimerized in tandem (38). Accuracy of gene segment identification depends upon completeness of the reference germline databases. Humans and mice have the most well defined Ig gene loci, but a map of all allelic variants is not complete (39). There have been efforts to address this with algorithms—such as TIgGER (40),

IgDiscover (39), IMPre (41), and a more recent allele prediction and validation tool (42)—that can be used to identify germline alleles for individual repertoires. Proper identification of non-template additions and deletions depends to a large degree on the accuracy and completeness of the reference database used.

Apart from IMGT and IgBlast, other software programs are available for analysis of the BCR and TCR repertoire data. A number of them also include preprocessing, annotation, and statistical analyses all in a single pipeline. Some of these programs along with their features are listed in **Table 2**.

DESCRIBABLE FEATURES OF B CELL REPERTOIRES

The expansive capacity of the BCR repertoire makes the probability of finding the same sequence within two individuals and even within two tissues of same organism extremely low, and this limits direct comparisons of specific sequences between individuals. However, it has recently been shown that human TCR repertoires can be grouped into functionally related categories that can be shared between individuals (73). The same algorithm, called GLIPH (Grouping of Lymphocyte Interactions by Paratope Hotspots), could also be used to group functional BCR repertoire but would have to include the additional complexity due to SHM. A number of other features have been used to quantitatively compare antigen receptor repertoires between individuals, groups, or experimental conditions. Below, we provide a brief survey of measurable repertoire features and some representative studies that have assessed them in the context of a variety of lines of inquiry.

V(D)J Segment Usage Frequencies

An Ig repertoire can be described in terms of the frequencies with which it uses the gene segments that make up the V exon, particularly the V segment, as it is the longest and most diverse. V gene segment frequencies, or VJ combinations frequencies, have been used to compare stages of immune responses, for example, to describe differences in B cell repertoires of avian flu (H7N9) patients at the time of infection and during recovery, where recovery was shown to utilize more diverse VJ combination frequencies (74). V gene usage frequency comparisons have also been used to describe age-related changes (75) as well as general population level descriptions (39, 40, 76).

Complimentary Determining Region 3 (CDR3) Properties

The CDR3 is the most variable region of an antibody and can be used to define clonal lineages. The CDR3 length and amino acid properties have been used to characterize a functional repertoire. The advantages and methods of CDR3 comparisons are reviewed elsewhere (77). There are many studies comparing CDR3 features in repertoire analysis. Comparisons of CDR3 lengths between cell groups expressing different IgH isotypes showed that IgM had longer CDR3s compared to all other isotypes examined (11), suggesting a potentially interesting link between a general V-region feature and IgH isotype. An analysis of BCR repertoire of naïve, IgM memory, and class switched memory B cells suggest that memory B cells may have shorter CDR3s with more positively charged amino acids. It was also found that IgM memory cells may have lower hydrophobic and aliphatic indexes compared to memory cells of other IgH isotypes (78). Antigen-experienced B cell repertoires appear to have a more exposed CDR3 region rich in charge (79). Antigen exposure also appears to be associated with a decrease in CDR3 length (80). IgM and IgA CDR3s tend to be longer with age (81). Systemic lupus erythematosus (SLE) patients were reported to have shorter CDR3 with higher arginine content (82).

Complimentary determining region 3 analysis also helped identify the "public" sequences. Public CDR3 (or public Ig) is a term used when similar or identical sequences are found in different individuals. They are usually reported in individuals who had been exposed to the same pathogen, like *Haemophilus influenzae* type B, tetanus toxoid, and influenza (83, 84). Public sequences are more common for IgL as compared to IgH (79). The public BCRs have also been observed in persistent diseases like autoimmunity and cancer (85). Understanding emergence of public CDR3s could help understand the process of affinity maturation and antibody development (86).

Mutation Analysis

Diversity due to somatic mutation is also a feature of the Ig repertoire. This includes insertions and deletions during V(D)J recombination and SHM. During SHM, AID targets at DGYW motifs (D = A/G/T, Y = C/T, W = A/T) (87, 88), which are also referred to as mutational hotspots. In general, mutations are analyzed as degree of divergence from germline sequences and give insight into the biological process of SHM and affinity maturation. Any nucleotide mutation can result in a different amino acid encoded at that position (replacement) or can result in no change (silent). Analysis of replacement versus silent mutation status at nucleotide positions can have implications for studies examining positions important for antibody selection (53, 89).

Somatic hypermutation analysis in twins has shown that genetic factors play a role in determining mutation frequency (90). Similar analysis showed that the level of SHM is reduced in older individuals (81). AID-mediated mutations tend to occur unequally across the V exon. CDRs have more hotspots and tend to mutate more than FWRs. Also, mutation selection pressure is different for the two regions. Mutations in the FWRs are more likely to be selected against, as these regions are important for structural fitness (91). Insertions and deletions occur during SHM, adding to the structural plasticity of the antibodies, but are relatively rarely found as they are more likely than mutations to cause negative selection from structural instability (92).

Somatic hypermutation studies have been employed to decipher why Ig loci are permissive for AID-mediated mutation compared to off target, non-Ig loci. This remains one of the most elusive questions in B cell biology. Studies examining a particular V gene segment in which certain AID-target hot spots were experimentally removed in a mutating human B cell line suggested that local sequence context may influence SHM of other

TABLE 2 | Softwares available for sequence error correction, annotation, and analysis of immunoglobulin (Ig) repertoire.

Name	Platform/ availability	Input format	Maximum sequence limit	Features	Reference
IMGT/V-QUEST IMGT/HighV-Quest	Online Online	Fasta Fasta	50 150,000	V(D)J Annotation, junction analysis; mutation; amino acid statistics; comparisons between two repertoires	(36, 43, 44) (45–48)
JOINSOLVER	Online/standalone	Fasta	–	Annotation; complimentary determining region 3 (CDR3); mutation; insertion deletion in human only	(49)
VDJSolver	Online	Fasta	500	Use hidden Markov model (HMM) or maximum likelihood to prediction V(D)J recombination	(50)
iHMMune-align	Online/standalone	Fasta		HMM to model the processes involved in human IGH gene rearrangement and maturation	(51)
VDJFasta	Standalone	Fasta	–	HMM-based CDR identification; translation and alignment; probabilistic germline classification	(52)
BASELINe	Online/standalone	Fasta	–	Quantifying selection based on somatic hypermutation (SHM) patterns	(53)
IgAT	Standalone (windows)	IMGT output files	150,000	Gene segments usage; CDR3; antigen selection based on SHM; the hydrophobicity of antigen-binding sites; structural properties of the CDR-H3 loop using Shirai's H3-rules	(54)
IgBlast	Online/standalone	Fasta	Online-1,000/ SA-none	V(D)J assignment; CDR3 identification; mutation; can use custom database in SA	(37)
pRESTO	Standalone	Fastq/Fasta	None	Merge; filter; error correction (with/without UMIs); annotation	(55)
Vidjil	Online/standalone	Fastq/Fasta	None	Extract V(D)J junctions; clonality	(56, 57)
The antibody mining toolbox	Standalone	Fastq	None	Analysis based on CDR3 as sequence identifiers	(58)
MIGEC	Standalone (Unix)	Fastq	None	Error correction and sequence assembly	(17)
IgRepertoireConstructor	Standalone	Fastq	None	Merge; filter; error correction (with/without UMIs); validation using mass spec; clonality; diversity	(59)
MiXCR	Standalone	Fastq	None	Merge; filter; PCR error correction; annotation; Gene segment usage; clonality; mutation	(60)
IMonitor	Standalone	Fastq/Fasta	None	Merge; filter; V(D)J assignment; gene usage frequency; CDR3; mutation; insertion and deletion	(61)
IgSCUEAL	Standalone	Fasta	None	V J annotation based on phylogeny; gene usage frequency; CDR3 length	(62)
Change-O	Standalone	IMGT/IgBlast Result	None	Gene usage; clonality; CDR3; diversity; phylogenetic; mutation; selection pressure; novel germline prediction	(63)
TIgGER	Standalone	Fasta	–	Predicts germline alleles	(40)
LymAnalyzer	Standalone	Fastq	None	V(D)J identification; CDR3; diversity; mutation; polymorphism analysis	(64)
sciReptor	Standalone	SFF/Fastq/Fasta	2,500	Single-cell analysis, annotation; maintains regional database; gene segment usage; clustering; mutation	(65)
repgenHMM	Standalone	Fasta	None	Predicts scenarios of V(D)J recombination	(66)
bcRep	Standalone (R)	IMGT output files	–	Gene usage frequency; clonality; diversity; mutations; repertoire comparison; visualization	(67)
IgDiscover	Standalone	Fastq	–	Identification of existing and novel germline V genes	(39)
Recon	Standalone	Frequency table (txt)	–	Diversity	(68)
IMPre	Standalone	Fasta	–	Predicts germline genes and alleles	(41)
ARResT/Interrogate	Standalone	IMGT output files	–	Calculation of statistics; visualization	(69)
Antigen Receptor Galaxy	Online	Fastq/Fasta	None	Demultiplex; annotation using IMGT/High V-Quest; V(D)J usage; SHM and CSR; Ag selection; clonality	(70)
IGoR	Standalone	Fasta	None	Calculates V(D)J recombination and mutation probabilities	(71)
ClonoCalc and ClonoPlot	Standalone	Fastq	–	GUI; Demultiplex; merge and annotate using MiXCR; analysis and plots using tcR package in R	(72)

regions within the V exon (93). Local sequence context was also shown to influence AID targeting on a passenger allele system, wherein a non-productive test allele was paired with a productive IgH knock-in to remove the effects of BCR-mediated cellular selection (88). DGYW motifs within CDR sequence regions were in general targeted more than DGYW motifs in framework regions (88). When the Ig passenger sequence was replaced with a non-Ig sequence, it was also targeted by AID, suggesting that the general location of the Ig V-region in the context of the IgH locus was an important feature of accessibility to SHM (88). This same passenger allele system was used to uncover sequence-intrinsic SHM-targeting rates of nucleotides across substrates representing maturation stages of an anti-HIV-1 broadly neutralizing antibody (94).

Isotype Analysis

Immunoglobulin repertoire analysis can provide insights into the biology of IgH isotypes. Each isotype has distinct biological functions governed by the C_H region domain. The sequences in a repertoire can be categorized into their respective isotypes if the experimental design accommodated for C region sequence in the library. Isotype analysis has included the categorization of Ig repertoire features, functions, or conditions to Isotype groups. As discussed above, sequencing data have shown that IgM is the least mutated and features the longest CDR3 in general compared to the other isotypes (11). Among memory cells, IgM has lower hydrophobic and aliphatic index compared to others (78), and SHM frequency has been reported to be higher in switched isotypes compared to IgM and IgD and varies between different subclasses of the same isotype (11). Isotype and SHM analysis has also been a key part of the concept of sequential switching. C_H regions for the various IgH isotypes are arranged in tandem along the IgH locus. Sequential switching occurs when CSR occurs first to Cμ-proximal C_H regions (e.g., to produce IgG3, IgG1, or IgA1), and then from these, to distally located isotypes (e.g., to IgG2, IgG4, or IgA2) (95). Studies have indicated that direct and indirect CSR can occur to distal isotypes (96, 97).

Clonal Relationship and Lineage Analysis

Lineage analysis and identification of clonal relationships between antibodies collected from an infected individual or during course of infection over time can track the evolutionary steps in the development of functional antibodies. This has been used in following HIV-1 bnAb VRC01 producing lineage for 15 years using peripheral B cell sampling for the rate of maturation and diversification in a single HIV-1-infected patient (98). A high substitution rate of 2 per 100 nucleotides per year resulted in extreme diversification in the context of chronic infection. Another study involving HIV-1 bnAbs found the intermediate antibodies to have reduced autoreactivity (99). PGT121-134 (100), PGT135-137 (16), and CH103 (101) are other bnAbs against HIV whose lineages have been studied in detail. Ig lineage and clone analysis has shown to have clinical relevance in the setting of lymphoma diagnostics. In this regard, lineage analysis at the time of diagnosis and relapse has revealed that B cells that reemerge are generally clonally related to the original cancer causing BCR (102, 103).

Network Based Analysis

A network is made from a group of entities (or nodes) connected to each other by links or edges if they share selected features. A B cell network may be based on mature antibody sequences clustered around the germline ancestor sequence. In this regard, all the nodes in a cluster would be the sequences identified to have come from that ancestor sequence, with edges connecting the nearest previous ancestor (**Figures 5A,B**). A healthy individual should have a very uniform network with each cluster of similar size and complexity (**Figure 5C**). An individual recently infected with a pathogen would have few expanded clusters corresponding to various versions of pathogen-reactive clones (**Figure 5D**). A uniformly distributed network versus a deformed network with few overly expanded V_H segments can identify chronic lymphocytic leukemia patients (104) (**Figure 5E**). A simpler network would be based on just the CDR3 region wherein homologous CDR3s are clustered together. Hepatitis B-infected patients harbor specific CDR3 sequences that may serve as identification signatures (105). General network properties—including reproducibility, robustness, and redundancy, have been studied for healthy Ig networks and can be evaluated *vis-a-vis* diseased Ig networks (106). The iGraph package in R can be used for network construction and visualization (107).

Paired Heavy and Light-Chain Analysis

Single-cell high-throughput sequencing of IgH and IgL together has been an important advance. With knowledge of IgH/IgL pairing, frequencies of paired usage of different V_H and V_L gene families can be determined together and a more authentic evaluation of antibody specificity can be achieved—as has been done in the evaluation of vaccine responses (30, 108, 109) as well as in autoimmune and inflammatory diseases (110). A comparison of single-cell sequences from naïve and antigen-experienced Ig repertoires uncovered several features related to how IgH and IgL pair together between these two groups (79). Single-cell sequencing can easily identify allelic inclusions, specifically noted by presence of both kappa and lambda light chains on the same B cell, as well as public V_H and V_L sequences. Single-cell sequencing has also shown that public V_Ls were able to pair with multiple V_H in multiple donors (31).

STATISTICAL ANALYSIS OF B CELL REPERTOIRES

Various forms of statistical tools have been applied on BCR and TCR sequences in a descriptive sense as well as to compare them in the context of experimental systems. Some or all of these methods can be used to describe and compare most of the BCR repertoire features discussed above. Below, we provide a brief survey of some of the analysis tools used in Ig repertoire studies.

Resampling

Resampling otherwise known as rarefication, or subset analysis, is a technique used to correct for differences in sequencing depth between samples. The sequencing reaction may generate more reads in certain libraries due to stochastic reasons and, depending

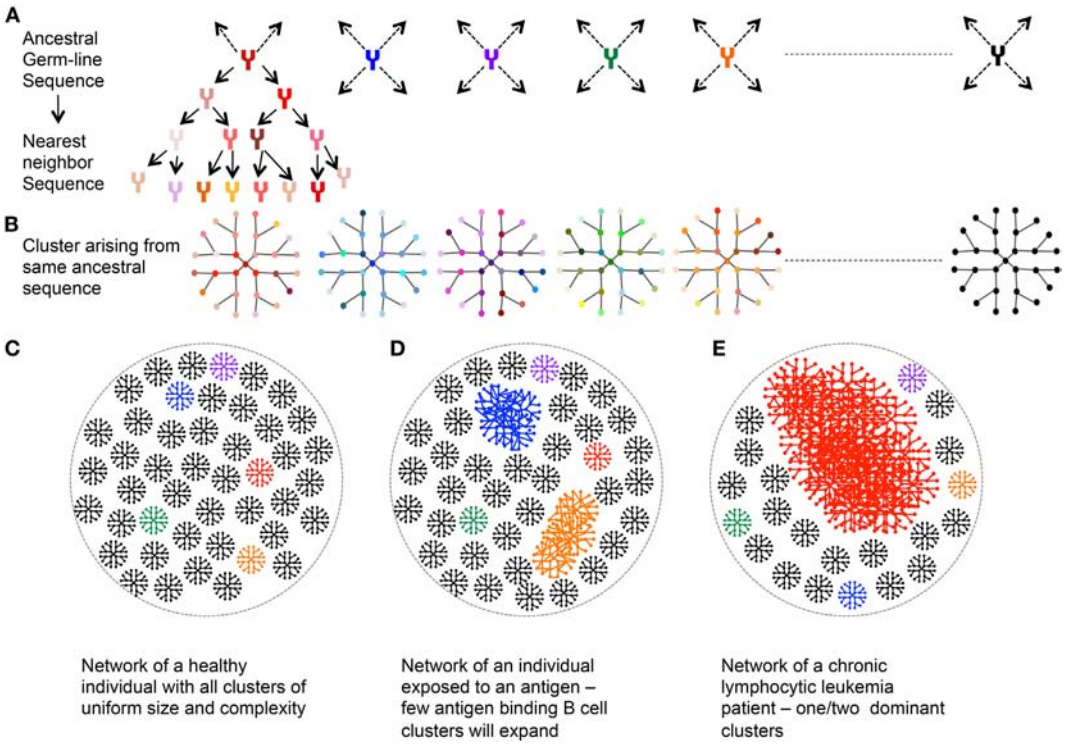


FIGURE 5 | Network analysis of immunoglobulin (Ig) repertoire—an explanatory model. **(A)** An example network arising from single germline sequence (Red). **(B)** Multiple clusters arising from different ancestral sequences. Each color represents cluster arising from different germline. **(C)** Representative network of a healthy individual: each cluster arising from an ancestral sequence is of uniform size and complexity. **(D)** Representative network of an individual exposed to an antigen: larger clusters represent the antibody, which recognizes the antigen and hence expands and mutates. **(E)** Representative Ig network of chronic lymphocytic leukemia patient with one dominant highly expanded cluster.

on how sequences are processed, has the potential to generate erroneous conclusions in the analysis. Subset sampling has been used in metagenomics studies where the number of sequences for all samples is reduced to the depth of that with the lowest read count. This step is designed to exclude any differences in the analysis that may be due to variable read depth, instead of the underlying biologic principle under investigation (84, 111, 112). However there are different views regarding use of rarefication. On one hand, subset analysis resolves randomly generated differences in sequence depth, but also results in discarding data, which leads to loss of assay power. This reduces the ability of finding difference between populations. In this light, it is important to run several control subsampling analyses to examine the degree to which the test subsamples reflect the properties of the whole. A sufficiently subsampled library from a whole library of sufficient depth should be essentially identical to the whole as well as other test subsets. Parallel comparisons of subsetted and whole data may be valuable to uncover read depth sufficiency. In general, we use subsetting when comparing averages of feature measurements from experiments repeated independently. If a test is used that considers only total counts (instead of averages of multiple experiments), such as the chi-square test, then we do not subset, as long as control comparisons of independently repeated tests indicate sufficient read depth of individual samples.

Chi-Square Test

Chi-square test for independence (113) checks if the proportions of two categorical variables are different from each other or not. It is a non-parametric test, which deals only with total counts—relative frequencies are not allowed. Here, the null hypothesis (H_0) states that the variables are independent while the alternate hypothesis states that they are dependent, i.e., knowledge of one variable can help predict the other variable. The test statistics for the Chi-square test is calculated as:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

where O is the observed frequencies; E is the expected frequencies, which, for each observation in the table is calculated as [(total observations in respective row)*(total observations in respect column)]/total number of observations in the table.

A limitation of the chi-square test is that it is extremely sensitive to sample size. The number of samples has to be large enough to have an expected value of at least 5 in each cell (113). Also, the test becomes more and more sensitive with increase in the sample size—eventually showing significance even with mild variation that can occur within assay error or repeat biological samples (114). This limits the use of chi-square test in high read

output platforms, such as the illumina systems. An example of deep sequencing data analyzed with chi-square test is in the comparison of V_H and V_L segment usage in developing B cells within weanling mouse bone marrow versus intestine. Chi-squared tests of pyrosequencing data showed significant differences in the V_L segment usage for the two groups, but not when randomly compared between biological repeats (115). Another such study used the chi-square test on Sanger sequencing data to compare the usage of V, D, and J segment families among patients with chronically evolving hepatitits C Virus (HCV) infection compared to resolved HCV, and healthy controls (116). They found that some of the families showed statistically significant association with the clinical groups for each of the three segments. HIV-1 specific and non-HIV-1 specific antibodies from an infected individual showed differences in the highly used V_H segments (117). A comparison of productive and non-productive antibody sequences revealed strong bias in the pairing of specific D and J segments due to multiple sequential D-to-J rearrangements (118). The function for calculation of chi-square statistics is available in the R package "stats."

Jensen–Shannon Divergence (JSD)

Jensen–Shannon divergence gives a measure of similarity between two probability distributions (119), and has also been used in Ig repertoire analysis. JSD is derived from Kullback–Leibler Divergence (KLD). For two probability distributions A and B, the JSD is calculated as:

$$JSD(A||B) = 0.5 * (KLD(A||M) + KLD(B||M)),$$

where $M = 0.5(A + B)$–midpoint of the two probability distributions and KLD($A||M$) and KLD($B||M$) are the KLD of A and M, and B and M, respectively. JSD is symmetric [JSD($A||B$) = JSD($B||A$)] and non-negative measure in contrast to KLD which is asymmetric [KLD($A||B$) ≠ KLD($B||A$)] and may be negative. JSD = 0, if $A = B$. JSD is also a non-parametric test. Since the test compares probability distributions of two populations, it is not affected by sample size. However, the effect of difference in sequencing depths leading to the differences in the probability distributions would still interfere with the results. Unlike KLD, it is symmetric, with values bound between 0 and 1 for both directions of comparison, which simplifies comparisons of multiple distributions.

Some studies, which included the use of JSD, calculated the distance between the repertoires under different conditions. JSD was used to compare TCR repertoires of cells with different epitope specificities (120). Ten epitope-specific TCR repertoires were characterized, and the JSD was used to compare gene frequency distributions for these repertoires with respect to the background distribution. A comparison of VJ combination and VJ-independent repertoires of peripheral blood mononuclear cells (PBMC) and tumor-infiltrating lymphocytes (TIL) in glioma patients revealed specific signature TCRs that were associated with PBMC of patients exhibiting low TIL divergence and which were depleted in patients with highly divergent TIL repertoires. This divergence, detectable in PBMC, can be used as a noninvasive technique for longitudinal monitoring of glioma (121). JSD has also been used to find similarity in isotype abundance in repertoires of individuals (95). The R package "tcR" includes a function to calculate the JS divergence for TCR and BCR repertoires (122).

Storer–Kim (SK) and Kulinskaya–Morgenthaler–Staudte (KMS)

Storer–Kim and KMS tests have been used recently to find statistically significant differences between two distributions (123, 124). Both tests assume non-parametric distribution. The second assumption might not be appropriate when considering affinity maturation and clonal expansion. SK test does not provide a confidence interval while KMS test does. Like JSD, these tests compare probability distributions and hence there is no limitation to number of sequences. A mouse study used SK and KMS tests to compare the V family usage within GC B cell repertoire of animals vaccinated with complex Ebola virus-like particle and unvaccinated controls (125). Enhanced use of IGHV8 was observed in the vaccinated group. The tests have been implemented in "WRS2" R package (126).

Repertoire Dissimilarity Index (RDI)

Repertoire dissimilarity index compares Ig repertoire based on usage of V, D, or J (127, 128). It is a non-parametric method, which tries to circumvent the problem of varying number of sequences in different samples. The first step involves subsampling the larger sample to the size of the smaller one. From these uniform samples, the feature of interest is counted; the frequency is normalized and transformed into probability distributions. Root mean square deviation (RMSD) is calculated between the two. Random subsampling is done multiple times and mean RMSD is calculated to get the RDI. This reduces sampling bias effects of rarefication to some degree. Since, for each comparison, the sample size is the lower of the two, RDI values between different samples are not comparable. Also, with decrease in sample size, RDI values increase. The RDI value gets closer to the true value as sample size increases. RDI was used to show that genetic bias effects VJ usage by analyzing BCR repertoire of monozygotic twins (127). RDI was validated by recapitulating known differences between T-cell subsets (128). R codes for calculation of RDI are available at https://bitbucket.org/cbolen1/rdicore (128).

Diversity

Diversity has frequently been used to describe lymphocyte antigen receptor repertoires. These indices come from ecology, where they are used to compare the diversities of ecosystems. With respect to the immune repertoire, diversity can be calculated in terms of use of V, D, and J gene segments as well as the use of individual CDR3s. Depending upon the kind of comparison diversity can be categorized into three types, namely, alpha, beta, and gamma. Alpha diversity is the diversity of an individual's repertoire, i.e., the total number of individual species (V_H or CDR3) present in the repertoire. This is also the species richness. Beta diversity gives a difference in repertoire of two individuals. It would be given by the sum of unique species in both the repertoires. Gamma diversity is a combine diversity of all the ecosystems or repertoires. Alpha, beta, and gamma diversities were compared between

patients with gastritis with (GHP) and without *Helicobacter pylori* (GNHP) background, gastric mucosa-associated lymphoid tissue lymphoma (MALT-L) (caused by GHP), and diffuse large B cell lymphoma (DLBCL) (may or may not be transformed MALT-L) (129). Contrary to the expectation, similar diversification was found in both GHP and GNHP, and MALT-L transformed DLBCL, and independent DLBCL. Also, MALT-L transformed DLBCL and MALT-L patients did not share any feature in their repertoires.

Species Richness

Species richness (alpha diversity) is the total number of unique species in a community. It is just a count and does not take into account the species abundance. It is the simplest way of describing diversity but is very sensitive to sampling depth. Greater sampling depth results in capture of more and more rare species resulting in higher species richness. Rarification can have a significant impact on this measure, as less represented species are usually lost during random subsampling. To account for the unseen species problem for under-sampled population, a number of measures have been devised which predict the actual species richness based on the sampled data, including Chao1 (130), abundance-based coverage estimators (ACE) (131), and DivE (132).

Chao1 and Abundance-based Coverage Estimators (ACE)

Chao1 and ACE have commonly been used in assessment of microbial species richness. These estimators add a correction factor to the number of observed species to account for the hidden/unsampled once (133). Chao1, for example, extrapolates the richness based on the number of rare species (count = 1 or 2) found in the samples.

$$Chao1 = S_{obs} + \frac{n_1^2}{2n_2},$$

where S_{obs} is the observed number of species, n_1 is the number of singletons (species with count = 1), and n_2 is the number of doubletons (species with count = 2).

Abundance-based coverage estimator, on the other hand, takes into account the number of species with count less than or equal to 10. It is calculated as:

$$ACE = S_{abund} + \left(\frac{S_{rare}}{C_{ACE}}\right) + \left(\frac{F_1}{C_{ACE}}\right)\gamma^2_{ACE,}$$

where S_{abund} is the number of species with count greater than 10; S_{rare} is the number of species with count less than or equal to 10; $C_{ACE} = 1 - F_1/N_{rare}$; F_1 is the number of species with count = 1

$$N_{rare} = \sum_{i=1}^{10} iF_i;\ F_i \text{ is the number of species with count}=i,$$

$$\gamma^2_{ACE} = max\left[\frac{S_{rare}\sum_{i=1}^{10} i(i-1)F_i}{C_{ACE}\left(N_{rare}\right)\left(N_{rare}-1\right)} - 1, 0\right];$$

Coefficient of variations of F_i's.

Even with the correction factors incorporated to calculate the true species richness, these estimators are still sensitive to sampling depth. A small change in the library preparation steps leading to increased sample quality or quantity may impact species diversity measurements. These factors are still unable to predict the real number of unseen species.

Diversity Estimator (DivE)

DivE (Diversity Estimator) is a diversity measure used originally in the calculation of TCR repertoire diversity (132). The initial step involves construction of rarefaction curves for multiple nested subsamples. A rarefaction curve is a plot of the number of species as a function of the number of sequences or sample size. A mathematical model, defining each of the rarefaction curves, is built and tested on all the nested samples. Each model is scored based on degree of fit using four criteria: Discrepancy (between the data points and the model), accuracy (of predicted versus actual species richness), similarity (between area between the curve fitted to the subsample and the complete data), and plausibility (the predicted number of species should increase or plateau off or the rate of increase of species should decrease or remain constant—any other scenario is not plausible). The top five scored models are extrapolated and combined to calculate a DivE. This estimator is unaffected by sample size and its accuracy is improved from the use of multiple models to predict diversity. The drawback is that the calculation process is lengthy and there is a requirement to fit multiple models. DivE has been used to calculate the species richness of T cell repertoires. With B cell repertoires being even more diverse, the computations are expected to be more complex. This species richness estimator was used to calculate the number of cells infected with human T-lymphotropic virus type 1 in patients, species richness in a TCR repertoire and fecal microbiota of infants (132).

These estimators have been adopted in analysis of diversity of BCR and TCR repertoires. Studies on the effect of aging on the B cell immune repertoire diversity on administration of influenza vaccine showed that the repertoires become more specialized and less plastic with age (134). Both naïve and antigen-experienced repertoires show reduced diversity with age. The Chao1 estimator was used to describe BCR repertoire differences within and between individuals (84). The R packages for estimation of DivE (132), Chao1, and ACE (135) are available.

Although species richness may be the most direct measure of diversity, evenness or the homogeneity/uniformity of species in the community also provides important information. Species evenness would describe the degree of clonal expansion in an immune repertoire. Two common indices calculated considering both richness and evenness, namely, the Shannon Index and the Simpson Indexes have different prospective for each (136).

Shannon Index (*H*)

Shannon index (*H*) calculations operate under the assumptions that individuals are randomly sampled from an infinitely large community, and that all species are represented in the sample.

The Shannon index increases as both richness and evenness of the community increase. The Shannon index is given by:

$$H = -\sum_{i=1}^{s} p_i \ln p_i,$$

where $p_i = n_i/N$ the proportion of individuals of the ith species; n_i is the number of individuals of the ith species; and N is the total number of individuals and s is the total number of species. Since this index is directly proportional to the species richness, it is sensitive to sampling depth.

Simpson Index of Diversity

Simpson Index of diversity is calculated as 1 − Dominance Index (D). D gives more weight to dominant species. It gives the probability that two individuals drawn from a population will belong to the same species. Thus, presence of rare species would not affect D and D increases with increase in dominance leading to decease in diversity. Simpson index of diversity ($1 - D$) gives the diversity value, which increases with decrease in dominance.

$$1 - D = 1 - \frac{\sum_{i=1}^{s} n_i (n_i - 1)}{N(N-1)},$$

where n_i is the number of individuals of the ith species and N is the total number of individuals, and s is the total number of species.

Shannon diversity has been used widely in antigen receptor diversity analysis. Some examples of this analysis in human studies include the comparison of TCR repertoires in colorectal tumors and adjacent healthy mucosa (137) and B cell repertoire of patients before and after hematopoietic stem cell transplantation (138). R packages are available for calculation of diversity indices like vegan (139) and BiodiversityR (140), with one developed specifically to characterize and analyze immune repertoires (122). Recon is another program developed to calculate the diversity measures (68).

Diversity 50 (D50)

Diversity 50 or D50 is the percentage of dominant unique species, which make up 50% of the total community. In terms of Ig repertoire, it is the percentage of distinct V_H segments or CDR3 constituting half of the total V_H or CDR3 in a population (141). A larger D50 value shows larger diversity. D50, like the Simpson index, is based on the number of dominant species and is not affected by the addition of rare species. The D50 has been used to compare the degree of clonal expansion/clonal dominance during infection. Both T and B cell repertoire diversity have been assessed *via* D50 analysis in human studies of viral infection (74), as well as in the characterization of TCR diversity in patients with Wiskott–Aldrich syndrome (142).

UniFrac Distance Matrix

In the context of microbial communities, this index includes environmental differences by taking into consideration phylogenetic information (143, 144). The branch lengths are deemed to differ based on genetic changes occurring due to environmental selection pressure. Thus, the branch lengths between two species in both communities are taken into account while considering the distance between the communities. Analogically, different selection pressure within repertoires of two organisms can be taken into account by including the phylogenetic information starting from the germline sequence (134). UniFrac distance is also sensitive to sequencing depth. Smaller number of sequences in a sample would be underrepresenting the rare species and this would artificially influence the distance between similar communities. UniFrac distance was used to calculate the difference between the Ig repertoires before and after immunization with influenza vaccine in old and young individuals. With age, Ig repertoires appear to become more specialized and less plastic, resulting in lower uniFrac distances, compared to younger individuals (134). R packages for calculation of uniFrac distance are available (145, 146).

Principal Component Analysis (PCA)

Principal component analysis is a way of simplifying the analysis of large datasets by reducing the dataset dimensionality. It does so by creating a new set of variables or principal components (PCs), which describe more complex variability in the data set. The first PC (PC1) explains the maximum variance of the dataset, followed by PC2, and so on. PCA can also help identify patterns in the data, which would otherwise not be prominent. PCA can be used to compare the Ig repertoire based on multiple variables. Using multiple variables like diversity, mutation rates, and others, to define Ig repertoire under different conditions, PCA has been used to find association patterns between these groups. A limitation of PCA is that it considers only linearly correlated data. Also, it discards smaller variance as noise, which may be important under certain conditions. Depending upon the variables being used to analyze the samples, PCA may or may not be dependent on sequencing depth. For example, having diversity as one of the variables would make PCA sample size dependent. PCA has been used on V(D)J usage among productive antibodies to explore the relationship between pre-B, FO, and MZ cells. A very clear clustering and gradient separation of pre-B, follicular, and marginal zone cell subsets was seen which was also observed with V usage analysis but not that of D and J (147). In a study comparing the effect of various influenza vaccines on B cell repertoire, PCA was applied to rarefaction analysis, diversity, V usage frequencies, and mutation rates for unimmunized and immunized groups (148). The basic stats package of R has function for PCA.

When it comes to the analysis of Ig repertoires, a standard protocol has yet to be set. The specific scientific question and the difference in the sequencing depth is one of the major concerns when selecting a statistical approach. Rarefication, a way to overcome differences in sequencing depth, works best when the number of sequences is not very different for each sample. This criterion is not always met. Chi-square test does not work well with sequencing depth of over a few thousand. The JSD, SK, and KMs approaches work are reasonable measures for large sequencing data sets. RDI addresses the problem of variable sequencing read depths by resampling multiple times and taking the mean. However, the RDI values for two different pairs of data are not

comparable. JSD on the other hand always gives a bounded value between 0 and 1 and can be relatively scaled between different comparisons. Diversity, though being the most common method used to assess and compare the Ig repertoire, is very susceptible to sequencing depth. Because each estimator used alone incompletely describes the diversity of a B cell repertoire, multiple parallel approaches are warranted.

CONCLUDING REMARKS

High-throughput sequencing provided immunology with a tool to enhance our understanding of lymphocyte antigen receptor repertoires. With increased application in human diagnostics—sample preparation, sequencing, and analysis techniques will continue to evolve to assist workers in describing lymphocyte antigen receptor repertoires. As large data sets become less expensive and more efficiently produced, necessities for more uniform and improved analysis methods are expected to drive further innovation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by the National Institutes of Health grants AI121394, and AI1113217 (to DW), a Ruth L. Kirschstein National Research Service Award Institutional Research Training Grant (AI007306-31) (to NC). DW holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a New Investigator Award from Food Allergy Research & Education (FARE).

REFERENCES

1. Schatz DG, Ji Y. Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* (2011) 11:251–63. doi:10.1038/nri2941
2. Alt FW, Oltz EM, Young F, Gorman J, Taccioli G, Chen J. VDJ recombination. *Immunol Today* (1992) 13:306–14. doi:10.1016/0167-5699(92)90043-7
3. Barreto V, Cumano A. Frequency and characterization of phenotypic Ig heavy chain allelically included IgM-expressing B cells in mice. *J Immunol* (2000) 164:893–9. doi:10.4049/jimmunol.164.2.893
4. Giachino C, Padovan E, Lanzavecchia A. kappa+lambda+ dual receptor B cells are present in the human peripheral repertoire. *J Exp Med* (1995) 181:1245–50. doi:10.1084/jem.181.3.1245
5. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* (2000) 102:553–63. doi:10.1016/S0092-8674(00)00078-7
6. Chaudhuri J, Tian M, Khuong C, Chua K, Pinaud E, Alt FW. Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature* (2003) 422:726–30. doi:10.1038/nature01574
7. De Silva NS, Klein U. Dynamics of B cells in germinal centres. *Nat Rev Immunol* (2015) 15:137–48. doi:10.1038/nri3804
8. Calis JJA, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol* (2014) 35:581–90. doi:10.1016/j.it.2014.09.004
9. Trepel F. Number and distribution of lymphocytes in man. A critical analysis. *Klin Wochenschr* (1974) 52:511–5. doi:10.1007/BF01468720
10. Bashford-Rogers RJM, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, et al. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol* (2014) 15:29. doi:10.1186/s12865-014-0029-0
11. Kitaura K, Yamashita H, Ayabe H, Shini T, Matsutani T, Suzuki R. Different somatic hypermutation levels among antibody subclasses disclosed by a new next-generation sequencing-based antibody repertoire analysis. *Front Immunol* (2017) 8:389. doi:10.3389/fimmu.2017.00389
12. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung M-W, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* (2013) 4:2680. doi:10.1038/ncomms3680
13. Yeku O, Frohman MA. Rapid amplification of cDNA ends (RACE). In: Nielsen H, editor. *RNA: Methods and Protocols*. Totowa, NJ: Humana Press (2011). p. 107–22.
14. Lin SG, Ba Z, Du Z, Zhang Y, Hu J, Alt FW. Highly sensitive and unbiased approach for elucidating antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113:7846–51. doi:10.1073/pnas.1608649113
15. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* (2012) 30:434–9. doi:10.1038/nbt.2198
16. Zhu J, O'Dell S, Ofek G, Pancera M, Wu X, Zhang B, et al. Somatic Populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Front Microbiol* (2012) 3:315. doi:10.3389/fmicb.2012.00315
17. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods* (2014) 11:653–5. doi:10.1038/nmeth.2960
18. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* (2012) 109:14508–13. doi:10.1073/pnas.1208715109
19. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) 11:1599–616. doi:10.1038/nprot.2016.093
20. Cole C, Volden R, Dharmadhikari S, Scelfo-Dalbey C, Vollmers C. Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier-guided amplicon assembly. *J Immunol* (2016) 196:2902–7. doi:10.4049/jimmunol.1502563
21. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh H-J, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) 2:e1501371. doi:10.1126/sciadv.1501371
22. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova OV, Sharonov GV, Staroverov DB, et al. Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J Immunol* (2015) 194:6155–63. doi:10.4049/jimmunol.1500215
23. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* (2017) 27:491–9. doi:10.1101/gr.209601.116
24. Briney B, Le K, Zhu J, Burton DR. Clonify: unseeded antibody lineage assignment from next-generation sequencing data. *Sci Rep* (2016) 6:23901. doi:10.1038/srep23901
25. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, Pevzner PA. Reconstructing antibody repertoires from error-prone immunosequencing reads. *J Immunol* (2017) 199:3369–80. doi:10.4049/jimmunol.1700485
26. Deakin CT, Deakin JJ, Ginn SL, Young P, Humphreys D, Suter CM, et al. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Res* (2014) 42:e129. doi:10.1093/nar/gku607
27. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp Cell Res* (2014) 322:12–20. doi:10.1016/j.yexcr.2014.01.008
28. Kelley DE, Perry RP. Transcriptional and posttranscriptional control of immunoglobulin mRNA production during B lymphocyte development. *Nucleic Acids Res* (1986) 14:5431–47. doi:10.1093/nar/14.13.5431

29. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-throughput pairing of T cell receptor α and β sequences. *Sci Transl Med* (2015) 7:301ra131. doi:10.1126/scitranslmed.aac5624
30. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* (2013) 31:166–9. doi:10.1038/nbt.2492
31. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2015) 21:86–91. doi:10.1038/nm.3743
32. Busse CE, Czogiel I, Braun P, Arndt PF, Wardemann H. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur J Immunol* (2014) 44:597–603. doi:10.1002/eji.201343917
33. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* (2015) 161:1202–14. doi:10.1016/j.cell.2015.05.002
34. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* (2010) 38:1767–71. doi:10.1093/nar/gkp1137
35. Lefranc MP. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* (2001) 29:207–9. doi:10.1093/nar/29.1.207
36. Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res* (2004) 32:W435–40. doi:10.1093/nar/gkh412
37. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* (2013) 41:W34–40. doi:10.1093/nar/gkt382
38. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol* (2012) 189:3221–30. doi:10.4049/jimmunol.1201303
39. Corcoran MM, Phad GE, Vázquez Bernat N, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
40. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* (2015) 112:E862–70. doi:10.1073/pnas.1417683112
41. Zhang W, Wang I-M, Wang C, Lin L, Chai X, Wu J, et al. IMPre: an accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol* (2016) 7:457. doi:10.3389/fimmu.2016.00457
42. Wendel BS, He C, Crompton PD, Pierce SK, Jiang N. A streamlined approach to antibody novel germline allele prediction and validation. *Front Immunol* (2017) 8:1072. doi:10.3389/fimmu.2017.01072
43. Brochet X, Lefranc M-P, Giudicelli V. 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* (2008) 36:W503–8. doi:10.1093/nar/gkn316
44. Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* (2011). doi:10.1101/pdb.prot5633
45. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V. IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In: Christiansen FT, Tait BD, editors. *Immunogenetics: Methods and Applications in Clinical Practice*. Totowa, NJ: Humana Press (2012). p. 569–604.
46. Li S, Lefranc M-P, Miles JJ, Alamyar E, Giudicelli V, Duroux P, et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* (2013) 4:2333. doi:10.1038/ncomms3333
47. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc M-P. IMGT/HighV-QUEST statistical significance of IMGT clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and T cell receptors. *PLoS One* (2015) 10:e0142353. doi:10.1371/journal.pone.0142353
48. Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc M-P. IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. *Front Immunol* (2016) 7:339. doi:10.3389/fimmu.2016.00339
49. Souto-Carneiro MM, Longo NS, Russ DE, Sun H, Lipsky PE. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J Immunol* (2004) 172:6790–802. doi:10.4049/jimmunol.172.11.6790
50. Ohm-Laursen L, Nielsen M, Larsen SR, Barington T. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology* (2006) 119:265–77. doi:10.1111/j.1365-2567.2006.02431.x
51. Gaëta BA, Malming HR, Jackson KJL, Bain ME, Wilson P, Collins AM. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* (2007) 23:1580–7. doi:10.1093/bioinformatics/btm147
52. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* (2009) 106:20216–21. doi:10.1073/pnas.0909775106
53. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) 40:e134. doi:10.1093/nar/gks457
54. Rogosch T, Kerzel S, Hoi KH, Zhang Z, Maier RF, Ippolito GC, et al. Immunoglobulin analysis tool: a novel tool for the analysis of human and mouse heavy and light chain transcripts. *Front Immunol* (2012) 3:176. doi:10.3389/fimmu.2012.00176
55. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30:1930–2. doi:10.1093/bioinformatics/btu138
56. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A, et al. Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics* (2014) 15:409. doi:10.1186/1471-2164-15-409
57. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F. Vidjil: a web platform for analysis of high-throughput repertoire sequencing. *PLoS One* (2016) 11:e0166126. doi:10.1371/journal.pone.0166126
58. D'Angelo S, Glanville J, Ferrara F, Naranjo L, Gleasner CD, Shen X, et al. The antibody mining toolbox: an open source tool for the rapid analysis of antibody repertoires. *MAbs* (2014) 6:160–72. doi:10.4161/mabs.27105
59. Safonova Y, Bonissone S, Kurpilyansky E, Starostina E, Lapidus A, Stinson J, et al. IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics* (2015) 31:i53–61. doi:10.1093/bioinformatics/btv238
60. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* (2015) 12:380–1. doi:10.1038/nmeth.3364
61. Zhang W, Du Y, Su Z, Wang C, Zeng X, Zhang R, et al. IMonitor: A robust pipeline for TCR and BCR repertoire analysis. *Genetics* (2015) 201:459–72. doi:10.1534/genetics.115.176735
62. Frost SDW, Murrell B, Hossain ASMM, Silverman GJ, Pond SLK. Assigning and visualizing germline genes in antibody repertoires. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140240. doi:10.1098/rstb.2014.0240
63. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* (2015) 31:3356–8. doi:10.1093/bioinformatics/btv359
64. Yu Y, Ceredig R, Seoighe C. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Res* (2016) 44:e31. doi:10.1093/nar/gkv1016
65. Imkeller K, Arndt PF, Wardemann H, Busse CE. sciReptor: analysis of single-cell level immunoglobulin repertoires. *BMC Bioinformatics* (2016) 17:67. doi:10.1186/s12859-016-0920-1
66. Elhanati Y, Marcou Q, Mora T, Walczak AM. repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from

sequence data. *Bioinformatics* (2016) 32:1943–51. doi:10.1093/bioinformatics/btw112
67. Bischof J, Ibrahim SM. bcRep: R package for comprehensive analysis of B cell receptor repertoire data. *PLoS One* (2016) 11:e0161569. doi:10.1371/journal.pone.0161569
68. Kaplinsky J, Arnaout R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* (2016) 7:11881. doi:10.1038/ncomms11881
69. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A, et al. ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. *Bioinformatics* (2017) 33:435–7. doi:10.1093/bioinformatics/btw634
70. IJspeert H, van Schouwenburg PA, van Zessen D, Pico-Knijnenburg I, Stubbs AP, van der Burg M. Antigen receptor galaxy: a user-friendly, web-based tool for analysis and visualization of T and B cell receptor repertoire data. *J Immunol* (2017) 198:4156–65. doi:10.4049/jimmunol.1601921
71. Marcou Q, Mora T, Walczak AM. IGoR: a tool for high-throughput immune repertoire analysis. *Q-Bio* (2017). Available from: http://arxiv.org/abs/1705.08246
72. Fähnrich A, Krebbel M, Decker N, Leucker M, Lange FD, Kalies K, et al. ClonoCalc and ClonoPlot: immune repertoire analysis from raw files to publication figures with graphical user interface. *BMC Bioinformatics* (2017) 18:164. doi:10.1186/s12859-017-1575-2
73. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* (2017) 547:94–8. doi:10.1038/nature22976
74. Hou D, Ying T, Wang L, Chen C, Lu S, Wang Q, et al. Immune repertoire diversity correlated with mortality in avian influenza A (H7N9) virus infected patients. *Sci Rep* (2016) 6:33843. doi:10.1038/srep33843
75. Martin V, Bryan Wu Y-C, Kipling D, Dunn-Walters D. Ageing of the B-cell repertoire. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140237. doi:10.1098/rstb.2014.0237
76. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) 184:6986–92. doi:10.4049/jimmunol.1000445
77. Miqueu P, Guillet M, Degauque N, Doré J-C, Soulillou J-P, Brouard S. Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol Immunol* (2007) 44:1057–64. doi:10.1016/j.molimm.2006.06.026
78. Wu Y-C, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* (2010) 116:1070–8. doi:10.1182/blood-2010-03-275859
79. DeKosky BJ, Lungu OI, Park D, Johnson EL, Charab W, Chrysostomou C, et al. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113:E2636–45. doi:10.1073/pnas.1525510113
80. Galson JD, Trück J, Fowler A, Clutterbuck EA, Münz M, Cerundolo V, et al. Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine* (2015) 2:2070–9. doi:10.1016/j.ebiom.2015.11.034
81. Wu YB, Kipling D, Dunn-Walters DK. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front Immunol* (2012) 3:193. doi:10.3389/fimmu.2012.00193
82. Liu S, Hou XL, Sui WG, Lu QJ, Hu YL, Dai Y. Direct measurement of B-cell receptor repertoire's composition and variation in systemic lupus erythematosus. *Genes Immun* (2017) 18:22–7. doi:10.1038/gene.2016.45
83. Trück J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J Immunol* (2015) 194:252–61. doi:10.4049/jimmunol.1401405
84. Galson JD, Trück J, Fowler A, Münz M, Cerundolo V, Pollard AJ, et al. In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Front Immunol* (2015) 6:531. doi:10.3389/fimmu.2015.00531
85. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos Trans R Soc B Biol Sci* (2015) 370:20140239. doi:10.1098/rstb.2014.0239
86. Hoehn KB, Fowler A, Lunter G, Pybus OG. The diversity and molecular evolution of B-cell receptors during infection. *Mol Biol Evol* (2016) 33:1147–57. doi:10.1093/molbev/msw015
87. Diaz M, Flajnik MF. Evolution of somatic hypermutation and gene conversion in adaptive immunity. *Immunol Rev* (1998) 162:13–24. doi:10.1111/j.1600-065X.1998.tb01425.x
88. Yeap L-S, Hwang JK, Du Z, Meyers RM, Meng F-L, Jakubauskaitė A, et al. Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* (2015) 163:1124–37. doi:10.1016/j.cell.2015.10.042
89. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) 20:683–94. doi:10.1093/intimm/dxn026
90. Wang C, Liu Y, Cavanagh MM, Le Saux S, Qi Q, Roskin KM, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc Natl Acad Sci U S A* (2015) 112:500–5. doi:10.1073/pnas.1415875112
91. Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140242. doi:10.1098/rstb.2014.0242
92. Briney BS, Willis JR, Crowe JE. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun* (2012) 13:523–9. doi:10.1038/gene.2012.28
93. Wei L, Chahwan R, Wang S, Wang X, Pham PT, Goodman MF, et al. Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc Natl Acad Sci U S A* (2015) 112:E728–37. doi:10.1073/pnas.1500788112
94. Hwang JK, Wang C, Du Z, Meyers RM, Kepler TB, Neuberg D, et al. Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc Natl Acad Sci U S A* (2017) 114:8614–9. doi:10.1073/pnas.1709203114
95. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, et al. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *Elife* (2016) 5:1–20. doi:10.7554/eLife.16578
96. Looney TJ, Lee J-Y, Roskin KM, Hoh RA, King J, Glanville J, et al. Human B-cell isotype switching origins of IgE. *J Allergy Clin Immunol* (2016) 137:579.e–86.e. doi:10.1016/j.jaci.2015.07.014
97. Wesemann DR, Magee JM, Boboila C, Calado DP, Gallagher MP, Portuguese AJ, et al. Immature B cells preferentially switch to IgE with increased direct Sμ to Sε recombination. *J Exp Med* (2011) 208:2733–46. doi:10.1084/jem.20111155
98. Wu X, Zhang Z, Schramm CA, Joyce MG, Do Kwon Y, Zhou T, et al. Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* (2015) 161:470–85. doi:10.1016/j.cell.2015.03.004
99. Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, Lu G, et al. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc Natl Acad Sci U S A* (2013) 110:6470–5. doi:10.1073/pnas.1219320110
100. Sok D, Laserson U, Laserson J, Liu Y, Vigneault F, Julien J-P, et al. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog* (2013) 9:e1003754. doi:10.1371/journal.ppat.1003754
101. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* (2013) 496:469–76. doi:10.1038/nature12053
102. Bashford-Rogers RJM, Nicolaou KA, Bartram J, Goulden NJ, Loizou L, Koumas L, et al. Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* (2016) 30:2312–21. doi:10.1038/leu.2016.142
103. Lee SE, Kang SY, Yoo HY, Kim SJ, Kim WS, Ko YH. Clonal relationships in recurrent B-cell lymphomas. *Oncotarget* (2016) 7:12359–71. doi:10.18632/oncotarget.7132
104. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* (2013) 23:1874–84. doi:10.1101/gr.154815.113
105. Chang Y-H, Kuan H-C, Hsieh TC, Ma KH, Yang C-H, Hsu W-B, et al. Network signatures of IgG immune repertoires In Hepatitis B associated chronic

infection and vaccination responses. *Sci Rep* (2016) 6:26556. doi:10.1038/srep26556
106. Miho E, Greiff V, Roskar R, Reddy ST. The fundamental principles of antibody repertoire architecture revealed by large-scale network analysis. *bioRxiv* (2017). doi:10.1101/124578
107. Csárdi G, Nepusz T. The igraph software package for complex network research. (2006). Available from: http://www.necsi.edu/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.p
108. Wang B, Lee C-H, Johnson EL, Kluwe CA, Cunningham JC, Tanno H, et al. Discovery of high affinity anti-ricin antibodies by B cell receptor sequencing and by yeast display of combinatorial VH:VL libraries from immunized animals. *MAbs* (2016) 8:1035–44. doi:10.1080/19420862.2016.1190059
109. Dai K, He L, Khan SN, O'Dell S, McKee K, Tran K, et al. Rhesus macaque B-cell responses to an HIV-1 trimer vaccine revealed by unbiased longitudinal repertoire analysis. *MBio* (2015) 6:e1375–1315. doi:10.1128/mBio.01375-15
110. Roy B, Neumann RS, Snir O, Iversen R, Sandve GK, Lundin KEA, et al. High-Throughput single-cell analysis of B cell receptor usage among autoantigen-specific plasma cells in celiac disease. *J Immunol* (2017) 199:782–91. doi:10.4049/jimmunol.1700169
111. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med* (2014) 6:248ra107. doi:10.1126/scitranslmed.3008879
112. Wendel BS, He C, Qu M, Wu D, Hernandez SM, Ma K-Y, et al. Accurate immune repertoire sequencing reveals malaria infection driven antibody lineage diversification in young children. *Nat Commun* (2017) 8:531. doi:10.1038/s41467-017-00645-x
113. McHugh ML. The chi-square test of independence. *Biochem Med (Zagreb)* (2013) 23:143–9. doi:10.11613/BM.2013.018
114. Bergh D. Sample size and chi-squared test of fit—a comparison between a random sample approach and a chi-square value adjustment method using Swedish adolescent data. *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg (2014). p. 197–211.
115. Wesemann DR, Portuguese AJ, Meyers RM, Gallagher MP, Cluff-Jones K, Magee JM, et al. Microbial colonization influences early B-lineage development in the gut lamina propria. *Nature* (2013) 501:112–5. doi:10.1038/nature12496
116. Racanelli V, Brunetti C, De Re V, Caggiari L, De Zorzi M, Leone P, et al. Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS One* (2011) 6:e25606. doi:10.1371/journal.pone.0025606
117. Li L, Wang X-H, Banerjee S, Volsky B, Williams C, Virland D, et al. Different pattern of immunoglobulin gene usage by HIV-1 compared to non-HIV-1 antibodies derived from the same infected subject. *PLoS One* (2012) 7:e39534. doi:10.1371/journal.pone.0039534
118. Volpe JM, Kepler TB. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res* (2008) 4:3. doi:10.1186/1745-7580-4-3
119. Lin J. Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theory* (1991) 37:145–51. doi:10.1109/18.61115
120. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* (2017) 547:89–93. doi:10.1038/nature22383
121. Sims JS, Grinshpun B, Feng Y, Ung TH, Neira JA, Samanamud JL, et al. Diversity and divergence of the glioma-infiltrating T-cell receptor repertoire. *Proc Natl Acad Sci U S A* (2016) 113:E3529–37. doi:10.1073/pnas.1601012113
122. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics* (2015) 16:175. doi:10.1186/s12859-015-0613-1
123. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. *J Am Stat Assoc* (1990) 85:146. doi:10.2307/2289537
124. Kulinskaya E, Morgenthaler S, Staudte RG. Variance stabilizing the difference of two binomial proportions. *Am Stat* (2010) 64:350–6. doi:10.1198/tast.2010.09080
125. Khavrutskii IV, Chaudhury S, Stronsky SM, Lee DW, Benko JG, Wallqvist A, et al. Quantitative analysis of repertoire-scale immunoglobulin properties in vaccine-induced B-cell responses. *Front Immunol* (2017) 8:910. doi:10.3389/fimmu.2017.00910
126. Mair P, Wilcox R. *Robust Statistical Methods in R Using the WRS2 Package*. (2014). Available from: https://cran.r-project.org/web/packages/WRS2/vignettes/WRS2.pdf
127. Rubelt F, Bolen CR, McGuire HM, Vander Heiden JA, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat Commun* (2016) 7:11112. doi:10.1038/ncomms11112
128. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. The repertoire dissimilarity index as a method to compare lymphocyte receptor repertoires. *BMC Bioinformatics* (2017) 18:155. doi:10.1186/s12859-017-1556-5
129. Michaeli M, Tabibian-Keissar H, Schiby G, Shahaf G, Pickman Y, Hazanov L, et al. Immunoglobulin gene repertoire diversification and selection in the stomach – from gastritis to gastric lymphomas. *Front Immunol* (2014) 5:264. doi:10.3389/fimmu.2014.00264
130. Chao A. Nonparametric estimation of the number of classes in a population. *Environ Ecol Stat* (1984) 11:265–70.
131. Chao A, Lee S-M. Estimating the number of classes via sample coverage. *J Am Stat Assoc* (1992) 87:210. doi:10.2307/2290471
132. Laydon DJ, Melamed A, Sim A, Gillet NA, Sim K, Darko S, et al. Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput Biol* (2014) 10:e1003646. doi:10.1371/journal.pcbi.1003646
133. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* (2001) 67:4399–406. doi:10.1128/AEM.67.10.4399
134. de Bourcy CFA, Angel CJL, Vollmers C, Dekker CL, Davis MM, Quake SR. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc Natl Acad Sci U S A* (2017) 114:1105–10. doi:10.1073/pnas.1617959114
135. Wang J-P. SPECIES: an R package for species richness estimation. *J Stat Softw* (2011) 40:1–15. doi:10.18637/jss.v040.i09
136. Hill MO. Diversity and evenness: a unifying notation and its consequences. *Ecology* (1973) 54:427–32. doi:10.2307/1934352
137. Sherwood AM, Emerson RO, Scherer D, Habermann N, Buck K, Staffa J, et al. Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. *Cancer Immunol Immunother* (2013) 62:1453–61. doi:10.1007/s00262-013-1446-2
138. Sethi MK, Thol F, Stadler M, Heuser M, Ganser A, Koenecke C, et al. VH1 family immunoglobulin repertoire sequencing after allogeneic hematopoietic stem cell transplantation. *PLoS One* (2017) 12:e0168096. doi:10.1371/journal.pone.0168096
139. O'Connor RJ. Multivariate analysis of ecological communities. *Trends Ecol Evol* (1988) 3:121. doi:10.1016/0169-5347(88)90124-3
140. Kindt R. *Package for Community Ecology and Suitability Analysis*. (2017).
141. Hou X-L, Wang L, Ding Y-L, Xie Q, Diao H-Y. Current status and recent advances of next generation sequencing techniques in immunological repertoire. *Genes Immun* (2016) 17:153–64. doi:10.1038/gene.2016.9
142. Wu J, Liu D, Tu W, Song W, Zhao X. T-cell receptor diversity is selectively skewed in T-cell populations of patients with Wiskott-Aldrich syndrome. *J Allergy Clin Immunol* (2015) 135:209–16. doi:10.1016/j.jaci.2014.06.025
143. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* (2005) 71:8228–35. doi:10.1128/AEM.71.12.8228-8235.2005
144. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *ISME J* (2011) 5:169–72. doi:10.1038/ismej.2010.133
145. McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* (2013) 8:e61217. doi:10.1371/journal.pone.0061217
146. Chen J. Generalized UniFrac distances version. *R Doc* (2012).
147. Kaplinsky J, Li A, Sun A, Coffre M, Koralov SB, Arnaout R. Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc Natl Acad Sci U S A* (2014) 111:E2622–9. doi:10.1073/pnas.1403278111

148. Cortina-Ceballos B, Godoy-Lozano EE, Téllez-Sosa J, Ovilla-Muñoz M, Sámano-Sánchez H, Aguilar-Salgado A, et al. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. *Genome Med* (2015) 7:124. doi:10.1186/s13073-015-0239-y

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.

The reviewer TT declared a shared affiliation, though no other collaboration, with the authors to the handling editor.

Copyright © 2018 Chaudhary and Wesemann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Immunology

PROTOCOLS
published: 29 June 2018
doi: 10.3389/fimmu.2018.01472

Computational Evaluation of B-Cell Clone Sizes in Bulk Populations

Aaron M. Rosenfeld[1], Wenzhao Meng[2], Dora Y. Chen[2], Bochao Zhang[1], Tomer Granot[3], Donna L. Farber[3], Uri Hershberg[1,4] and Eline T. Luning Prak[2]**

[1] School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, United States, [2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [3] Columbia Center for Translational Immunology, Columbia University, New York, NY, United States, [4] Department of Microbiology and Immunology, Drexel College of Medicine, Drexel University, Philadelphia, PA, United States

B cell clones expand and contract during adaptive immune responses and can persist or grow uncontrollably in lymphoproliferative disorders. One way to monitor and track B cell clones is to perform large-scale sampling of bulk cell populations, amplifying, and sequencing antibody gene rearrangements by next-generation sequencing (NGS). Here, we describe a series of computational approaches for estimating B cell clone size in NGS immune repertoire profiling data of antibody heavy chain gene rearrangements. We define three different measures of B cell clone size—copy numbers, instances, and unique sequences—and show how these measures can be used to rank clones, analyze their diversity, and study their distribution within and between individuals. We provide a detailed, step-by-step procedure for performing these analyses using two different data sets of spleen samples from human organ donors. In the first data set, 19 independently generated biological replicates from a single individual are analyzed for B cell clone size, diversity and sampling sufficiency for clonal overlap analysis. In the second data set, B cell clones are compared in eight different organ donors. We comment upon frequently encountered pitfalls and offer practical advice with alternative approaches. Overall, we provide a series of pragmatic analytical approaches and show how different clone size measures can be used to study the clonal landscape in bulk B cell immune repertoire profiling data.

Keywords: B cell, clone, antibody, immune repertoire, next generation sequencing, immunoglobulin, diversity

OPEN ACCESS

Edited by:
Gregory C. Ippolito, University of Texas at Austin, United States

Reviewed by:
Sai T. Reddy, ETH Zürich, Switzerland
Kay L. Medina, Mayo Clinic, United States

****Correspondence:***
Uri Hershberg
uh25@drexel.edu;
Eline T. Luning Prak
luning@pennmedicine.upenn.edu

Specialty section:
This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology

Received: *07 December 2017*
Accepted: *13 June 2018*
Published: *29 June 2018*

Citation:
Rosenfeld AM, Meng W, Chen DY, Zhang B, Granot T, Farber DL, Hershberg U and Luning Prak ET (2018) Computational Evaluation of B-Cell Clone Sizes in Bulk Populations. Front. Immunol. 9:1472. doi: 10.3389/fimmu.2018.01472

INTRODUCTION

The accurate measurement of clone size is fundamental to many immunological studies. B cells that are clonally related derive from a common progenitor cell. B cell clones can be viewed as the unit of selection in an immune response (1); the successful recruitment of clones results in diversification and expansion of cells with the appropriate antigen specificity and effector function (2, 3). Longitudinal studies of B cell responses, such as those tracking influenza-binding B cell clones (4, 5) require methods for measuring clone sizes and comparing them at different time points. Tracking B cell clones over time is also important for the diagnosis and monitoring of lymphoproliferative disorders such as chronic lymphocytic leukemia (6). Determining if a clone is likely to be present or absent in a population, as is the case for minimal residual disease testing (7), requires knowing or defining the analysis on the expected size of the clone and powering the analysis to detect clones of that size in the population (8, 9). Further complicating the analysis, the human

B-cell repertoire contains a diverse collection of B cell clones of different sizes (9). Hence, clone tracking methods need to take several factors into account, including the level of sampling (the number of B cells being studied), the depth of sequencing (including the number of independently generated sequencing libraries per sample), and the distribution of clone sizes in the population being studied.

Here, we describe a series of computational procedures for estimating B-cell clone sizes in bulk populations using next-generation sequencing (NGS) data on antibody heavy chain gene rearrangements in genomic DNA (gDNA). The analysis of gDNA is the most parsimonious means of studying clone sizes on a large scale as each cell has only one template and many cells can be efficiently queried. Clonal overlap analysis and clone tracking typically require extensive sampling (10). Genomic DNA also provides information on non-productive gene rearrangements, providing a potential second target to identify a clone in B cells with two heavy chain gene rearrangements. Furthermore, DNA is less likely to be degraded than RNA, making it more versatile for suboptimal samples, such as those having low viability or those being derived from fixed tissues or cells. The analysis of the antibody heavy chain is most informative for clone identification and tracking because it has the most diverse CDR3 sequence (by virtue of the D gene segment, two rearrangement junctions and higher levels of non-templated additions and deletions at the junctions). IgH rearrangements amplified from gDNA are also the most often used sample type in the clinical setting, where parsimonious and robust assays are required.

With respect to the data generation, there are already several excellent protocols for immune repertoire profiling by NGS from DNA, RNA, and single cells (11–18). These different methods can be compared against one another on the same sample, along with procedures such as digital droplet PCR to perform experimental estimates on clone size (8). Single cell PCR methods, performed in emulsions or on beads provide a quantitative means of counting individual cells. These methods rely upon cDNA synthesis, either with reconstruction from RNAseq libraries or target capture-based approaches [reviewed in Ref (19)]. In addition to more straightforward quantitation (counting individual cells), single cell approaches can provide paired heavy and light chain IgH/IgL data from the same cell, providing additional fidelity for clonal assignment. One potential drawback of the single cell approach is that the efficiency of IgH/IgL amplification differs in different B cell subsets due to differences in RNA template abundance. The subset can be controlled by sorting or it may be possible to correct for these differences by measuring the recovery of IgH/IgL pairs from different subsets that are identified using other information about the cells (such as RNA transcript profiling within the same experiment). Of note, there have been recent advances in the generation of algorithms that deduce IgH/IgL rearrangements from single cell RNAseq data (20).

With bulk cell samples, one approach to measuring clone size experimentally is to use molecular calibrators (16, 21). With molecular calibrators, one or preferably several cloned standards are spiked into the reaction at known concentrations. For better quantification, multiple dilutions of standards are used, yielding a standard curve against which values of unknown rearrangements can be compared. In the log-linear range of the curve, quantification is most accurate. Standards can correct for differences in amplification efficiency of different VH or Vβ genes (21). One challenging aspect of molecular calibrators is that antibody genes can undergo somatic hypermutation (SHM). In fact, clones of interest often harbor somatic mutations when one is studying an immune response or certain forms of B cell neoplasia (such as follicular lymphoma or multiple myeloma) (22, 23). If mutations occur in the region of primer binding, the use of germline gene standards may not accurately model the PCR efficiency. For RNA-based libraries, a series of RNA spike-in standards has been developed that includes different murine and human VH genes, different lengths, different concentrations, and different levels of SHM (16, 24). The use of these standards following a protocol termed molecular amplification fingerprinting, allows for correction in PCR amplification efficiency and bias (16). While quite useful for understanding the nature of bias and error in the PCR amplification and sequencing steps (24), such calibrators have not yet been validated for broad use, particularly with pauci-cellular or suboptimal samples such as formalin-fixed paraffin-embedded specimens, where the calibrators may out-compete the lower quality sample templates.

Another method for evaluating clone size in bulk populations is limiting dilution analysis. In this method, one prepares serial dilutions of the sample and assays the rearrangements (or antibodies) of multiple replicates at different sample inputs (25, 26). The key to doing this well is to sequence several replicate libraries at each dilution factor. At limiting dilution, the event of interest is counted as present or absent and its frequency in the sample can be modeled using the Poisson distribution (27). As with single cell sequencing, this approach is expensive and requires extensive sampling and sequencing.

In our view, it is quite difficult to establish a "gold standard" for clone size estimation in bulk cell samples. Samples that contain mixed populations of cells with varying levels of SHM present a complex mixture of different templates for amplification. There can be PCR jackpot events, which can result in spurious clonal expansions. While many applications of clone tracking focus on large differences in clone size, with smaller clones or more subtle shifts in clone size, other factors come into play such as differences in sampling or library quality and sequencing depth. The advent of high-throughput sequencing has radically increased the number of cells we study when we analyze immune repertoires. Nonetheless, we still must assume in nearly all cases that our experiments are under-sampling the full diversity of the repertoires we are studying. To address this issue and ask questions about diversity and sampling of immune repertoires, we and others have turned to ecology for tools and methods (28–40).

In this Protocol, we focus on sample-based computational procedures for evaluating clone sizes in bulk B cell antibody sequencing libraries. We describe and illustrate the use of metrics that rely on the analysis of individual sequencing libraries and repeating the analysis with multiple libraries per sample. We define three different metrics of B-cell clone size based upon

sequence copies, instances, and unique sequences. To illustrate these procedures, we use two data sets from human spleen. We chose the spleen because it contains a complex mixture of B cell clones ranging in size (9). The spleen also contains abundant populations of memory B cells, providing a diverse mixture of B-cell clonal types, over a range of different SHM levels (41, 42). The spleen is also large, providing an ample supply of diverse clones for demonstrating clone size metrics that require different degrees of sampling. We describe our procedures using a large number of independently amplified sequencing libraries from the spleen of one organ donor and in a newly generated data set of spleen samples from eight different organ donors (**Figure 1**). Using these deep and survey-level sequencing data sets, we illustrate measures of within- and between-individual clonal size and diversity analysis. We propose a step-by-step approach that we hope will be useful for investigators who study clones in a variety of settings ranging from immune responses to malignancy.

MATERIALS AND EQUIPMENT

Donors

Human tissues used in this research were obtained from deceased organ donors through an approved research protocol and material transfer agreement with LiveOnNY, the organ procurement organization for the New York metropolitan area, as described previously (43). This type of research has been determined by IRBs of both the University of Pennsylvania and Columbia University to be non-human subjects research and, hence, ethics approval was not required, per institutional and national guidelines. A summary of donor information is provided in **Table 1**.

Sample Processing

Spleen samples were maintained in cold saline and brought to the laboratory at the University of Columbia within 4 h of organ procurement. Samples from D207 were processed as described [Experiment 1 in **Figure 1** (9)] All other donor samples were rapidly processed to obtain lymphocyte populations, as described in detail (43, 44) and cryopreserved. Frozen cells were shipped on dry ice to the University of Pennsylvania. On the day of experiment 2 (see **Figure 1**), all of the cryopreserved samples were thawed and processed. Each sample was split into two aliquots. Genomic DNA was extracted from the first aliquot using a Qiagen Gentra Puregene cell kit following the manufacturer's directions (Qiagen, Valencia, CA, USA, Cat. No. 158388). Flow cytometry was performed on the second cell aliquot to obtain the B-cell fraction. The following antibody-fluorophore combinations were used: FITC anti-CD19 (HIB19), PE anti-CD20 (2H7), APC anti-CD3 (HIT3a). Data were acquired on an LSRII flow cytometer (BD Biosciences, San Jose, CA, USA) and analyzed using FlowJo version 7.6.5 software (Treestar Inc., Ashland, OR, USA). The B cell fraction (CD19+CD20+CD3− divided by the total cells) for each donor spleen sample (except D207) is shown in **Table 2**.

Antibody Heavy Chain Gene Rearrangement Amplification

The D207 sequencing libraries were generated as described previously (only the FR1 + JH amplified samples are included in this analysis) (9). For all samples from donors other than D207, sequencing libraries were amplified using a cocktail of VH1, VH2, VH3, VH4, VH5, and VH6 family specific primers in FR1 and

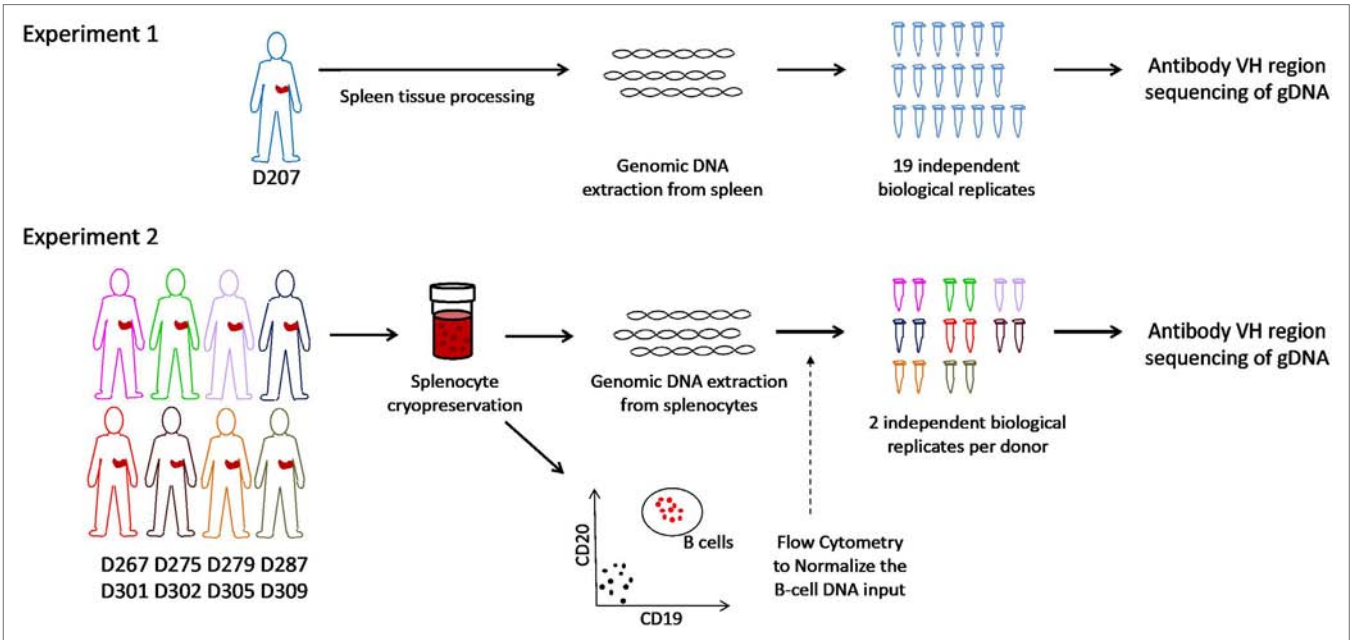


FIGURE 1 | Schematic of experiments. In experiment 1, genomic DNA is extracted from spleen tissue of an organ donor. Immunoglobulin heavy chain gene rearrangements were amplified and sequenced from genomic DNA in 19 biological replicates. In experiment 2, splenocytes from eight different organ donors were cryopreserved at the time of organ recovery. All samples were thawed on the same day and analyzed for the B cell fraction by flow cytometry. Genomic DNA (normalized for the B cell fraction) was separately amplified and sequenced (two biological replicates per subject).

TABLE 1 | Demographic characteristics of the organ donors.

Donor	Age	Sex	Race	Cause of death	WBC final	HCV	CMV	EBV
267	70	F	Black	CVA	11.3	0	1	1
275	31	M	White	Anoxia	17.8	0	0	1
279	73	F	White	CVA	25.4	0	0	1
287	34	M	White	Head trauma	5.6	0	1	1
301	33	F	Hispanic	Anoxia	29.8	0	1	1
302	56	M	Hispanic	Anoxia	16.1	0	1	1
305	28	F	White	Anoxia	9.0	0	0	0
309	45	F	Black	CVA	27.2	0	1	1
207	23	M	Hispanic	Head trauma	15.7	0	1	1

Donor numbers are assigned by the Farber Lab. Age is in years. Cause of death is classified as cerebrovascular accident (CVA), head trauma, or anoxia. WBC, white blood cell count in thousands per microliter. Serologic status (IgG) for hepatitis C virus (HCV), cytomegalovirus (CMV), and Epstein–Barr Virus (EBV). 1 = positive; 0 = negative.

TABLE 2 | B cell percentages in spleen cell samples.

Donor	B cell (% total)	B cell (% ly gate)
267	8.48	33.76
275	12.2	18.27
279	25.4	39.76
287	37	43.65
301	6.32	13.67
302	15.6	28.12
305	35.6	54.17
309	7.11	28.6

The B-cell (CD19+CD20+CD3– lymphocyte) percentage is calculated either out of the total mononuclear cells processed by Ficoll density gradient (total) or from the lymphocyte gate (ly gate). The % total was used to normalize the B-cell content between donors.

TABLE 3 | PCR primers with Illumina adapters for human IgH rearrangement sequencing.

NexteraR2-Hu-VH1-FW1 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACA GGGCCTCAGTGAAGGTCTCCTGCAAG-3′
NexteraR2-Hu-VH2-FW1 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGGTCTGGTCCTACGCTGGTGAAACCC-3′
NexteraR2-Hu-VH3-FW1 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGCTGGGGGGTCCCTGAGACTCTCCTG-3′
NexteraR2-Hu-VH4-FW1 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGCTTCGGAGACCCTGTCCCTCACCTG-3′
NexteraR2-Hu-VH5-FW1 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGA CAGCGGGGAGTCTCTGAAGATCTCCTGT-3′
NexteraR2-Hu-VH6-FW1 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGTCGCAGACCCTCTCACTCACCTGTG-3′
NexteraR1-Hu-JHmix1 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TACGTNCTTACCTGAGGAGACGGTGACC-3′
NexteraR1-Hu-JHmix2 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGC TGCNCTTACCTGAGGAGACGGTGACC-3′
NexteraR1-Hu-JHmix3 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG AGNCTTACCTGAGGAGACGGTGACC-3′

While there is only one consensus JH primer, three primers with different spacers are used to generate sequencing diversity during sequencing from the JH end.

one consensus JH region primer, adapted from the BIOMED2 primer series (45). Primers were synthesized by Integrated DNA Technologies (Coralville, IA, USA) and their sequences are provided in **Table 3**. The input DNA for amplification was normalized to the B cell fraction in the sample. For example, for D305, the B cell fraction (out of total cells) was 35%. To amplify 50 ng-equivalents of B cell gDNA from D305 spleen, we used 142.8 ng of input DNA (50 ng/0.35 = 142.8 ng). For each 25 µL amplification, primers were used at a concentration of 0.6 µM, gDNA normalized to represent 50 ng equivalents of B cell DNA, 0.2 mM dNTPs, and 1× PCR buffer with 1.5 mM $MgCl_2$ using the Qiagen Multiplex PCR kit (Qiagen, Valencia, CA, USA; Cat. No. 206143) in molecular biology grade water (Millipore Sigma, St. Louis, MO, USA; Cat. No. W4502-1L). Amplification conditions for the PCR were primary denaturation at 95°C for 7 min, followed by cycling at 95°C 45 s, 60°C for 45 s, extension at 72°C for 90 s for 35 cycles, and a final extension step at 72°C for 10 min, using a Veriti 96-well thermal cycler (Life Technologies Corporation, Carlsbad, CA, USA) in 96-well plates (Denville, Holliston, MD, USA; Cat. No. C18080-10) sealed with Microseal B adhesive seal (BioRad, Cat. No. MSB1901). Amplicons were visualized on 1.5% agarose gels (Invitrogen/ThermoFisher, Waltham, MA, USA; Cat. No. 16500500) in TAE buffer, prepared fresh from 50× stock solution (Quality Biological, Gaithersburg, MD, USA; Cat. No. 351-008-491).

Library Preparation and Sequencing

Amplicons were purified using the Agencourt AMPure XP beads system (Beckman Coulter, Inc., Indianapolis, IN, USA; Cat. No. A63882) in a 1:1 ratio of beads to sample and eluted in 40 µL of TE (0.1 mM EDTA) buffer. 96-well plates with purified samples were sealed with adhesive aluminum sealing foil (RPI, Mount Prospect, IL, USA; Cat. No. 202502) and saved at −20°C if the second-round PCRs were not performed immediately following purification. Second-round PCRs (to generate the sequencing libraries with individual sample barcodes) were carried out using 4 µL of the first-round PCR product and 2.5 µL each of NexteraXT Index Primers S5XX and N7XX, using the Qiagen Multiplex PCR kit in a reaction volume of 25 µL. Amplification conditions for the library PCR were primary denaturation at 95°C for 10 min, followed by cycling at 95°C 30 s, 60°C 30 s, extension at 72°C 45 s for eight cycles, and a final extension step at 72°C for 10 min. Library amplicons were pooled and then subjected to two rounds of purification using the AMPure XP beads system. In both rounds of purification a 1:1 ratio of beads to sample was used, as before. After the first round of purification, the beads were eluted in TE buffer (1× Solution pH 8.0 with low EDTA, Affymetrix, Santa Clara, CA, USA; Cat. No. J75793-AP). Then, an equal volume of beads and TE eluate were mixed together and repurified. DNA concentrations of purified library preparations were measured using the Qubit 3.0 instrument (Invitrogen/ ThermoFisher) with the Qubit dsDNA HS Assay Kit following the manufacturer's instructions (Invitrogen, Cat. No. Q32851). Pooled libraries with a final concentration of 15 pM and PhiX control (titrated to be 10% of the concentration of the sequencing libraries; PhiX V3 Kit, Illumina Cat. No. FC-110-3001) were loaded onto an Illumina MiSeq instrument in the Human Immunology Core Facility at the University of Pennsylvania. 2 × 300 bp paired end kits were used (MiSeq Reagent Kit v3-600 cycle, Illumina, San Diego, CA, USA; Cat. No. 102-3003).

Sequencing Run QC

Once the sequencing data are available, it is important to evaluate the quality of the sequencing run (46). We use the following metrics and cut-offs for run quality: (1) the percentage of clusters passing the Illumina sequencer instrument filter (%PF) is 90% or greater and (2) the percentage of sequences above the Phred quality score of Q30 (which is equivalent to the probability of an incorrect base call of 1 in 1,000) is 70% or greater.[1] The use of paired sequences yields more information over a longer stretch of the V region than a single unpaired read. Paired reads also provide a consensus sequence at the termini of the reads, where the sequence quality tends to decline.

Sequence Data Quality Filtering

Prior to using the ImmuneDB pipeline (31), pRESTO [described in Ref. (28)] was used for quality filtering of raw Illumina MiSeq sequences. First, each sequence was analyzed with a sliding window of 10 base pairs. If at any point, the average quality score within the window fell below 20, the sequence was trimmed from its beginning to the end of the window. To correct for single bases with low-quality, any base that had a quality score less than 20 was replaced with an "N," indicating the uncertainty of the base call. Any sequence with more than 10 such N's or a total length of less than 100 bases was discarded from the analysis. **Code 1** shows the script used to run pRESTO with these parameters.

Code 1 | Sequencing data quality control: The bash script used to run pRESTO.

```
FilterSeq.py trimqual -s *.fastq
PairSeq.py \
    --coord illumina \
    -1 *R1*trimqual-pass.fastq \
    -2 *R2*trimqual-pass.fastq
AssemblePairs.py align \
    --rc tail --coord illumina \
    -1 *R1*pair-pass.fastq \
    -2 *R2*pair-pass.fastq
FilterSeq.py length -n 100 -s *assemble-pass.fastq
FilterSeq.py maskqual -s *length-pass.fastq
FilterSeq.py missing -s *maskqual-pass.fastq
```

Alignment and Generation of Unique Sequences

After the QC steps described in Section "Sequence Data Quality Filtering," each sequencing library has a set of high-quality sequences based upon the Phred quality scores. The next step was to use ImmuneDB (version 0.23.0) to determine the closest corresponding germline V- and J-gene for each sequence using an anchoring method (47). Once these V- and J-gene assignments were known, any sequence with less than 60% V-gene germline identity was discarded. Further, sequences were trimmed to IMGT position 150 to avoid primer biasing mutational analysis, and any sequence beginning after position 150 was removed.

[1]https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf (Accessed: November, 2017).

ImmuneDB allows for multiple V- or J-genes to be assigned each sequence. This assignment can occur in two ways. First, if two germline gene sequences are equally and maximally similar to the input sequence, both will be assigned. Alternatively, if the maximally similar gene(s) is statistically indistinguishable from other genes, given the average mutation and length of sequences in the sample, all such genes will be assigned to the sequence. If any sequence has multiple V-gene annotations that are not from the same family, the sequence is discarded, as this likely indicates the sequence contains errors such as a hybrid PCR product. Sequences with cross-family J-gene annotations are not discarded; however, because many of these genes, especially human IGH J1, J4, and J5, are very similar to each other (47). At this stage of the process, the number of total reads and the fraction of valid reads are computed for each sequencing library (replicate). If the total number of valid reads (those containing identified V and J genes and passing quality, length, and primer trimming) from one replicate is very different (five or more times lower) than the other replicate, the sample in question is subjected to re-amplification and re-sequencing. The fraction of valid antibody heavy chain gene rearrangement sequences depends upon the stringency of filtering, but with the above-described parameters is typically 75–90%.

Once sequences are assigned V- and J-genes, the unique sequences are collapsed across the entire subject. Two sequences are considered the same if they differ only in positions where either sequence contains an N. This results in a set of sequences, which are unique within the subject, and have a corresponding copy number. The next step is to group sets of unique sequences, which likely share a common progenitor cell into clones. To prevent spurious clones from being constructed, sequences with a copy number <2 across the subject, those containing a stop codon in the CDR3, or those having any window of 30 nucleotides falling below 60% germline identity (indicating a potential uncorrected insertion/deletion) are excluded from clonal assignment. For the remaining sequences to be included in a common clone, they must share the same V-gene, J-gene, and CDR3 nucleotide length. Further, each pair of sequences within the clone must share at least 85% CDR3 amino-acid similarity by Hamming distance. The script to run the ImmuneDB pipeline with these parameters is shown in **Code 2**.

Code 2 | ImmuneDB pipeline: The script used to run ImmuneDB. The germline files are included as supplemental files.

```
immunedb_admin create frontiers ~/configs
immunedb_identify ~/configs/frontiers.json imgt_human_v.fasta \
    imgt_human_j.fasta --trim-to 150 --max-padding 150
immunedb_collapse ~/configs/frontiers.json
immunedb_clones ~/configs/frontiers.json similarity
immunedb_clone_stats ~/configs/frontiers.json
immunedb_sample_stats ~/configs/frontiers.json
```

Tools for Immune Repertoire Visualization

The code for D20, cosine similarity, Hill number diversity plots, sample-based rarefaction curves, and clone metrics can be found at https://github.com/DrexelSystemsImmunologyLab/

frontiers-clone-size-scripts. Resampling plots were created based on the method in Ref. (48) as implemented in https://github.com/bochaozhang/sampleRarefaction, which directly query ImmuneDB. For the clonal overlap analysis string plot, clones were exported from ImmuneDB using the immunedb_export command. Using this, clone tracking plots were generated using VDJtools requiring the CDR3 amino acids and V gene assignment to match. The command line for clone tracking in VDJtools is:

```
$VDJTOOLS TrackClonotypes --i aaV \
    [sample1.txt sample2.txt sample3.txt . . . ] output_prefix
```

Data were converted to Boolean values that, in turn, were used to generate string plots in CIMminer.[2] For D20 and individual-based rarefaction analysis, VDJtools was used both to pool individual libraries (replicates were exported from ImmuneDB) and to generate the associated plots.

Data and Method Sharing

Raw data and accompanying sample data are available on SRA under BioProject number PRJNA476510. In compliance with the Adaptive Immune Receptor Repertoire (AIRR) standard (49), steps for processing the raw data with pRESTO and ImmuneDB are available on Zenodo.[3] Sequences annotated with ImmuneDB (those with VH gene and JH gene calls) are available *via* associated GenBank entries. All code to generate the database is provided in Code 2. Post-pipeline analysis scripts to generate data for plotting are available at (see https://github.com/DrexelSystemsImmunologyLab/frontiers-clone-size-scripts). Scripts used to generate plots are available upon request.

STEPWISE PROCEDURES

Two data sets were generated for this analysis (see overview in **Figure 1**). In experiment 1, 19 independently amplified antibody heavy chain gene rearrangement libraries were generated from the spleen of D207. These libraries are part of a much larger data set on human organ donor tissues that was described previously. In experiment 2, survey-level sequencing of antibody heavy chain gene rearrangements was performed in eight additional organ donors in a newly created data set specifically for this study. In experiment 2, spleen samples from each donor were subjected to flow cytometry to determine the B cell fraction and antibody gene rearrangements were separately amplified in duplicate from the same number of input B cells in all eight donors. After generation of survey-level and deep antibody heavy chain gene rearrangement sequencing data and initial quality filtering, gene alignment and grouping of related sequences into clonal lineages (see "Materials and Methods"), we are ready to evaluate the clonal landscape, which is the focus of this Protocol.

Different Metrics of Clone Size

The clonal landscape of a B cell population can be viewed as a continuum of information content, ranging from maximal information with individual sequence copies to minimal information wherein each clone is only counted once. Borrowing terms from ecology, one can view each sequence copy as an individual and each clone as a species. The unique sequence variants within the clones generated by SHM are akin to quasi-species. We consider three different metrics of clone size in bulk population data: copies, instances, and unique sequences. **Figure 2** illustrates these metrics for two hypothetical B cell clones in two separate sequencing libraries (replicates). The copies are like the individual B cells (although some B cells may have more than one sequence copy and some may have none, depending upon the depth of sequencing and on the level of sampling). Also, as discussed in the Section "Introduction," it should be emphasized that extrapolating from copies to cells is challenging with bulk sequencing methods because there can be primer amplification bias. The next level down in information content is to ignore the copies and only count the number of times each unique sequence variant appears in each of the sequencing libraries. We call this measure instances. If the same sequence appears in both replicates, it is counted twice. If it appears in only one of the two libraries, it is only counted once. This measure is less sensitive to PCR amplification bias because the same bias has to occur with the same clone in independent replicates. But this measure is also affected by the depth of sequencing and the level of sequencing error, which can introduce spurious mutations that may be counted as unique sequence variants, depending upon how the data are filtered. The next level down is unique sequences. With unique sequences, all of the identical sequences from a single subject are grouped together and each unique sequence is only counted once. Here again, the measure can be influenced by sequencing depth and sequencing error. Finally,

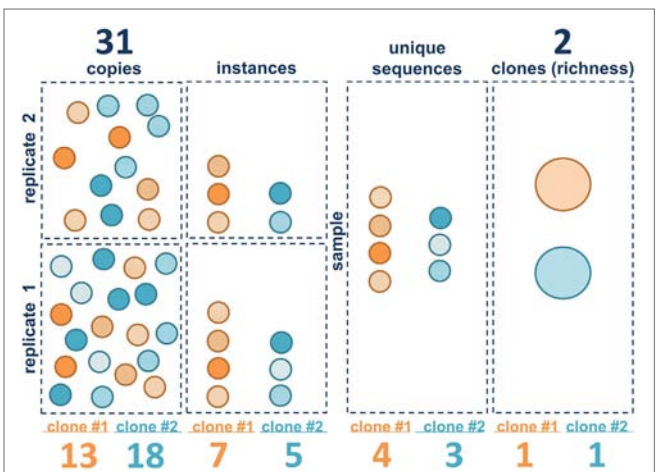


FIGURE 2 | Clone size measure definitions. The first three vertical panels represent the three measures of clone size in this paper: copies, instances, and unique sequences. Each circle represents a B cell. Sequence variants within clone 1 (orange) and clone 2 (blue) are represented by different patterns in the circles. The right panel contains the clones (large ovals). Squares represent within the left two panels represent individual sequencing libraries whereas the entire collection of replicates from a single sample is shown in the right panels. Summed values (for both clones) for each metric are given above. Values for each individual clone are given below.

[2]https://discover.nci.nih.gov/cimminer/oneMatrix.do (Accessed: November, 2017).
[3]DOI: 10.5281/zenodo.1292010.2017.

the most minimalist measure is to simply count the number of different clones, counting each clone only once, analogous to species richness. Richness does not capture information about clone size, only clone number.

It is important to emphasize that we are focusing here on measures of clonal size in bulk sequencing data. There are other metrics that could be used to estimate clone size, but they tend to be impractical in bulk libraries. For example, one could count instances of clones (rather than instances of unique sequences). Counting clone instances is a Boolean metric (presence or absence) wherein each clone is counted only once per sequencing library, if it is present. Thus, if there were two sequencing libraries, Boolean instances for all of the sampled clones would be either 1 or 2. A Boolean instances measure requires a very large number of sequencing libraries to be sensitive to differences in clone size (9). Such a metric is especially useful if single cell or digital droplet PCR is being used to measure clone size because, in those cases, hundreds or thousands of "libraries" can be queried.

Initial Rearrangement Metadata Assessment

As an initial check of the sequencing data, we evaluate the numbers of copies, instances, unique sequences, and the number of clones (clones that are found in more than one sequencing library are only counted once). In **Table 4**, we present these data [obtained through the ImmuneDB pipeline (31)] in aggregate form for all eight donors at two libraries per donor and for one donor (D207) at 19 libraries.

We begin by comparing the estimated number of B cells to the number of unique sequences. Under conditions of maximal diversity [with each B cell in the population under study harboring at least one different heavy chain rearrangement, and assuming 100% yield and 1.4 VDJ rearrangements per cell on average because some cells will harbor more than one rearrangement (50)], we use the following formula to approximate the maximal number of B cell rearrangements per ng of input DNA:

$$\text{max number of rearrangements} = \frac{\text{ng} \times 1000\ \text{pg/ng} \times 1.4\ \text{rearrangements/cell}}{6.7\ \text{pg/cell}} \quad (1)$$

From a pure (flow cytometrically sorted) B-cell population, another useful estimator of the maximum number of cells (assuming 1.4 rearrangements/cell) is:

$$1\ \text{ng B cell DNA} \sim 150\ \text{cells} \quad (2)$$

In experiment 2, we accomplished normalization of the sample size by measuring the B-cell fraction using flow cytometry (**Table 2**) and then used the B-cell fraction to create the same number of B cell equivalents for each amplification. We used 50 ng of B-cell equivalent DNA in each replicate. Using the equations above, the maximum number of rearrangements that should be found in any one sample is 50 ng × 2 replicates × 1,000 pg/ng × 1.4 rearr/(6.7 pg/cell) = 20,895 unique rearrangements. All donors exceed this predicted maximum number of unique rearrangements. 20–40% of all of the sequences are present in one copy (**Figure 3**). Many of these sequences represent sequencing errors, whereas others represent infrequent clones. Because sequence copies are computed across all of the replicates, D207, with 19 replicates, has the lowest fraction of single copy sequences. In contrast to the excess of unique sequences, the number of clones is much closer to the theoretical maximum number of rearrangements.

Reducing Sequencing Errors

To reduce the contribution of sequencing errors to clone size measures, one can employ several different strategies, often in combination (19, 46). The first is to use more stringent quality thresholds to filter the data. We can use quality scores of 30 or higher for bulk sequencing runs in which SHM is being analyzed.

TABLE 4 | Sequencing data summary.

Subject	Libraries	Copies	Instances	Uniques	Clones
D207	19	5,526,691	1,921,080	1,895,669	136,876
D267	2	632,949	508,219	507,559	21,717
D275	2	446,178	340,315	339,787	15,161
D279	2	467,435	353,311	351,088	12,405
D287	2	537,427	360,998	360,967	9,111
D301	2	388,677	261,876	259,982	8,861
D302	2	504,337	380,141	379,360	14,333
D305	2	477,282	371,135	371,071	20,225
D309	2	412,672	361,225	360,807	17,186

Each library was generated from separately amplified aliquots of DNA. Copies, instances, unique sequences, and clone numbers (see text) were compiled for all VDJ rearrangements (productive and non-productive) across all of the libraries from each donor using ImmuneDB. No copy number cut-off was used when computing the numbers of unique sequences or clones. The data from D207 have been published previously (9), but, here, we are only showing the data from that donor that were generated with the FR1 + JH primers, so only some of the libraries that were generated for the previous publication are shown here.

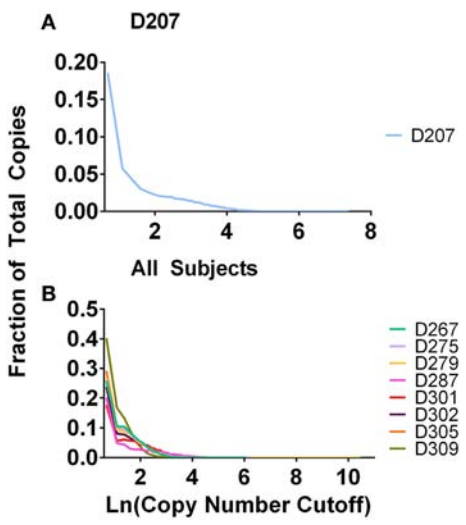


FIGURE 3 | Fraction of total copies vs. clone copy number cut-off. **(A)** D207 (19 replicates); **(B)** all other subjects (survey-level sequencing).

In addition, filtering of primer sequences, length trimming, and treatment of in/dels are important. The second approach to minimize sequencing error is to use a copy number filter. We typically use a minimum copy number of 2, which eliminates a lot of low copy sequences that are generated through sequencing error, but also eliminates valid low-copy sequences. One can rescue some of these valid low-copy sequences by computing clonal lineages across different replicates, which ImmuneDB can do (31). Thus, even if a sequence has a copy number of 1 in one library, if that same sequence is found in another library, it now has a copy number of at least 2 and can pass the filter. A third approach is to filter the data based upon instances or the number of times that the same unique sequence is found in different replicates. A fourth approach is to employ molecular barcoding (51). Molecular barcoding is typically performed on RNA samples and introduced *via* primers with variable sequence tags ("barcodes") at the cDNA synthesis step [for a detailed method that can be applied to bulk populations, see Ref. (11)]. At sufficient sequencing depth, alignment of sequences with the same barcode is performed and used to generate a consensus sequence that is virtually free of sequencing errors. RNA-based assays tend to be lower throughput and require far more sequencing, increasing cost. Another approach to minimizing sequencing errors is to perform rolling circle amplification (52).

A final consideration involves processing of the sequencing data, which may affect how somatic mutations are identified and counted. For example, if the subject has one or more novel V gene alleles, rearrangements with these V genes may be scored as being mutated rather than matching the novel germline sequence. Software tools have been developed to search for novel alleles within individual samples, providing an individual reference database of germline and putative germline alleles against which sequences from the same individual are compared for mutation [see arXiv:1711.05843 (q-bio.PE), https://github.com/psathyrella/partis, and (53, 54)]. Filtering of low copy number sequence variants can eliminate valid sequences. To recover some of these sequences, one can construct algorithms that identify sequence variants that are shared (even among single copy sequences) in separate replicates from the same individual. ImmuneDB can do this handily because it takes all of the sequences from an individual into account when in constructs clonal lineages.

Copy Number Cut-Offs and Clone Numbers

Table 5 shows the numbers of clones at different clone size thresholds (i.e., the clone size cut-off used in this illustration is instances) in each donor. As one would expect, the number of clones decreases as the threshold increases. One could envision setting the copy number threshold to be near a number that corresponds to the maximum number of unique rearrangements. An alternative approach is to discard clones at some fractional cut-off. For example, one could discard clones having sequences that fall below 50% of the mean copy number frequency of the sample. Either approach results in biases in the data. In the case of an absolute copy number cut-off, one runs the risk of discarding infrequent clones and the stringency of this cut-off will vary based upon the depth of sequencing: samples that are not as deeply sequenced (and have lower average copy numbers per template molecule) will lose more data than samples that are deeply sequenced. On the other hand, a relative copy number cut-off can be influenced by the copy number distribution of the sample. If a sample has very large clones in it, these large clones can skew the average copy number value and lead to excessively stringent filtering. Despite the fact that this experiment was controlled for differences in the B cell fraction, different numbers of clones were observed in different donors. Some of these differences appear to be due to intrinsic biological differences between subjects in their clonal landscape. Consistent with this idea, **Table 5** also shows that the distribution of clones at different size cut-offs is not uniform across the different donors.

TABLE 5 | Clone numbers with different instance cut-offs.

Subject	C1	C2	C3	C4	C5	C10
D207	136876	34441	12419	6,170	3,598	891
D267	21717	9,174	2,770	944	373	30
D275	15161	6,794	2,316	942	436	65
D279	12405	5,706	2,010	804	378	34
D287	9,111	5,910	3,669	2,047	1,195	98
D301	8,861	5,310	1,969	849	434	90
D302	14333	7,378	2,882	1,190	590	81
D305	20225	7,823	2,505	754	260	7
D309	17186	3,467	721	248	100	10

The numbers of clones that were amplified with FR1 + JH primers are shown for different clone size cut-offs in each of the donors. The clone size cut-offs are given in instances (the minimum number of replicates that contain at least one sequence from the clone). C1 (at least one instance) is equivalent to no cut-off and contains all clones irrespective of size. C2 is two or more instances, C3 is three or more instances, and so on.

Visualization of Large Clones

A quick way of drilling down on the largest clones in a sample is to determine the fraction of the total copies that is comprised of the 20 highest copy number clones, also known as D20, which is defined in Eq. 3 where c_i is the copy count of clone i, and T is the total clone copy count.

$$\mathrm{D20} = \sum_{i=1}^{20} \frac{c_i}{T} \tag{3}$$

The D20 percentage can be over 90% in a patient with B-cell malignancy. Conversely, a blood sample from a healthy adult will tend to have many smaller clones and a corresponding D20 value of 1–2% or less. **Figure 4A** shows the D20 fraction and **Figure 4B** shows the copy number fraction of the 20 top copy rearrangements in survey-level sequencing from each of the nine organ donors. D279 and D301 have D20 values that exceed 18% of total copies. In the case of D279, a single rearrangement comprises over 15% of total copies. Furthermore, both of these donors have a rearrangement that is at least three times more frequent than the next most frequent rearrangement. In addition, two other donors, D207 and D287, each have top copy rearrangements that exceed the next most frequent rearrangement by more than threefold, but neither of these rearrangements exceeds 5% of total copies. The combined use of a frequency cut-off (such as 5%) and fold-change cut-off relative

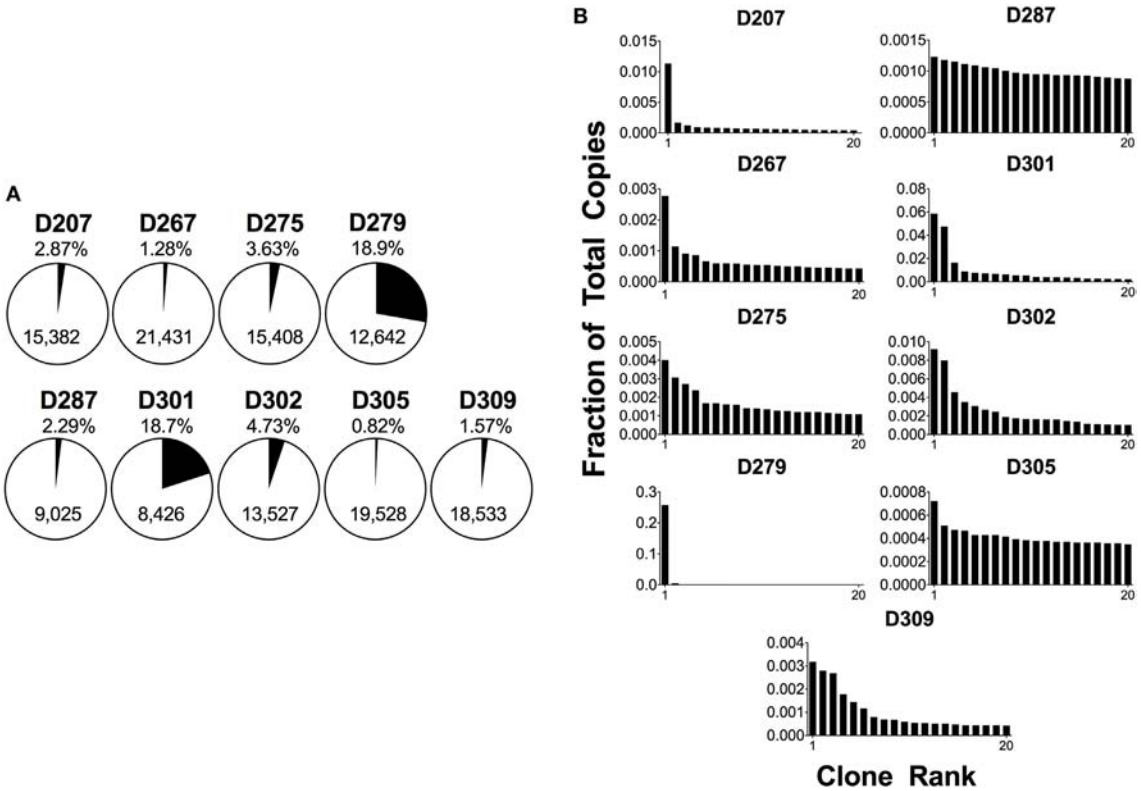


FIGURE 4 | Analysis of 20 top-copy clones in different donors. **(A)** The 20 clones with the most copies in each subject are presented as a proportion of total copies in a given sample; the percentage is shown above the pie chart. The total number of clones per sample is in the body of the pie. **(B)** Histogram plots of the top 20 copy number rearrangements. The fraction of total copies for each rearrangement is plotted vs. the clone rank.

to the polyclonal background (such as threefold) provides greater confidence in declaring a true clonal expansion from oligoclonality. Furthermore, in both D279 and D301, there are several thousand B-cell clones, favoring clonal expansion over oligoclonality. At low B-cell numbers, PCR can be less efficient and jackpots (the disproportionate amplification of one dominant sequence) are more likely to occur (55). To further evaluate if these are bonafide clonal expansions, we measured the fractions of total copies in each individual replicate. In both cases, the fraction of total copies for each of the top two rearrangements was similar between the two replicates (the top copy rearrangement in D279 was 16% of total copies for replicate 1 and 17% for replicate 2; the top copy rearrangement in D301 was 5% in replicate 1 and 4% in replicate 2). The reproducibility of these values suggests that these rearrangements are present in one or two large expanded clones. If this analysis were being performed on peripheral blood samples, finding rearrangements of this size would be considered worrisome for pathologic clonal expansion, but we do not yet know the normal "reference range" of clone sizes in human tissues.

Figure 5 shows how different size measures compare for the top 20 ranked clones in D207. While many of the clones in the top 20 are found in all three ranking systems, their position in the ranking can shift and some clones are only found in a single rank. For example, clone #180721 (ranked eleventh by unique sequences) is not found in the top 20 clones ranked by instances or ranked by copies. One contributing factor to the difference in ranking is that instance and unique sequence-based measures have a much smaller dynamic range than copy number measures. A clone may have 10 times as many copy numbers but the same number of instances. There are far more "ties" in the instance-based measure. Biological differences may influence the rankings. For example, if there is a very large clone with minimal SHM, it may rank higher in a copy number-based rank than in a unique sequence-based rank.

Diversity of Clones

In order to visualize the clonal landscape at different clone size ranges, one can plot the diversity of clones and give different weights to clones that are smaller or larger in size, as described in Ref. (39, 48). Here, the true diversity is given by:

$$ {}^{q}D = \left(\sum_{i=1}^{R} p_i^q \right)^{1/(1-q)} \tag{4} $$

The equation for, ${}^{q}D$, true diversity at order (Hill number) q. R is richness, in this case the total number of clones, and p_i is the proportional abundance of clone i. The abundance can be the proportional number of copies, unique sequences, or instances. In this equation, diversity is a unitless number that

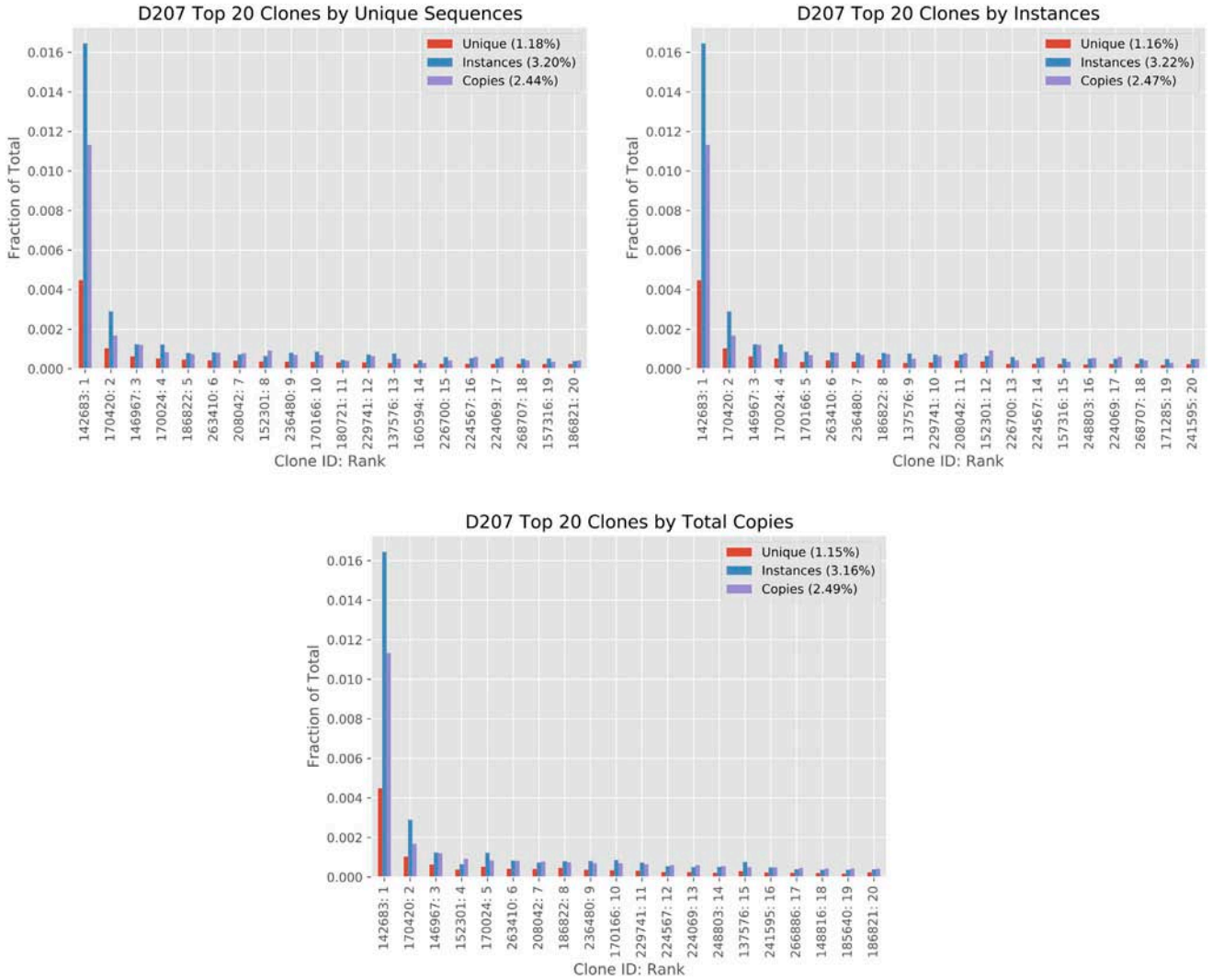


FIGURE 5 | Analysis of top 20 clones by different clone size measures in D207. The top 20 clones are ordered in the three columns by total copies, instances, and unique sequences using the 19-library data set on D207. The *x*-axis shows the Clone ID (in ImmuneDB) followed by the rank. The clone size measures are given in fractions. The percentage within each chart legend indicates the sum of the top 20 ranked clones in the chart for each measure (please refer to the text for clone size definitions).

refers to the "effective" number of different clones in the population [see discussion in Ref. (39)]. Diversity is weighted by the parameter q. When q is 0, D is the number of different clones in the population. When q approaches 1, the diversity of each clone is proportional to its abundance (i.e., it is the weighted geometric mean). When q is greater than 1, larger clones are given more weight.

In **Figure 6A**, diversity is calculated for all clones (clones with 1 or more instances, marked C1) in all of the donors except 207. As the order increases, the diversity diminishes because there are far fewer large clones than there are small clones. Unlike the copy number cut-off plots, the diversity plots provide greater resolution of the representation of large clones in the population. At higher orders, D287 has a longer tail of medium to large-sized clones than the other donors. But among clones with at least five instances (C5 clones), D279 has the greatest diversity at higher orders.

In **Figure 6B**, diversity is calculated for clones of different size cut-offs in D207 (C1–C5 and C10 instances). Although we are using instances to filter the clones being considered, we then re-analyze the clone sizes of all of the clones meeting the instance cut-off using the three different size metrics: unique sequences, instances, and copies. Our test of sampling sufficiency is based on resampling and thus considers clone sizes in unique instances. However, at the clone size, we deem relevant when analyzing clonal diversity or overlap, there is still added information in considering different aspects of clone size. As stated above, both copy number and unique sequence number (and thus also instances) can be affected by PCR and sequencing artifacts. However, both also represent different indications

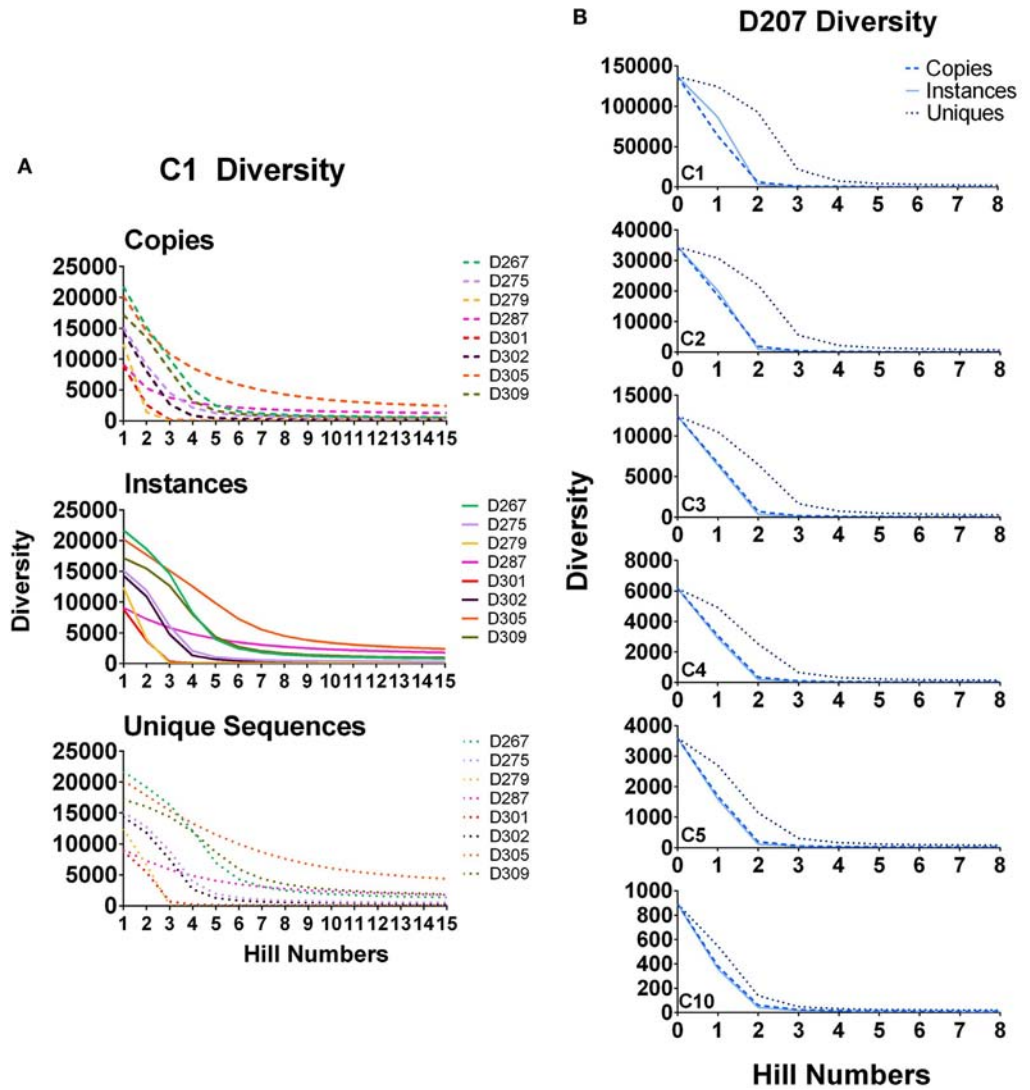


FIGURE 6 | Diversity at different orders. **(A)** Diversity of C1 clones (all clones) calculated for all donors at survey level by instances, copies, and unique sequences; **(B)** Diversity of C1, C2, C3, C4, and C5 clones in D207. True diversity is calculated using copies, instances, and unique sequences (see text). Calculations are performed at different orders (Hill numbers). Higher orders give more weight to larger clones.

of expansion—diversification by mutation for unique numbers and proliferation of specific types for copy numbers and instances. Thus, while they may not be totally faithful only to these dynamics, we do count and compare them, along with simply tracking the level of presence of clones in different samples.

Descriptive Measures of Clonal Diversity and Evenness

Beyond the analysis of very large clones, one can study the distribution of clones within a sample using various descriptive measures of diversity or evenness or both (56). It is important to consider a sample may not be a single replicate, but multiple replicates from a common source. Shown in **Figure 7** are analyses using 19 independent spleen replicates, which were analyzed from one subject (D207) and stratified by different clone size cut-offs based upon instances. The simplest metric is to count the number of species (a.k.a., richness, R) at each clone size cut-off using different clone metrics. Richness, which is equivalent to diversity of order 0, specifies the total number of species in a sample. As expected, richness decreases with increasing clone size cut-offs and decreases more rapidly for unique sequences and clones than for instances and total copies. Note that richness alone does not account for clone size; two samples with the same number of clones but drastically different clone sizes will still have the same richness. Therefore, it is useful to examine other metrics, which measure clonal size distribution in addition to species diversity. The Shannon entropy [H (57)] takes the number of individuals of each species (i), the proportion of sequences in a given clone over all of the different clones being measured

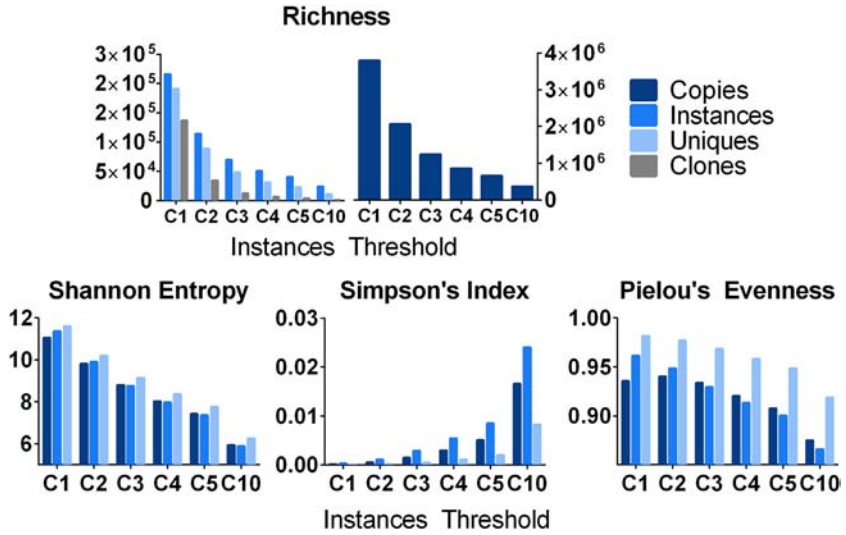


FIGURE 7 | Various metrics of diversity and evenness. Count: the total number of clones and the number of instances, unique sequences, and copies that comprise them. Shannon Diversity: the Shannon diversity of clones when their size is defined as instances, unique sequences, and copies. Pielou's Evenness: measures how evenly distributed clone sizes are. A value of 1 indicates all clones are of the same size, and a value of 0 indicates maximal diversity in size. It is defined as the Shannon Diversity divided by the maximal diversity [equal to ln(*S*)] where *S* is the number of clones. Simpson index: diversity at Hill order 2. This can be interpreted as the probability of randomly drawing two copies/instances/unique sequences that belong to the same clone.

(p), and the number of different clones (R) into account, as shown in Eq. 5:

$$H = -\sum_{i=1}^{R} p_i \ln p_i \tag{5}$$

The Shannon entropy can be computed using copies, instances, or unique sequences as metrics for the numbers of individuals in each species (clone). For a given threshold (instance cut-off), all clones failing to meet the cut-off are filtered out and all of the remaining clones are used to compute p_i. If nearly all of the sequences in a sample are found in one clone, the Shannon entropy approaches 0. Conversely, if all of the clones are equally abundant, the Shannon entropy approaches the natural logarithm of R. The Shannon entropy can also be computed with different logarithm bases.

Simpson's index measures the true diversity (Eq. 4) at Hill order of 2. Simpson's index measures the likelihood of encountering two sequences derived from the same clone when sequences are drawn at random from a given sequencing library or, in this case, a collection of 19 sequencing libraries from D207. It is defined by Eq. 6, where, as before, R is the richness and p_i is the proportional abundance of each clone.

$$\lambda = \sum_{i=1}^{R} p_i^2 \tag{6}$$

Counts, the Shannon entropy, and Simpson's index are all influenced by sampling and by the depth of sequencing (which can cause spurious concentrations of sequences within individual clones in over-sequenced samples). Another frequently used metric, Clonality, takes on normalized diversity values ranging from 0 (maximally diverse) to 1 (monoclonal). Unlike entropy, clonality, C, measures the loss of diversity and can be represented as the inverse of entropy (H). One can also quantify how uniform clone sizes are using a measure of "evenness." (58) Pielou's Evenness is defined by Eq. 7:

$$P = \sum_{i=1}^{R} \frac{p_i \ln p_i}{\ln R} \tag{7}$$

Samples where most clones are of similar size will have an evenness measure closer to one whereas samples with predominant rearrangements will have a lower value. In D207, there are a few large clones, but the majority of clones are small, hence the overall evenness is very high. As smaller clones are excluded, the evenness decreases. When R is not known or if the clone size copy number cut-off is uncertain (resulting in variable inclusion/exclusion of low copy number clones across different sample types), this ratio can fluctuate and other ratio-based measures of evenness such as those described by Peet (59), may perform better (60).

Figure 7 shows that different metrics of richness and evenness (or hybrid measures of both) yield different results at different clone size cut-offs. Furthermore, when comparing results on different populations, the results from one measure do not necessarily translate intuitively to the results of another measure because the size distributions of clones in the different samples vary. For example, **Figure 4** shows that D309 has a steep copy number cut-off curve at low copy number counts with a high proportion of low copy number clones, but also a shorter tail of higher copy number cut-off clones compared to most of the other donors. Conversely, D287 has fewer low copy clones, fewer intermediate size clones, but a longer tail of larger clones. For these reasons, we and others recommend analyzing the clonal landscape with several different metrics as well as plotting clonal diversity at different Hill numbers to visualize the clonal

landscape in different clone size ranges (46). Some have also advocated using a collection of scalable and normalized diversity metrics to study immune repertoires (61).

Rarefaction Analysis to Power Clone Size for Clone Tracking

Tracking clones through different samples requires powering the analysis to detect clones of a given size. If one does not do this, then, the lack of clonal overlap between two samples could be due to insufficient sampling. The increment in finding new clones with additional sampling can be evaluated using rarefaction analysis (39, 40). Stratifying clones by size, one can generate rarefaction-based estimates of sampling adequacy, as illustrated in **Figure 8**. When there are only modest amounts of sequencing data (as we have here with only two replicates per donor), one typically relies on individual-based rather than sample-based rarefication analysis (**Figure 8A**). In this analysis, the diversity is the richness (number of different clones, analogous to the number of different species) and the sample size is the number of sequence copies (analogous to the number of sampled individuals). Different donors have different levels of diversity, despite the fact that we controlled for the B cell content. In all donors, the diversity is substantially lower than the number of sequences because several of the same or highly similar sequences comprise each clone. In some donors, such as D287, there many more sequences per clone than other donors (such as D267). One caveat to this analysis is that using individual-based rarefication can be unreliable when we count clone size by copy number: copy numbers may be inaccurate if sequencing depth and PCR amplification efficiency are not properly controlled.

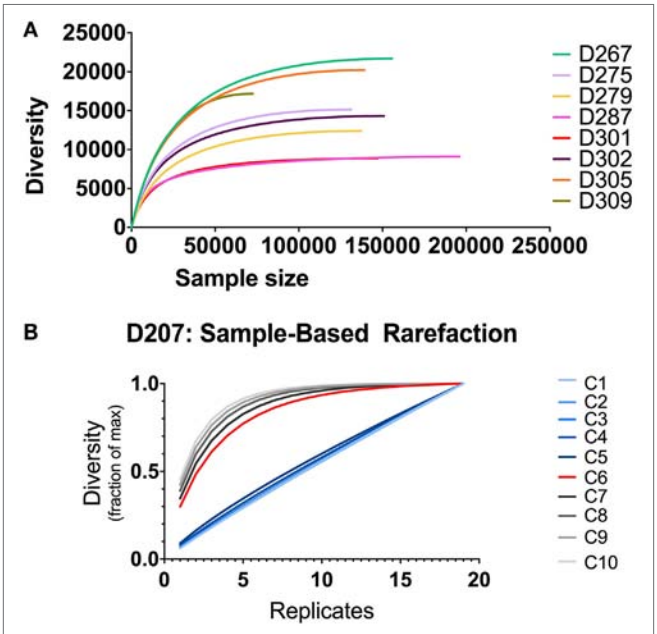


FIGURE 8 | Rarefaction analysis. **(A)** Individual-based rarefaction with different sample sizes. Individual-based rarefaction plots were computed with VDJtools using unique sequences (no copy number cut-off) and subsampling to impute expected clone numbers. **(B)** Rarefaction analysis at different clone size cut-offs in D207. The expected number of clones in the population is given on the *y*-axis and the number of samples (replicates) is shown on the *x*-axis. Each line depicts rarefaction curves for clones with a given size cut-off. Clone size *C* is measured in instances (see text).

Our most accurate indication of clone size and sequence abundance is the number of times we observe things in independently generated sequencing replicates. Beyond its ability to correct for sample-specific copy number inaccuracies, counting instances has the added advantage of being less influenced by samples that derive from a non-homogenous population (9). With more extensive sampling, we can perform sampling-based rarefaction analysis, as illustrated for D207 in **Figure 8B**. In this analysis, with larger clones (corresponding to size cut-offs of 6–10 instances), the curves level off. Rarefaction analysis, coupled with clone size, can be used for power detection of clonal overlap or clone tracking between samples. In this example, clones with an instance cut-off of 6 (C6, red curve) are the optimal size for overlap analysis in this data set: they are the smallest size clone with a rarefaction curve that levels off. As one would expect, there is a trade-off between the amount of sampling required and the likelihood of capturing a clone. However, when two populations have very large numbers of overlapping clones or very large clones, a lower capture threshold may be sufficient to adequately sample clones of interest.

Clonal Overlap Analysis

To determine if two samples contain overlapping clones, the most straightforward thing to do is to count the number of clones that overlap and compare that number to the total number of clones in each of the samples. This type of counting is the basis for generating a Venn diagram. However, such Venn diagrams are hard to compare quantitatively as they do not take different sizes of clones into account and they become visually cumbersome when more than two samples are being compared. In lieu of this approach, several metrics have been developed to quantify overlap, giving weight to clone sizes in the samples, including the relative overlap diversity, the geometric mean of relative overlap frequencies, and the clonotype-wise sum of the geometric mean frequencies. This metric is easy to calculate and is not overly influenced by the relative sizes of clones and, as it ranges from 0 to 1, it is easy to compare across experiments. Here, we illustrate the use of the cosine similarity metric for quantifying overlap between two-sample pairs. The equation for the cosine similarity metric is:

$$cosine\ similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (8)$$

The cosine similarity is computed for two samples *A* and *B*, for example, replicate 1 from D207 and replicate 2 from D207. Both *A* and *B* are vectors of length *R*, where *R* is the number of unique clones across the two samples. The value of A_i or B_i is the abundance of clone *i* in sample *A* or *B*. The abundance can be any of the measures of clone size such as copies, unique sequences, or instances.

The cosine metric is useful for evaluation of clonal overlap between two samples, but does not provide a means of comparing overlap of clones that span multiple samples.

To visualize clones found in three or more samples, we use string plots. String plots based on Boolean values (presence or absence of a clone in a replicate), are shown in **Figure 9** for all clones (n = 14,543) that overlap in at least two replicates in D207. In these plots, the strings (horizontal lines) represent the individual overlapping clones. The overlapping clones comprise 11% of all C1 clones in the 19 libraries of D207 (overlapping clones have been removed from this total C1 clone number). Note that over 90% of these overlapping clones in the entire data set have already been discovered in the first 10 sampled replicates. In these plots, the strings can also be colored based upon a metric of clone size such as percentage of copies within a sequencing library (34).

One can add other dimensions to string plots, as we did with "line circle plots" in our analysis of clonal representation in different tissues (9). In line circle plots, each line represented a clone, circles indicated membership of that clone in a tissue, the size of the circle indicated the number of copies of the clone in the tissue samples, and colored wedges in the circle indicated what fraction of sequencing libraries from that tissue contained members of the clone. Thus, in line circle plots, one can display three different features of a clone—its tissue membership, its copy number in the different tissues, and its instance number—within replicate libraries from each tissue. The plots can of course be modified to display different parameters, depending upon the comparisons of interest. String plots provide a means of visualizing overlapping clones, but it is important to remember that they focus exclusively on the overlapping clones. The appearance of a string plot can be misleading if the samples that are included in the plot are of unequal size or clonal composition. The total number of observed clones within each sample being compared needs to be considered to determine if the numbers of overlapping clones reflect meaningful overlap or merely sampling differences. To visualize all of the clones in the samples being compared, Venn diagrams can be used.

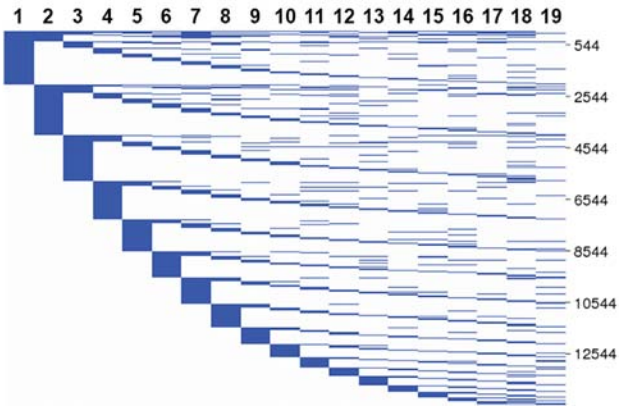


FIGURE 9 | Tracking of shared clones across replicates in D207. All clones that were found in two or more replicates are shown. Each column is a replicate. Each horizontal line is a clone. The numbers of shared clones are given on the scale to the right.

ANTICIPATED RESULTS (PITFALLS, ARTIFACTS, AND TROUBLESHOOTING)

Overview of Clone Size Evaluation

Here, we present a series of analytical approaches to evaluate B-cell clone size in bulk populations. We describe three basic types of immune repertoire measures that are impacted by clone size. The first is the fraction of total copies that harbor the one or two-most frequent rearrangements. This fraction is typically higher than 0.05 for a malignant clone with a polyclonal background and may require an even higher cut-off with an oligoclonal background. We also determine the fold increase of the most frequent rearrangement compared to the next most frequent rearrangement in the sample. Ideally, this change should be threefold or greater. These criteria tend to be sufficient for relative or approximate measures of clone size. For smaller clones or finer resolution of clone size, calibration to an external standard (such as cloned rearrangements that are spiked into the sequencing reaction), single cell counting, or limiting dilution analysis may be required. We perform sequencing reactions in at least duplicate to rule out spurious clonal expansions due to oligoclonality (accompanied by poor reproducibility in the replicates) or PCR jackpots.

The second measure is clonal diversity, which can provide insights into immune competence or robustness of a targeted immune response, such as tumor infiltrating lymphocytes in a biopsy specimen. At the bulk population level, clonal diversity needs to be visualized at different orders (Hill numbers) to give weight to clones of different size in the population. If different samples are to be compared, it is important to normalize the input DNA of the sequencing library for the B cell content in different samples. This can be accomplished by sorting cells with a specific phenotype or by performing FACS analysis and determining the B cell fraction in the cell suspension, as we did here. If we had not controlled for the B cell content, then differences in diversity could have reflected differences in sampling rather than true differences in diversity. A second important consideration with diversity analysis is to visualize the clonal landscape at different clone sizes. We describe two ways to accomplish this: first, plot the number of clones at different clone size cut-offs, and second, plot the true diversity at different orders, giving different weight to clones of different size ranges. Two individuals may have very similar small clones, but one person may have many intermediate-size clones while another may have a few really large clones. If one only visualized the data with a single diversity measure, one might miss features of the clonal landscape that distinguish one sample from another.

The third measure is clone tracking, including clonal overlap analysis. The ability to track a clone depends upon its frequency in the population; thus, the respective sizes of the clone and the

population in which the clone resides both matter. Starting with sequences from bulk populations, we recommend using sample-based rarefaction analysis to determine the clone size that can be detected reliably and the number of sequencing replicates needed to adequately sample the clone. The null hypothesis of this analysis is that the clone size is the same in the two populations being sampled and evaluated for clonal overlap, which often is not the case. We recommend tailoring the clone size metric to the types of clones being compared. For clones that harbor substantial somatic mutations, the number of unique sequence variants per clone may be a useful clone size metric. On the other hand, if samples differ in their sequencing depths, it may be useful to deploy an instance or even a Boolean (presence vs. absence) metric for clone size, although such metrics typically require very extensive sampling and may be impractical. Finally, if amplification efficiency, sequencing depth, and B cell content are well controlled, it may be possible to use copy number fractions. To visualize clonal overlap, we recommend using the cosine statistic for two-sample comparisons and string plots for three or more sample clone tracking experiments.

Real-Life Limitations and Alternative Approaches

Samples may be limited in quantity or quality, or we may not know the B-cell content. Small and low-quality samples can be encountered in fixed tissue samples. Due to poor DNA quality, it may only be possible to generate short amplicons from such samples, potentially reducing the fidelity of V gene assignment (47). Furthermore, modest numbers and/or fractions of B cells in such samples can increase the likelihood of PCR jackpots and the accumulation of sequencing errors due to over-sequencing of the few templates that are present. Clone size measurements in such samples may not be reliable or even possible.

To judge the adequacy of the library, we use the quality metrics described in Sections "Sequencing Run QC" and "Sequence Data Quality" Filtering and look at the number of clones. If the sample has fewer than 50 clones and/or 1,000 valid sequences, it may be challenging to identify a dominant clone unless the sample is nearly monoclonal. Replicate amplifications from oligoclonal samples lacking clones will tend to reveal different clones in the replicate, whereas samples with clonal expansions will reveal consistent amplification of the same clones if they are large enough. Sometimes it is not possible to know the B-cell content in a sample. If the B-cell content is unknown, we calculate the copy number distribution and average copy number, and then use a fraction of the average copy number as the copy number cut-off. For example, if the average copy number is 100, the copy number cut-off might be as high as 10 or 20, whereas if the average copy number is 2, there might be no cut-off or only sequences with a copy number of 1 may be eliminated.

Community Efforts to Validate and Standardize Repertoire Analysis Tools

All of the clonal size measurements introduced for studying bulk populations of B cells assume that the annotation of clones (and, therefore, genes) is correct. There are many tools that claim to achieve this, including ImmuneDB, which was used for this paper. However, there has been no robust method for determining how well the clonal associations produced approximate the true clonal landscape. It may be possible to validate tools under specific conditions when the clones are known *a priori*, but currently there is no universal standard by which tools can be tested.

There are at present several hurdles to creating such an estimate, some of which may be insurmountable. We start our germline association of sequences from an expanded population whose somatic and germline history is ill-defined. To estimate the accuracy of our association, we would need to know which germline genes, and what copy numbers are present in the subject and what kinds of selection pressures have created the gene segment usage in the active repertoire. This final requirement is quite difficult, as selection has been shown to skew repertoires significantly and in a very individual way (62). In addition, there are differences that simply cannot be detected after the fact, such as discrimination between germline genes that are too similar to tell apart (47).

There is a goal of generating such a standard within the AIRR Community, with active discussion in the B-T.cr forum.[4] An alternative method is to generate sequences *in silico* with known V- and J-gene assignments (63), run a gene inference tool on the sequences, and see how well the results match the input. However, this approach assumes that we can write software that adequately mimics the underlying biological processes.

Use of Multiple Tools on the Same Data Set

One approach to validation of an analysis method, when lacking a "gold standard" for comparison, is to use multiple clonal assignment tools and compare the results. Even though the results will likely not be identical, at least for the large clones, one would expect similar clonal assignments to be produced. The size analyses from this paper could then be applied at the threshold of clone size to which the tools generally agree.

Starting with the sequencing data table, we begin with a back-of-the-envelope equation on the maximum predicted number of gene rearrangements. If the number of unique sequences in the sample exceeds this value, we look at other quality metrics to determine if there is adequate filtering of the data to remove sequencing errors. If the number of unique sequences in the sample is 10-fold or more below the maximum predicted number of rearrangements (Eq. 1), we review the experiment to determine if there is any explanation for the low recovery of rearrangements, such as a low-quality sample or perhaps an unexpectedly low B cell fraction. We also look more closely at the data filtering to see if we are discarding too much data in the filtering process.

While beyond the scope of this paper, the method for grouping related sequences into clonal lineages could also influence

[4]https://b-t.cr/t/how-to-decide-on-an-example-data-set-to-use-for-testing-software/129 (Accessed: November, 2017).

downstream measurements of clonal diversity, size, and overlap. Several different approaches for associating antibody sequences into clonal lineages exist (29, 32, 35, 38, 64). Different methods as well as different parameters within individual methods can be tested to determine if the findings are robust. These methods make different assumptions and can result in different stringencies of clonal association under different conditions [such as different levels of SHM, discussed further in Ref. (9)]. We often try processing data with two or more different pipelines, such as MiXCR (34) and ImmuneDB (31). Getting similar answers with both pipelines encourages us that the result is more likely to be robust. When there are discrepancies, they can be due to differences in sequence quality filtering or the clonal lineage assignment steps. In ImmuneDB, one can compare the total number of sequences to the number of "identifiable" sequences, which are those to which a V and a J gene (or gene tie) have been assigned. If there is a massive loss of sequences going from total to valid sequences, the pipeline may be filtering out sequences of interest or there may be poor sequence quality. Another quick sanity check is to look at the fraction of sequences with productive rearrangements. In mature B cells, a low fraction of productive rearrangements (<75%) is usually an indicator of poor sequence quality.

For clone size analysis, if only clones with at least six instances (C6) are inferred similarly with multiple tools, it may be possible to limit the analysis to clones of size C6 and above. However, it may not always be desirable to focus only on a small set of large clones, as the discarded clones typically comprise the vast majority of the sampled repertoire. Additionally, different tools make different basic assumptions, sometimes making comparisons difficult. There is also the question of how stringent one needs to be to declare two clonal assignments "similar." Finally, in some cases, differences observed with different tools are not due to problems with the data or the analysis but rather are due to bonafide differences in what the analytical tools are actually measuring in the data set. For example, comparing diversity at different orders can result in different answers because they give different weights to clones of different sizes. In the end, the best course of action is to pick the metric that best captures the clones of interest in the population.

Replication

As with most experiments, one of the most reliable methods for determining if results are valid is to replicate them by making additional measurements or by performing additional experiments. In the data we present here, we illustrate two forms of replication. The first is that we perform the same bulk sequencing analysis on nine different organ donors. This analysis reveals certain features that are shared in all donors (such as the preponderance of small clones having 1–2 sequences per clone) and other features that vary between different donors (such as the proportion of very large clones, D20, or the distribution of clone sizes). If we had only analyzed two donors (such as D267 and D309), we might have concluded the different individuals have rather similar clone size distributions in the spleen.

Replication is also achieved by making additional measurements on the same sample, as we showed here with 19 replicates from D207 spleen. With antibody gene rearrangement sequencing from gDNA, performing additional amplifications and sequencing on the same DNA aliquot is analogous to sampling additional cells from the same cell population since each cell only has one template molecule. As we show here and in Ref. (9), for clone tracking studies, it is important to power the analysis on both the clone size and the degree of sampling.

CONCLUDING REMARKS

In this paper, we demonstrate the importance of choosing the appropriate combination of experimental approaches and analytical tools to measure B-cell clone size. One has to know what scale of clone sizes is of interest, which means visualizing the repertoire as a whole on a diversity or clone copy number cut-off plot. Additional considerations that guide the choice of clone size metric include the prevalence of SHMs, the possibility of uncontrolled differences in sequencing depth, and the availability of replicate libraries. For clonal overlap analysis, there are approaches that quantify the degree of overlap and others that focus on the similarity of the overlapping clones themselves. Finally, analysis tools and experimental approaches, especially in the single cell realm, are undergoing rapid evolution (19, 20). Members of the AIRR and RepSeq communities, including many of the research teams that have contributed to this special research topic in Frontiers, are contributing to experimental approaches, data analysis and data sharing as methods and providing recommendations (49, 65). We look forward to a future for clonal analysis that is filled with promise and complexity.

ETHICS STATEMENT

Human tissues used in this research were obtained from deceased organ donors through an approved research protocol and material transfer agreement with LiveOnNY, the organ procurement organization for the New York metropolitan area. This type of research has been determined by IRBs of both the University of Pennsylvania and Columbia University as non-human subjects research and, hence, an ethics approval was not required as per institutional and national guidelines.

AUTHOR CONTRIBUTIONS

EP, UH, AR, and WM contributed to the conception and design of the study; DF directs the organ donor tissue resource for acquisition of tissue samples; TG assisted with organ donor sample processing, shipping, and provided input into experimental design; WM generated the flow cytometry and sequencing data and contributed to the data analysis; AR organized the database and performed the clonal rank, rarefaction, and diversity analyses; BZ developed the software to measure resampling; DC generated the clone size cut-off and individual rarefaction plots and designed all of the figures. EP wrote the first draft of the manuscript and oversaw the overall study; UH, AR, WM, and DC wrote sections of the manuscript.

All authors contributed to manuscript revision, read and approved the submitted version.

ACKNOWLEDGMENTS

We thank LiveOnNY and the donor families for making this study possible. We acknowledge the Penn Flow Cytometry Shared Resource and the Human Immunology Core facility. This work was supported by NIH P01 AI106697 and NIH P30 CA016520.

REFERENCES

1. Cohn M, Langman RE. The immune system: a look from a distance. *Front Biosci* (1996) 1:d318–23. doi:10.2741/A134
2. Sablitzky F, Wildner G, Rajewsky K. Somatic mutation and clonal expansion of B cells in an antigen-driven immune response. *EMBO J* (1985) 4:345–50.
3. McKean D, Huppi K, Bell M, Staudt L, Gerhard W, Weigert M. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc Natl Acad Sci U S A* (1984) 81:3180–4. doi:10.1073/pnas.81.10.3180
4. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci U S A* (2014) 111:4928–33. doi:10.1073/pnas.1323862111
5. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat Med* (2016) 22:1456–64. doi:10.1038/nm.4224
6. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL, et al. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc Natl Acad Sci U S A* (2011) 108:21194–9. doi:10.1073/pnas.1118357109
7. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* (2012) 120:5173–80. doi:10.1182/blood-2012-07-444042
8. Robins HS, Ericson NG, Guenthoer J, O'Briant KC, Tewari M, Drescher CW, et al. Digital genomic quantification of tumor-infiltrating lymphocytes. *Sci Transl Med* (2013) 5:214ra169. doi:10.1126/scitranslmed.3007247
9. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, et al. An atlas of B-cell clonal distribution in the human body. *Nat Biotechnol* (2017) 35:879–84. doi:10.1038/nbt.3942
10. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32:158–68. doi:10.1038/nbt.2782
11. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc* (2016) 11:1599–616. doi:10.1038/nprot.2016.093
12. Wang C, Liu Y, Roskin KM, Jackson KJ, Boyd SD. Laboratory and data analysis methods for characterization of human B cell repertoires by high-throughput DNA sequencing. *Methods Mol Biol* (2015) 1343:219–33. doi:10.1007/978-1-4939-2963-4_17
13. von Boehmer L, Liu C, Ackerman S, Gitlin AD, Wang Q, Gazumyan A, et al. Sequencing and cloning of antigen-specific antibodies from mouse memory B cells. *Nat Protoc* (2016) 11:1908–23. doi:10.1038/nprot.2016.102
14. Murugan R, Imkeller K, Busse CE, Wardemann H. Direct high-throughput amplification and sequencing of immunoglobulin genes from single human B cells. *Eur J Immunol* (2015) 45:2698–700. doi:10.1002/eji.201545526
15. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* (2015) 21:86–91. doi:10.1038/nm.3743
16. Khan TA, Friedensohn S, de Vries ARG, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* (2016) 2:e1501371. doi:10.1126/sciadv.1501371
17. Lin SG, Ba Z, Du Z, Zhang Y, Hu J, Alt FW. Highly sensitive and unbiased approach for elucidating antibody repertoires. *Proc Natl Acad Sci U S A* (2016) 113:7846–51. doi:10.1073/pnas.1608649113
18. Vergani S, Korsunsky I, Mazzarello AN, Ferrer G, Chiorazzi N, Bagnara D. Novel method for high-throughput full-length IGHV-D-J sequencing of the immune repertoire from bulk B-cells with single-cell resolution. *Front Immunol* (2017) 8:1157. doi:10.3389/fimmu.2017.01157
19. Friedensohn S, Khan TA, Reddy ST. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol* (2017) 35:203–14. doi:10.1016/j.tibtech.2016.09.010
20. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. *Nat Biotechnol* (2017) 35:908–11. doi:10.1038/nbt.3979
21. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* (2013) 4:2680. doi:10.1038/ncomms3680
22. Bakkus MH, Heirman C, Van Riet I, Van Camp B, Thielemans K. Evidence that multiple myeloma Ig heavy chain VDJ genes contain somatic mutations but show no intraclonal variation. *Blood* (1992) 80:2326–35.
23. Zuckerman NS, McCann KJ, Ottensmeier CH, Barak M, Shahaf G, Edelman H, et al. Ig gene diversification and selection in follicular lymphoma, diffuse large B cell lymphoma and primary central nervous system lymphoma revealed by lineage tree and mutation analyses. *Int Immunol* (2010) 22:875–87. doi:10.1093/intimm/dxq441
24. Friedensohn S, Lindner JM, Cornacchione V, Iazeolla M, Miho E, Zingg A, et al. Synthetic standards combined with error and bias correction improves the accuracy and quantitative resolution of antibody repertoire sequencing in human naive and memory B cells. *bioRxiv* (2018). doi:10.1101/284810
25. Klinman NR, Aschinazi G. The stimulation of splenic foci in vitro. *J Immunol* (1971) 106:1338–44.
26. Bachmann MF, Kundig TM, Kalberer CP, Hengartner H, Zinkernagel RM. How many specific B cells are needed to protect against a virus? *J Immunol* (1994) 152:4235–41.
27. Trumble IM, Allmon AG, Archin NM, Rigdon J, Francis O, Baldoni PL, et al. SLDAssay: a software package and web tool for analyzing limiting dilution assays. *J Immunol Methods* (2017) 450:10–6. doi:10.1016/j.jim.2017.07.004
28. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30:1930–2. doi:10.1093/bioinformatics/btu138
29. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *J Immunol* (2017) 198:2489–99. doi:10.4049/jimmunol.1601850
30. Schwartz GW, Hershberg U. Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. *Front Immunol* (2013) 4:357. doi:10.3389/fimmu.2013.00357
31. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U. ImmuneDB: a system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* (2017) 33:292–3. doi:10.1093/bioinformatics/btw593
32. Dunn-Walters DK, Belelovsky A, Edelman H, Banerjee M, Mehr R. The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees. *Dev Immunol* (2002) 9:233–43. doi:10.1080/10446670310001593541

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/article/10.3389/fimmu.2018.01472/full#supplementary-material.

TABLE S1 | Fasta files of IMGT human JH genes used for germline gene assignment.

TABLE S2 | Fasta files of IMGT human VH genes used for germline gene assignment.

33. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree: creating immunoglobulin variable region gene lineage trees. *J Immunol Methods* (2008) 338:67–74. doi:10.1016/j.jim.2008.06.006
34. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* (2015) 12:380–1. doi:10.1038/nmeth.3364
35. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res* (2013) 2:103. doi:10.12688/f1000research.2-103.v1
36. Kepler TB, Munshaw S, Wiehe K, Zhang R, Yu JS, Woods CW, et al. Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *Front Immunol* (2014) 5:170. doi:10.3389/fimmu.2014.00170
37. Volpe JM, Cowell LG, Kepler TB. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics* (2006) 22:438–44. doi:10.1093/bioinformatics/btk004
38. Ralph DK, Matsen FA. Likelihood-based inference of B cell clonal families. *PLoS Comput Biol* (2016) 12:e1005086. doi:10.1371/journal.pcbi.1005086
39. Schwartz GW, Hershberg U. Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics. *Phys Biol* (2013) 10:035005. doi:10.1088/1478-3975/10/3/035005
40. Kaplinsky J, Arnaout R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* (2016) 7:11881. doi:10.1038/ncomms11881
41. Mamani-Matsuda M, Cosma A, Weller S, Faili A, Staib C, Garcon L, et al. The human spleen is a major reservoir for long-lived vaccinia virus-specific memory B cells. *Blood* (2008) 111:4653–9. doi:10.1182/blood-2007-11-123844
42. Bagnara D, Squillario M, Kipling D, Mora T, Walczak AM, Da Silva L, et al. A reassessment of IgM memory subsets in humans. *J Immunol* (2015) 195:3716–24. doi:10.4049/jimmunol.1500753
43. Sathaliyawala T, Kubota M, Yudanin N, Turner D, Camp P, Thome JJ, et al. Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. *Immunity* (2013) 38:187–97. doi:10.1016/j.immuni.2012.09.020
44. Thome JJ, Yudanin N, Ohmura Y, Kubota M, Grinshpun B, Sathaliyawala T, et al. Spatial map of human T cell compartmentalization and maintenance over decades of life. *Cell* (2014) 159:814–28. doi:10.1016/j.cell.2014.10.026
45. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* (2003) 17:2257–317. doi:10.1038/sj.leu.2403202
46. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* (2015) 7:121. doi:10.1186/s13073-015-0243-2
47. Zhang B, Meng W, Prak ET, Hershberg U. Discrimination of germline V genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment. *J Immunol Methods* (2015) 427:105–16. doi:10.1016/j.jim.2015.10.009
48. Colwell RK, Chao A, Gotellis NJ, Lin S-Y, Mao CX, Chazdon RL, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* (2012) 5:3–21. doi:10.1093/jpe/rtr044
49. Rubelt F, Busse CE, Bukhari SAC, Burckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* (2017) 18:1274–8. doi:10.1038/ni.3873
50. Alt FW, Yancopoulos GD, Blackwell TK, Wood C, Thomas E, Boss M, et al. Ordered rearrangement of immunoglobulin heavy chain variable region segments. *EMBO J* (1984) 3:1209–19.
51. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci U S A* (2013) 110:13463–8. doi:10.1073/pnas.1312146110
52. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A* (2013) 110:19872–7. doi:10.1073/pnas.1319590110
53. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci U S A* (2015) 112:E862–70. doi:10.1073/pnas.1417683112
54. Corcoran MM, Phad GE, Vazquez Bernat N, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
55. Cha RS, Thilly WG. Specificity, efficiency, and fidelity of PCR. *PCR Methods Appl* (1993) 3:S18–29. doi:10.1101/gr.3.3.S18
56. Hill MO. Diversity and evenness: a unifying notation and its consequences. *Ecology* (1973) 54:427–32. doi:10.2307/1934352
57. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* (1948) 27:379–423,623–656. doi:10.1002/j.1538-7305.1948.tb00917.x
58. Jost L. The relation between evenness and diversity. *Diversity* (2010) 2:207–32. doi:10.3390/d2020207
59. Peet RK. The measurements of species diversity. *Ann Rev Ecol Syst* (1974) 5:285–307. doi:10.1146/annurev.es.05.110174.001441
60. Heip CHR, Herman PMJ, Soetaert K. Indices of diversity and evenness. *Oceanis* (1998) 24:61–87.
61. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* (2015) 7:49. doi:10.1186/s13073-015-0169-8
62. Anderson SM, Khalil A, Uduman M, Hershberg U, Louzoun Y, Haberman AM, et al. Taking advantage: high-affinity B cells in the germinal center have lower death rates, but similar rates of division, compared to low-affinity cells. *J Immunol* (2009) 183:7314–25. doi:10.4049/jimmunol.0902452
63. Safonova Y, Lapidus A, Lill J. IgSimulator: a versatile immunosequencing simulator. *Bioinformatics* (2015) 31:3213–5. doi:10.1093/bioinformatics/btv326
64. Lee DW, Khavrutskii IV, Wallqvist A, Bavari S, Cooper CL, Chaudhury S. BRILIA: integrated tool for high-throughput annotation and lineage tree assembly of B-cell repertoires. *Front Immunol* (2017) 7:681. doi:10.3389/fimmu.2016.00681
65. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front Immunol* (2017) 8:1418. doi:10.3389/fimmu.2017.01418

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.

Copyright © 2018 Rosenfeld, Meng, Chen, Zhang, Granot, Farber, Hershberg and Luning Prak. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership