De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of B-cell transcripts

Jiang Zhu[a,1], Xueling Wu[a,2], Baoshan Zhang[a], Krisha McKee[a], Sijy O'Dell[a], Cinque Soto[a], Tongqing Zhou[a], Joseph P. Casazza[a], NISC Comparative Sequencing Program[b,3], James C. Mullikin[b], Peter D. Kwong[a,4], John R. Mascola[a,4], and Lawrence Shapiro[b,c,4]

[a]Vaccine Research Center and [b]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892; and [c]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032

Edited by Rino Rappuoli, Novartis Vaccines and Diagnostics Srl, Siena, Italy, and approved August 30, 2013 (received for review April 3, 2013)

Next-generation sequencing of antibody transcripts provides a wealth of data, but the ability to identify function-specific antibodies solely on the basis of sequence has remained elusive. We previously characterized the VRC01 class of antibodies, which target the CD4-binding site on gp120, appear in multiple donors, and broadly neutralize HIV-1. Antibodies of this class have developmental commonalities, but typically share only ∼50% amino acid sequence identity among different donors. Here we apply next-generation sequencing to identify VRC01 class antibodies in a new donor, C38, directly from B cell transcript sequences. We first tested a lineage rank approach, but this was unsuccessful, likely because VRC01 class antibody sequences were not highly prevalent in this donor. We next identified VRC01 class heavy chains through a phylogenetic analysis that included thousands of sequences from C38 and a few known VRC01 class sequences from other donors. This "cross-donor analysis" yielded heavy chains with little sequence homology to previously identified VRC01 class heavy chains. Nonetheless, when reconstituted with the light chain from VRC01, half of the heavy chain chimeric antibodies showed substantial neutralization potency and breadth. We then identified VRC01 class light chains through a five-amino-acid sequence motif necessary for VRC01 light chain recognition. From over a million light chain sequences, we identified 13 candidate VRC01 class members. Pairing of these light chains with the phylogenetically identified C38 heavy chains yielded functional antibodies that effectively neutralized HIV-1. Bioinformatics analysis can thus directly identify functional HIV-1–neutralizing antibodies of the VRC01 class from a sequenced antibody repertoire.

antibodyomics | cross-donor phylogenetic analysis | DNA sequencing | humoral immune response | sequence signature

Significance

An extraordinary influx of sequencing information is revolutionizing biological inquiry. While sequences of entire antibody repertoires are straightforward to obtain, understanding antibody function on the basis of sequence alone has remained elusive. Can bioinformatics identify function-specific antibodies within the ocean of B cell transcripts representing unrelated specificities? We undertook the challenge of identifying antibodies of the VRC01 class. These antibodies individually neutralize up to 90% of HIV-1; although they share less than 50% sequence identity they do have characteristic sequence motifs and evolutionary relatedness. Our bioinformatics methods identified heavy and light chains from a new donor that could form functional antibodies and neutralize HIV-1 effectively. Identification of HIV-1 neutralizing antibodies of the VRC01 class can thus occur solely on the basis of bioinformatics analysis of a sequenced antibody repertoire.

The heavy and light chain sequences of an antibody determine its antigen-specific recognition (1–3), and a long-standing problem in structural bioinformatics has been to predict the recognition of an antibody based solely on its sequence. This problem of sequence-based recognition can be separated into two structural components (1): determining recognition from structure and (2) determining structure from sequence. Both of these components remain active areas of inquiry, with the latter representing the famous "protein-folding problem" (4, 5). For antibodies, the overall structure of immunoglobulins is known, and recognition is generally determined by six loops, the complementarity-determining regions (CDRs). Despite this reduced complexity, antibodies display diversity $>10^{12}$ in each individual and distinguish epitopes with high precision. Thus, although the general problem of predicting recognition from sequence remains intractable, a number of strategies are now being developed to determine recognition from antibody sequence.

First, population-based strategies: if a particular antibody sequence is highly prevalent, biological considerations can suggest a particular function. For example, Reddy et al. (6) used the population-specific metric of frequency to identify prevalent lineages of antigen-specific antibodies from bone marrow plasma cells of immunized mice (6). Second, sequence signature-based strategies: sequence characteristics can clearly be used to delineate antibodies with similar recognition when the identity is high (e.g., >90%). Moreover, structurally defined sequence signatures can be effective for identifying select elements within more divergent sequences (e.g., as low as 30% identity) that specify related recognition. Third, evolution-based strategies: evolutionary similarity often reveals functional relationships between proteins. In the particular case of antibodies, the overall function is recognition, and evolutionary similarity can reveal details of this recognition, as demonstrated for the VRC01 class

Author contributions: J.Z., X.W., K.M., S.O., T.Z., J.C.M., P.D.K., J.R.M., and L.S. designed research; J.Z., X.W., B.Z., K.M., S.O., C.S., T.Z., J.P.C., and N.I.S.C.C.S.P. performed research; X.W. contributed new reagents/analytic tools.; J.Z., X.W., B.Z., K.M., S.O., C.S., T.Z., J.P.C., N.I.S.C.C.S.P., J.C.M., P.D.K., J.R.M., and L.S. analyzed data; and J.Z., X.W., P.D.K., J.R.M., and L.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Raw Roche 454 pyrosequencing data sets used in this study were deposited in the NCBI Sequence Read Archive (SRA) (accession no. SRP026397). Verified neutralizing heavy and light chain variable region sequences have been deposited in the GenBank database (accession nos. KF306044–KF306069).

[1] Present address: Department of Immunology and Microbial Science, The Scripps Research Institute, La Jolla, CA 92037.

[2] Present address: Aaron Diamond AIDS Research Center, New York, NY 10016.

[3] A complete list of the National Institutes of Health Intramural Sequencing Center (NISC) Comparative Sequencing Program can be found in the *Supporting Information*.

[4] To whom correspondence may be addressed. E-mail: pdkwong@nih.gov, jmascola@nih.gov, or lss8@columbia.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1306262110/-/DCSupplemental.

antibodies (7–12). Named after the antibody VRC01, this class includes some of the most effective HIV-1 neutralizers and has been identified in multiple HIV-1–infected donors. Antibodies of the VRC01 class share a number of features including a common gp120 binding mode that incorporates heavy chain mimicry of the CD4 receptor, heavy chain origin from the IGHV1-2 germ-line gene, and a light chain characterized by a CDR L3 region of five amino acids in length. Despite these similarities, sequence differences between VRC01 class heavy chains often exceed 50%; nonetheless, Wu et al. (7) used evolutionary similarity to identify VRC01 class antibodies in donors already known to have VRC01 class antibodies. However, it was unclear whether these evolutionary techniques could be applied to new donors, in which template VRC01 class sequences were unknown.

Advances in next-generation sequencing (13–15) allow for the routine determination of millions of antibody heavy and light chain sequences from a sample of donor B cells, providing the potential for a much more detailed analysis of the expressed antibody repertoire—the antibodyome (6, 16–18). Here we ask whether the information acquired from next-generation sequencing can be combined with recognition/sequence strategies of population metrics, sequence signatures, and evolution to identify VRC01 class antibodies in an individual, donor C38, with no previously characterized antibodies. The results indicate that, in cases such as with antibodies of the same class that share donor-independent characteristics, knowledge-based sieving of B-cell transcripts determined by next-generation sequencing can identify antibodies with specific recognition from sequence alone.

Results

Serum Analysis of Donor C38 Sample. VRC01 class antibodies precisely target the CD4-binding site of the gp120 envelope glycoprotein to achieve broad neutralization of HIV-1 (8, 19). To date, heavy chains of all VRC01 class antibodies identified originate from IGHV1-2 family genes. We selected an HIV-1–infected donor (C38), whose sera exhibited broad and potent neutralization of HIV-1 (*SI Appendix*, Table S1). Serum neutralization of a panel of eight HIV-1 isolates was assessed in the presence of a resurfaced HIV-1 gp120 core protein (RSC3), a probe that binds VRC01 (19), and a negative control (ΔRSC3) that contains a single amino acid deletion in the VRC01 binding site. In the presence of RSC3 but not ΔRSC3, sera from donor C38 showed a mean 22% reduction in neutralization. Although this reduction was less than seen in donors 45, 74, and 0219, from which we had previously identified VRC01 class antibodies (7) (Fig. 1*A*), the data nonetheless suggested the presence of VRC01 class or other types of CD4-binding site-directed antibodies in this donor.

454 Pyrosequencing and Sequence-Based Identification of VRC01-Class Antibodies. We performed next-generation sequencing of donor C38 B-cell transcripts using PCR to amplify IgG and IgM heavy chain sequences from the VH1 family. mRNA from an estimated 5 million peripheral blood mononuclear cells (PBMCs) was used for reverse transcription to produce template cDNA. We initially used primers that overlapped the end of the V-gene leader sequence and the start of the V region (H1 primers) (7, 20) and then switched to more upstream primers that annealed to the start of the V-gene leader sequence (G1 primers; *SI Appendix*, Table S2) (10); by avoiding the variable region, the G1 primers were better able to capture sequences with increased somatic mutation (10).

Roche 454 pyrosequencing provided 460,706 (H1 primers) and 574,027 (G1 primers) heavy chain reads for donor C38. After initial processing using our previously described bioinformatics pipeline (7, 21) (*SI Appendix*, Table S3), 138,523 sequences in the H1 primer data set and 168,365 sequences in the G1 primer data set were assigned to the IGHV1-2*02 allelic origin, the heavy chain germ-line gene for VRC01 class antibodies

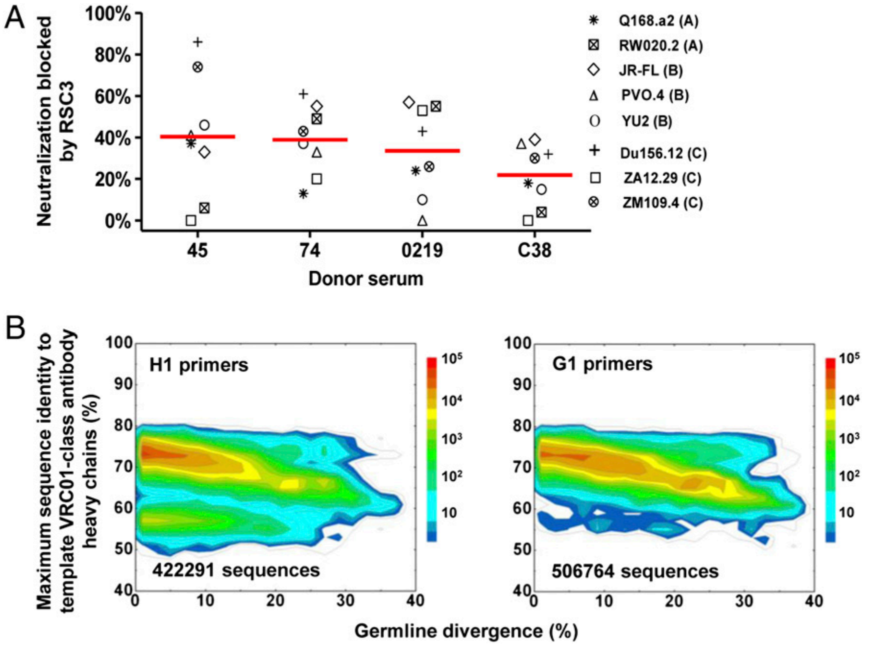


Fig. 1. Donor C38 serum analysis and identification of HIV-1–neutralizing antibodies based on sequence identity. (*A*) Reduction in neutralization ID_{50} resulting from RSC3 vs. ΔRSC3 competition against eight HIV-1 strains for a 2008 serum sample of donor C38. Sera from donors 45, 74, and 0219 are included in this analysis for comparison. Red bars indicate the mean reduction. (*B*) Divergence/identity analysis of donor C38 heavy chain sequences generated from 454 pyrosequencing with H1 (*Left*) and G1 (*Right*) primers. Heavy chain sequences are plotted as a function of maximal sequence identity to the heavy chains of 13 template VRC01 class antibodies (VRC01, VRC02, VRC03, NIH45-46, VRC-PG04, VRC-PG04b, VRC-CH30, VRC-CH31, VRC-CH32, 3BNC60, 3BNC117, 12A12, and 12A21) and of sequence divergence from inferred germ-line alleles.

(7, 19). These sequences were then compared with a set of known template VRC01 class antibodies isolated from other donors, which included eight from our previous studies (7, 19) (VRC01-03, VRC-PG04/04b, and VRC-CH30-32) and five identified by Nussenzweig and coworkers (10) (12A12, 12A21, 3BNC60, 3BNC117, and NIH45-46) (*SI Appendix*, Fig. S1). Because all antibody variable domains, heavy and light chains alike, share a similar scaffold with conserved sequence motifs in the framework regions that are responsible for the heavy/light chain complexation, the sequence identity of irrelevant antibody chains is usually higher than that of functionally distinct protein domains. Therefore, an identity cutoff of 80%, instead of 30–40% as is generally used for unrelated proteins, was used here to detect sequence homology between antibody chains. No sequence from either the H1 or G1 primer data set was found to be >80% identical to any of the heavy chains from template antibodies from other donors (Fig. 1*B*). These results suggested that VRC01 class antibodies, if they did exist in the 454 pyrosequencing–derived repertoire of donor C38, could not be simply recognized by sequence homology to a known antibody.

Sequence Prevalence-Based Identification of VRC01 Class Heavy Chains. Given the low homology of donor sequences to the known VRC01 class antibodies, we sought to use a prevalence-based method to interrogate the donor C38 repertoire. One implementation of antibody prevalence analysis, the lineage rank method (6), is based on the supposition that the desired antibodies form highly prevalent lineages in the total antibody repertoire (Fig. 2*A*). Because the heavy chains of all known VRC01 class antibodies can complement the VRC01 light chain to form neutralizing antibodies, we tested the lineage rank approach on heavy chain sequences. The heavy chain complementary-determining region 3 (CDR H3), which encompasses the site of V(D)J recombination—V-D junction, D segment, D-J junction, and part of J gene preceding the conserved WGXG motif, provides a sequence signature for antibody lineage definition. This definition of an antibody clonal lineage may be more robust than analyses based on the inferred germ-line use and junctional sequences, due to the relatively low accuracy of D gene assignment, as well as other complicating factors such as repertoire diversity and sequencing errors. Both CDR H3 similarity and VH gene characteristics were considered in our lineage definition.

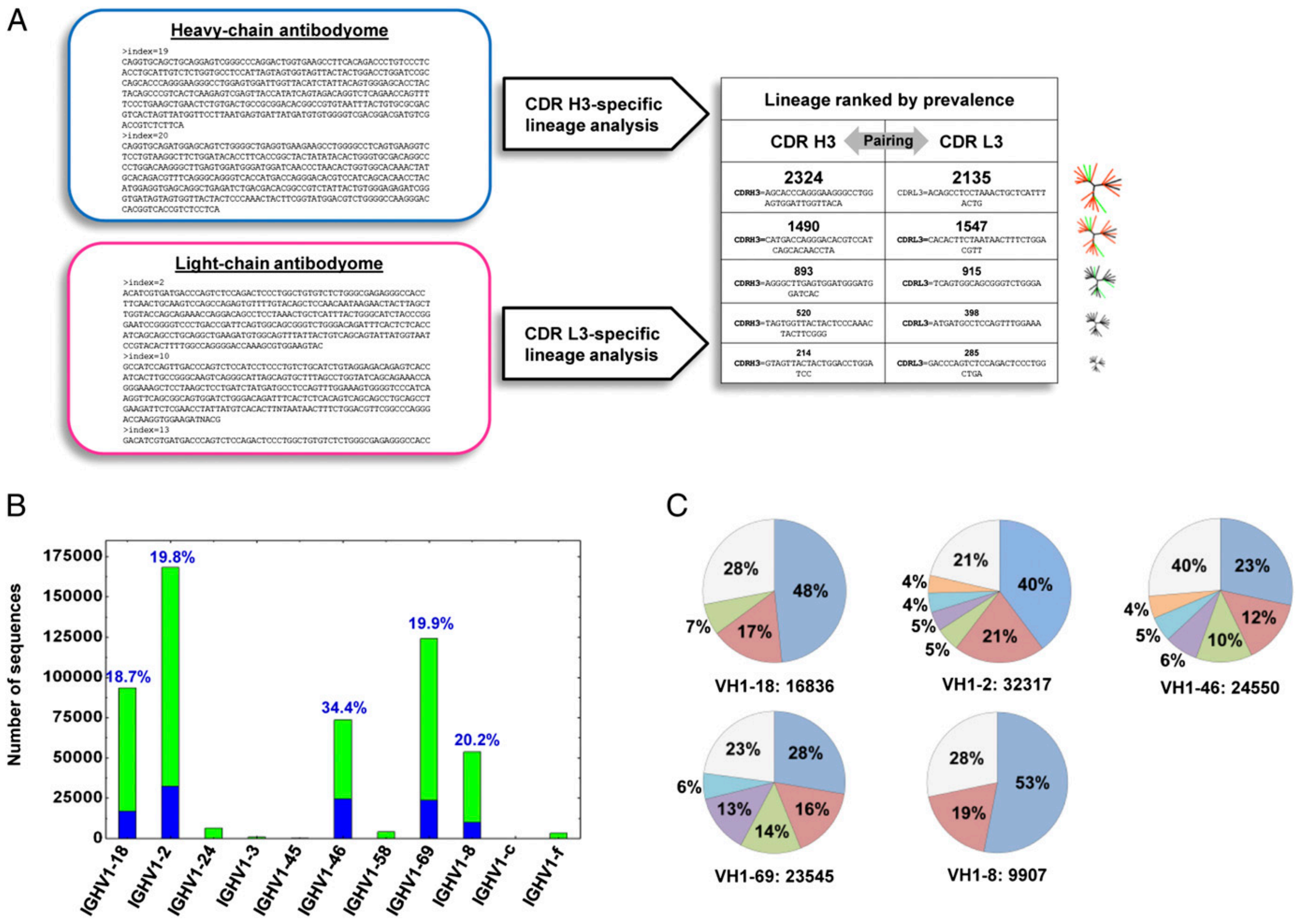

A
Heavy-chain antibodyome
>index=19
CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCACAGACCCTGTCCCTC
ACCTGCATTGTCTCTGGTGCCTCCATTAGTAGTGGTAGTTACTACTGGACCTGGATCCGC
CAGCACCCAGGGAAGGGCCTGGAGTGGATTGGTTACATCTATTACAGTGGGAGCACCTAC
TACAGCCCGTCACTCAAGAGTCGAGTTACCATATCAGTAGACAGGTCTCAGAACCAGTTT
TCCCTGAAGCTGAACTCTGTGACTGCCGCGGACACGGCCGTGTAATTTACTGTGCGCGAC
GTCACTAGTTATGGTTCCTTAATGAGTGATTATGATGTGTGGGGTCGACGGACGATGTCG
ACCGTCTCTTCA
>index=20
CAGGTGCAGATGGAGCAGTCTGGGGCTGAGGTGAAGAAGCCTGGGGCCTCAGTGAAGGTC
TCCTGTAAGGCTTCTGGATACACCTTCACCGGCTACTATATACACTGGGTGCGACAGGCC
CCTGGACAAGGGCTTGAGTGGATGGGATGGATCAACCCTAACACTGGTGGCACAAACTAT
GCACAGACGTTTCAGGGCAGGGTCACCATGACCAGGGACACGTCCATCAGCACAACCTAC
ATGGAGGTGAGCAGGCTGAGATCTGACGACACGGCCGTCTATTACTGTGGGAGAGATCGG
GTGATAGTAGTGGTTACTACTCCCAAACTACTTCGGTATGGACGTCTGGGGCCAAGGGAC
CACGGTCACCGTCTCCTCA
Light-chain antibodyome
>index=2
ACATCGTGATGACCCAGTCTCCAGACTCCCTGGCTGTGTCTCTGGGCGAGAGGGCCACC
TTCAACTGCAAGTCCAGCCAGAGTGTTTTGTACAGCTCCAACAATAAGAACTACTTAGCT
TGGTACCAGCAGAAACCAGGACAGCCTCCTAAACTGCTCATTTACTGGGCATCTACCCGG
GAATCCGGGGTCCCTGACCGATTCAGTGGCAGCGGGTCTGGGACAGATTTCACTCTCACC
ATCAGCAGCCTGCAGGCTGAAGATGTGGCAGTTTATTACTGTCAGCAGTATTATGGTAAT
CCGTACACTTTTGGCCAGGGGACCAAAGCGTGGAAGTAC
>index=10
GCCATCCAGTTGACCCAGTCTCCATCCTCCCTGTCTGCATCTGTAGGAGACAGAGTCACC
ATCACTTGCCGGGCAAGTCAGGGCATTAGCAGTGCTTTAGCCTGGTATCAGCAGAAACCA
GGGAAAGCTCCTAAGCTCCTGATCTATGATGCCTCCAGTTTGGAAAGTGGGGTCCCATCA
AGGTTCAGCGGCAGTGGATCTGGGACAGATTTCACTCTCACAGTCAGCAGCCTGCAGCCT
GAAGATTCTCGAACCTATTATGTCACACTTNTAATAACTTTCTGGACGTTCGGCCCAGGG
ACCAAGGTGGAAGATNACG
>index=13
GACATCGTGATGACCCAGTCTCCAGACTCCCTGGCTGTGTCTCTGGGCGAGAGGGCCACC
CDR H3-specific lineage analysis
CDR L3-specific lineage analysis
Lineage ranked by prevalence
CDR H3
Pairing
CDR L3
2324
CDRH3=AGCACCCAGGGAAGGGCCTGGAGTGGATTGGTTACA
2135
CDRL3=ACAGCCTCCTAAACTGCTCATTTACTG
1490
CDRH3=CATGACCAGGGACACGTCCATCAGCACAACCTA
1547
CDRL3=CACACTTCTAATAACTTTCTGGACGTT
893
CDRH3=AGGGCTTGAGTGGATGGGATGGATCAC
915
CDRL3=TCAGTGGCAGCGGGTCTGGGA
520
CDRH3=TAGTGGTTACTACTCCCAAACTACTTCGGG
398
CDRL3=ATGATGCCTCCAGTTTGGAAA
214
CDRH3=GTAGTTACTACTGGACCTGGATCC
285
CDRL3=GACCCAGTCTCCAGACTCCCTGGCTGA
B
Number of sequences
175000
150000
125000
100000
75000
50000
25000
0
18.7%
19.8%
34.4%
19.9%
20.2%
IGHV1-18
IGHV1-2
IGHV1-24
IGHV1-3
IGHV1-45
IGHV1-46
IGHV1-58
IGHV1-69
IGHV1-8
IGHV1-c
IGHV1-f
C
28%
48%
7%
17%
VH1-18: 16836
21%
40%
4%
4%
5%
5%
21%
VH1-2: 32317
40%
23%
4%
12%
5%
10%
6%
VH1-46: 24550
23%
28%
6%
13%
16%
14%
VH1-69: 23545
28%
53%
19%
VH1-8: 9907

Fig. 2. Identification of functional VRC01 class antibody heavy chains by frequency-based method, lineage rank. (*A*) Flowchart of lineage rank method. Starting from the heavy and light chain antibodyomes, after a primary analysis to calculate parameters such as germ-line divergence, a CDR H3 or L3-specific lineage analysis is carried out to identify all heavy and light chain lineages, which are then ranked by their prevalence (the number of sequences in a lineage), and representative heavy and light chains from the top-ranking lineages are paired for experimental assessment, e.g., neutralization. Because the VRC01 light chain can complement all VRC01 class antibody heavy chains, here lineage rank is only tested with the C38 heavy chains. (*B*) Sequence distribution of IGHV1 germ-line genes for G1 primer–derived data set. Blue bars (and percentage values above each bar) indicate the sequences with 20% or greater divergence from inferred germ line, which are considered heavy chains of mature antibodies and subjected to the lineage rank analysis. (*C*) Prevalent CDR H3 lineages identified within the five major IGHV1 germ-line gene families: IGHV1-18, IGHV1-2, IGHV1-46, IGHV1-69, and IGHV1-8. None of the 35 heavy chain sequences selected from these 22 lineages shows HIV-1 neutralization when paired with VRC01 light chain.

Within a germ-line VH gene family, sequences having a divergence of 20% or greater were first clustered into groups such that sequences within each group have no more than five nucleotide differences in the CDR H3 region. These heavy chain sequence groups are henceforth referred to as CDR H3 groups; lineages were then constructed by merging similar CDR H3 groups based on three criteria: (*i*) their CDR H3 sequences were of the same length; (*ii*) they shared >80% amino acid sequence identity; and (*iii*) divergence in V gene did not increase significantly on group merging (to prevent merging two different heavy chain lineages with coincidentally similar CDR H3s). Prevalent lineages (those with >1,000 sequences in the current case) could then be subjected to experimental validation. Lineage rank was tested on both H1 and G1 primer-derived data acquired from donor C38 and H1 primer-derived data from donor 74, which was acquired in our previous study (7) and served here as a benchmark. The intermediate output of lineage rank analyses, such as CDR H3 groups and lineages constructed from these groups, is detailed for donor 74 and donor C38 in the *SI Appendix*, Figs. S2 and S3 and Tables S4–S9.

Analysis of IGHV1-2 family heavy chain sequences derived from donor 74, from whom we previously isolated mAb VRC-PG04, revealed four CDR H3 groups (*SI Appendix*, Table S4). The two most prevalent of these (*SI Appendix*, Table S5) were found to correspond to the broadly neutralizing VRC-PG04–like antibodies designated classes 7 and 8 in our earlier paper (7) (*SI Appendix*, Fig. S2). Analysis of the donor C38 IGHV1-2 sequences derived from H1 primer amplification revealed a lower than expected population (<6%) of VH sequences that were more than 20% diverged from their germ-line ancestor, suggesting that H1 primers were suboptimal for amplification of matured antibodies in this donor (*SI Appendix*, Fig. S3 and Tables S6 and S7), and thus the lineage rank analysis focused on the mature sequences from VH1 gene families that were amplified using the G1 primers (Fig. 2*B*). For each identified lineage, the predicted sequence at the center of the largest sequence cluster was selected as a representative heavy chain, which was synthesized and cloned into an expression vector; the full antibody was reconstituted with the VRC01 light chain and tested for neutralization. As stated above, we chose the VRC01 light chain because prior data demonstrated that it complemented other VRC01 class heavy chain sequences (7). If total V gene variation within the lineage was 7% or greater, a second representative sequence was also tested. A total of 35 heavy chain sequences were selected from 22 lineages (Fig. 2*C*; *SI Appendix*, Tables S8 and S9), comprising >15% of the whole data set and 73% of the heavy chain population with a divergence of >20%. Thirty-three of these VH region sequences, when reconstituted as full heavy chains and paired with a VRC01 light chain, could be expressed as soluble IgGs (*SI Appendix*, Table S10), but none of these showed neutralization against HIV-1. Thus, analysis of the most prevalent lineages in donor C38 antibodyome did not identify VRC01 class neutralizing antibody sequences, suggesting that either this donor does not contain VRC01 class antibodies or the prevalence assumption of the lineage rank method did not apply for this donor.

Phylogeny-Based Identification of VRC01 Class Antibody Heavy Chains. We next tested an evolution-based method for identification of VRC01 class antibody heavy chains termed “cross-donor phylogenetic analysis.” We previously demonstrated that VRC01 class antibody heavy chains from different donors evolve in a similar way to achieve precise recognition of the CD4-binding site and developed a sieving method—cross-donor phylogenetic analysis—based on phylogenetic similarity to capture heavy chain sequences with similar maturation pattern (7). In this analysis, a phylogenetic tree was rooted by the IGHV1-2*02 germ line, and donor VRC01 class heavy chain sequences segregated with exogenously added, known VRC01 class antibody heavy chains from other donors. Our prior analysis of donor 74 identified 5,047 heavy chain sequences in the VRC01-like subtree. (Note that VRC01-like refers to sequences that possess the key characteristics of VRC01 class enumerated above but have not been experimentally confirmed to have the same function.) Twenty-four of these were members of a set of sequences chosen by other means (identity/divergence analysis) that were reconstituted and shown in our prior paper to neutralize HIV-1 (7). Neutralization by these sequences suggests that the phylogeny-based method can be used to directly identify VRC01 class antibody heavy chains from a donor antibodyome. However, it is worth noting that the cross-donor analysis of donor 74 was performed with prior knowledge of VRC01 class antibodies isolated from the same donor, as opposed to the donor C38, for whom such information is not available.

To test this possibility, we performed cross-donor phylogenetic analysis on the donor C38 heavy chain sequences derived from H1 primer amplification (131,108 total sequences) and identified 11 sequences in the VRC01-like subtree. Two of these were expressed with the VRC01 light chain, and one of them (gVRC-H1$_{dC38}$) showed broad neutralization of HIV-1 (Fig. 3*A*). Using the G1 primers (163,108 sequences), 93 nonredundant heavy chain sequences were identified as segregating in the VRC01-like subtree, including the sequence designated gVRC-H1$_{dC38}$. Among these, 10 representative sequences were manually selected from different branches of the VRC01-like subtree for expression with the VRC01 light chain. Nine of the reconstituted antibodies displayed HIV-1 neutralization, yielding a 90% success rate in identification (Fig. 3*B*). In total, 10 functional VRC01 class antibody heavy chains were identified from donor C38 (Fig. 4*A*; *SI Appendix*, Table S11). Identity/divergence analysis (Fig. 4*B*) showed that none of these 10 heavy chains was >75% identical to any of the 13 template VRC01 class antibody heavy chains. Furthermore, a significant number of unrelated sequences with higher sequence homology to the template sequences were found in both H1 and G1 primer-derived antibodyomes. A detailed gene family and junction analysis of 10 heavy chains (*SI Appendix*, Figs. S4 and S5) revealed four different CDR H3 groups, suggesting that they might belong to different lineages or alternatively that they evolved from the same ancestor but diverged significantly after a long maturation process. The antibody gVRC-H1$_{dC38}$/VRC01L was tested on a panel of 153 HIV-1 strains and neutralized >80% of these at concentrations of <50 μg/mL (*SI Appendix*, Fig. S6 and Table S12).

Identification of VRC01 Class Antibody Light Chains from Donor C38. An antibody consists of two paired heavy and light chains. Identification of light chain partners for the 10 neutralizing heavy chains allows for the reconstitution of functional VRC01 class antibodies from donor C38. Because the germ-line origin of C38 light chains was unclear, we designed primers to amplify both λ and κ germ-line V genes (*SI Appendix*, Table S2). With the first 454 pyrosequencing experiment, 211,830 full-length light chain sequences, 54,079 λ chains and 157,751 κ chains, were obtained from 257,910 raw reads (*SI Appendix*, Table S3). Recently, West et al. (12) analyzed structural and sequence data for known VRC01 class antibodies and found that these antibodies possess a CDR L3 loop of five amino acids and a glutamine (Q) or glutamate (E) at position 96 (Kabat numbering). Separately, we investigated the B-cell ontogeny of VRC01 class antibodies from multiple donors with next-generation sequencing and arrived at a similar conclusion to the definition of CDR L1 signature. Based on these studies, we adopted a sequence-specific motif—a CDR L3 length of five amino acids and Q or E at position 96—as a simple signature, along with a requirement of maturation from the germ line greater than 10%, to identify VRC01-like

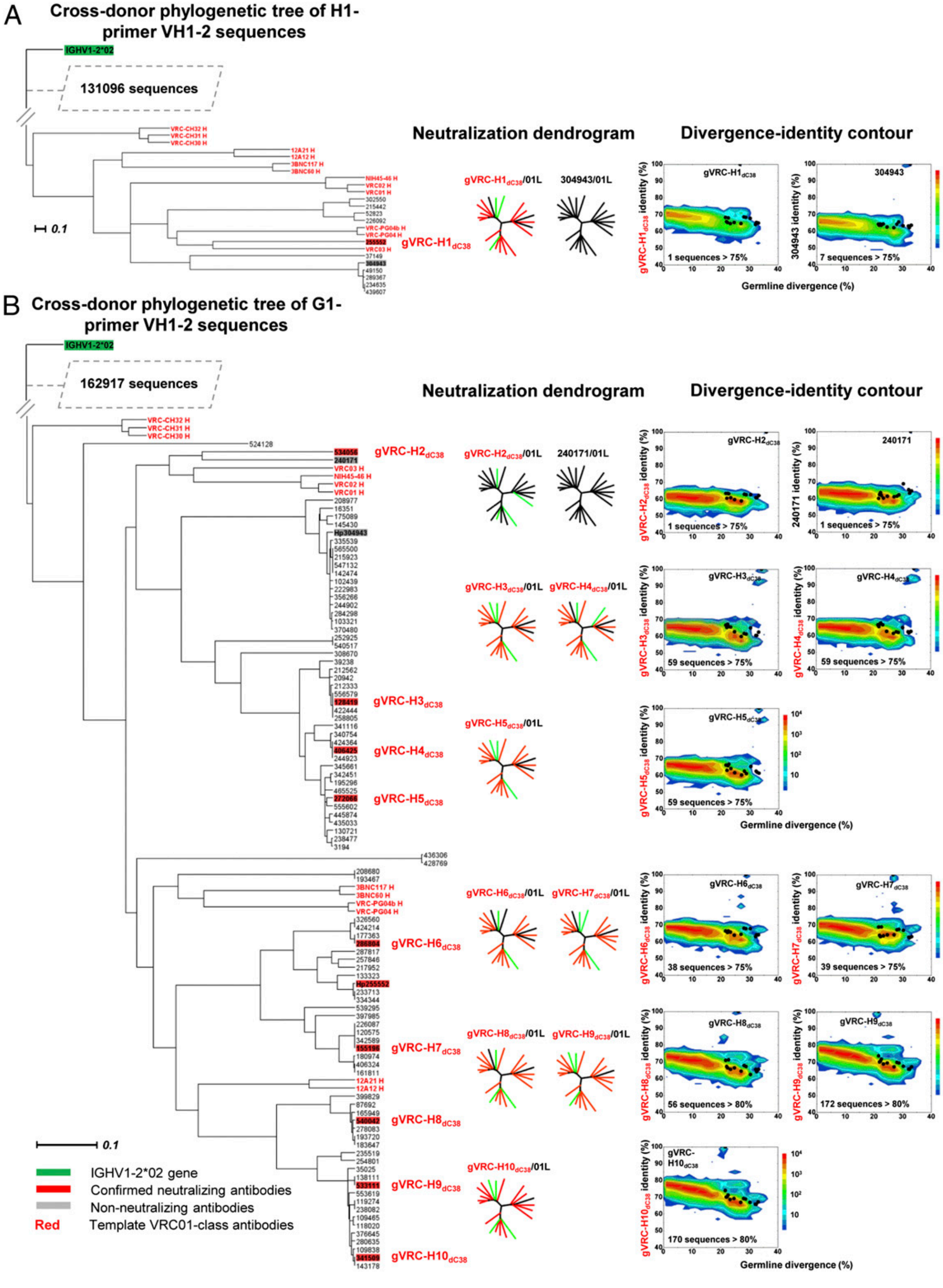
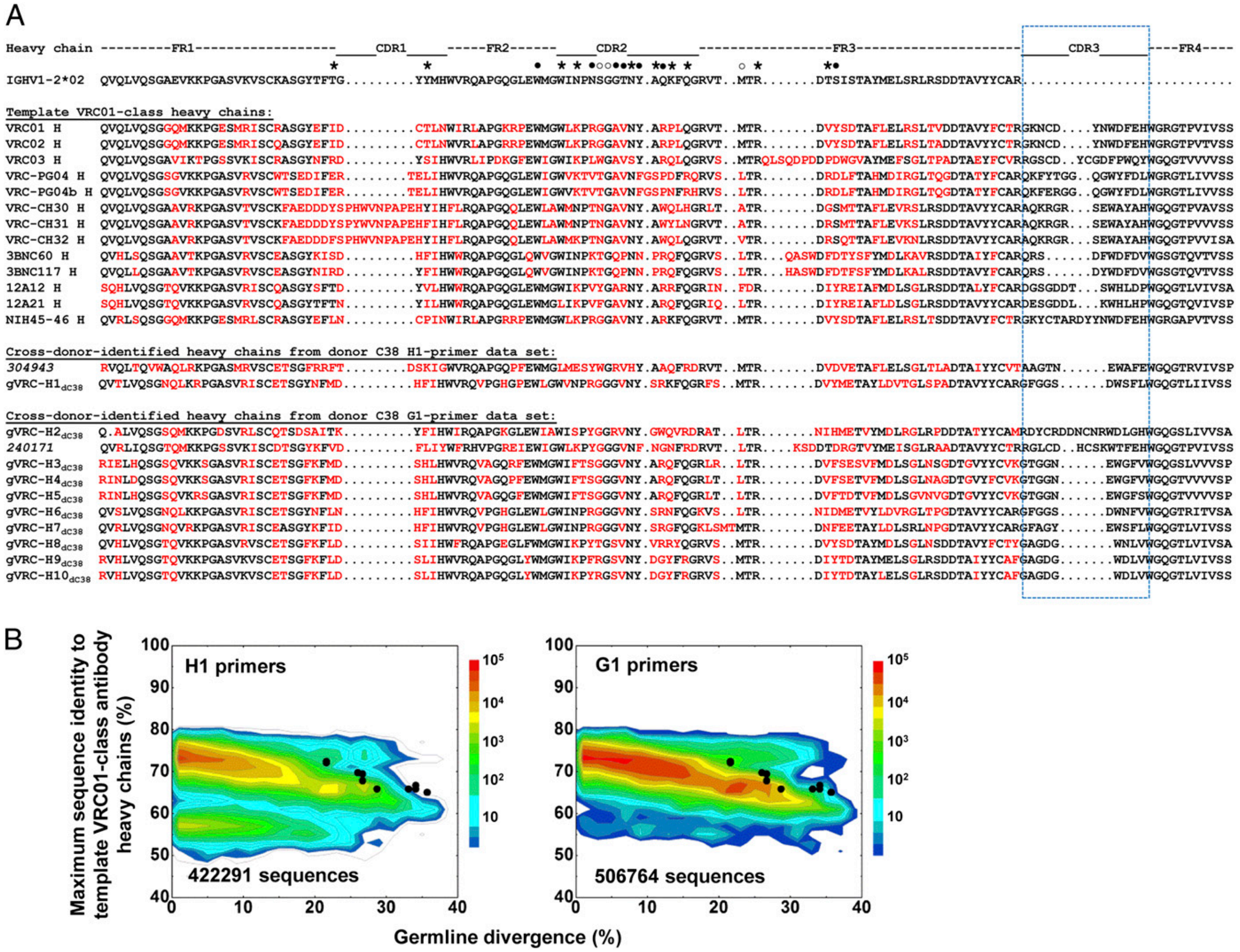


Fig. 3. Identification of VRC01 class antibody heavy chains from donor C38 antibodyome by cross-donor phylogenetic analysis. Cross-donor phylogenetic trees of C38 heavy chain sequences generated by H1 (*A*) and G1 (*B*) primer sets. For both data sets, maximum-likelihood trees of variable domain sequences of the IGHV1-2*02 origin from donor C38, along with 13 known VRC01 class antibody heavy chain sequences from five other donors, are rooted by the germ-line gene sequence. VRC01-like donor sequences segregate with these known VRC01 class antibodies. Bars representing 0.1 changes per nucleotide site are shown. Two and 10 heavy chain sequences were selected from H1 and G1 primer data sets, respectively, for functional assessment and identity/divergence analysis. The neutralization profiles of selected heavy chain sequences reconstituted with the VRC01 light chain [named gVRC-H(n)$_{dC38}$] are depicted with 20 isolate neutralization dendrograms. Explicit neutralization IC$_{50}$s are provided in *SI Appendix*, Table S10. In the rightmost panels, the repertoire of heavy chain sequences is plotted as a function of sequence identity to a selected heavy chain sequence and of sequence divergence from inferred germ-line alleles. The number of sequences >75% (or 80%) identical to the selected sequence, which are likely the somatic variants of the reference heavy chain sequence, is provided for each plot. Notably, the identified heavy chains from donor C38 do not display significant sequence similarity to previously identified VRC01 class antibodies. Color coding indicates the number of sequences.



```
A
Heavy chain ---------FR1-----------------____CDR1____-----FR2------____CDR2_________----------------FR3--------------------______CDR3______----FR4----
                                      *            *         •  * * •○○••*•  *•* *     ○ *        *•
IGHV1-2*02   QVQLVQSGAEVKKPGASVKVSCKASGYTFTG.........YYMHWVRQAPGQGLEWMGWINPNSGGTNY.AQKFQGRVT..MTR.......DTSISTAYMELSRLRSDDTAVYYCAR..........................

Template VRC01-class heavy chains:
VRC01 H      QVQLVQSGGQMKKPGESMRISCRASGYEFID.........CTLNWIRLAPGKRPEWMGWLKPRGGAVNY.ARPLQGRVT..MTR.......DVYSDTAFLELRSLTVDDTAVYFCTRGKNCD....YNWDFEHWGRGTPVIVSS
VRC02 H      QVQLVQSGGQMKKPGESMRISCQASGYEFID.........CTLNWVRLAPGRRPEWMGWLKPRGGAVNY.ARPLQGRVT..MTR.......DVYSDTAFLELRSLTADDTAVYYCTRGKNCD....YNWDFEHWGRGTPVTVSS
VRC03 H      QVQLVQSGAVIKTPGSSVKISCRASGYNFRD.........YSIHWVRLIPDKGFEWIGWIKPLWGAVSY.ARQLQGRVS..MTRQLSQDPDDPDWGVAYMEFSGLTPADTAEYFCVRRGSCD..YCGDFPWQYWGQGTVVVVSS
VRC-PG04 H   QVQLVQSGSGVKKPGASVRVSCWTSEDIFER........TELIHWVRQAPGQGLEWIGWVKTVTGAVNFGSPDFRQRVS..LTR.......DRDLFTAHMDIRGLTQGDTATYFCARQKFYTGG..QGWYFDLWGRGTLIVVSS
VRC-PG04b H  QVQLVQSGSGVKKPGASVRVSCWTSEDIFER........TELIHWVRQAPGQGLEWIGWVKTVTGAVNFGSPNFRHRVS..LTR.......DRDLFTAHMDIRGLTQGDTATYFCARQKFERGG..QGWYFDLWGRGTLIVVSS
VRC-CH30 H   QVQLVQSGAAVRKPGASVTVSCKFAEDDDYSPHWVNPAPEHYIHFLRQAPGQQLEWLAWMNPTNGAVNY.AWQLHGRLT..ATR.......DGSMTTAFLEVRSLRSDDTAVYYCARAQKRGR...SEWAYAHWGQGTPVAVSS
VRC-CH31 H   QVQLVQSGAAVRKPGASVTVSCKFAEDDDYSPYWVNPAPEHFIHFLRQAPGQQLEWLAWMNPTNGAVNY.AWYLNGRVT..ATR.......DRSMTTAFLEVKSLRSDDTAVYYCARAQKRGR...SEWAYAHWGQGTPVVVSS
VRC-CH32 H   QVQLVQSGAAVRKPGASVTVSCKFAEDDDFSPHWVNPAPEHYIHFLRQAPGQQLEWLAWMKPTNGAVNY.AWQLQGRVT..VTR.......DRSQTTAFLEVKNLRSDDTAVYYCARAQKRGR...SEWAYAHWGQGTPVVISA
3BNC60 H     QVHLSQSGAAVTKPGASVRVSCEASGYKISD.........HFIHWWRQAPGQGLQWVGWINPKTGQPNN.PRQFQGRVS..LTR...QASWDFDTYSFYMDLKAVRSDDTAIYFCARQRS.....DFWDFDVWGSGTQVTVSS
3BNC117 H    QVQLLQSGAAVTKPGASVRVSCEASGYNIRD.........YFIHWWRQAPGQGLQWVGWINPKTGQPNN.PRQFQGRVS..LTR...HASWDFDTFSFYMDLKALRSDDTAVYFCARQRS.....DYWDFDVWGSGTQVTVSS
12A12 H      SQHLVQSGTQVKKPGASVRISCQASGYSFTD.........YVLHWWRQAPGQGLEWMGWIKPVYGARNY.ARRFQGRIN..FDR.......DIYREIAFMDLSGLRSDDTALYFCARDGSGDDT...SWHLDPWGQGTLVIVSA
12A21 H      SQHLVQSGTQVKKPGASVRVSCQASGYTFTN.........YILHWWRQAPGQGLEWMGLIKPVFGAVNY.ARQFQGRIQ..LTR.......DIYREIAFLDLSGLRSDDTAVYYCARDESGDDL...KWHLHPWGQGTQVIVSP
NIH45-46 H   QVRLSQSGGQMKKPGESMRLSCRASGYEFLN.........CPINWIRLAPGRRPEWMGWLKPRGGAVNY.ARKFQGRVT..MTR.......DVYSDTAFLELRSLTSDDTAVYFCTRGKYCTARDYYNWDFEHWGRGAPVTVSS

Cross-donor-identified heavy chains from donor C38 H1-primer data set:
304943       RVQLTQVWAQLRKPGASMRVSCETSGFRRFT........DSKIGWVRQAPGQPFEWMGLMESYWGRVHY.AAQFRDRVT..MTR.......DVDVETAFLELSGLTLADTAIYYCVTAAGTN.....EWAFEWGQGTRVIVSP
gVRC-H1dc38  QVTLVQSGNQLKRPGASVRISCETSGYNFMD.........HFIHWVRQVPGHGPEWLGWVNPRGGGVNY.SRKFQGRFS..MTR.......DVYMETAYLDVTGLSPADTAVYYCARGFGGS.....DWSFLWGQGTLIIVSS

Cross-donor-identified heavy chains from donor C38 G1-primer data set:
gVRC-H2dc38  Q.ALVQSGSQMKKPGDSVRLSCQTSDSAITK.........YFIHWIRQAPGKGLEWIAWISPYGGRVNY.GWQVRDRAT..LTR.......NIHMETVYMDLRGLRPDDTATYYCAMRDYCRDDNCNRWDLGHWGQGSLIVVSA
240171       QVRLIQSGTQMKKPGSSVKISCDTSGYKFVD.........FLIYWFRHVPGREIEWIGWLKPYGGGVNF.NGNFRDRVT..LTR....KSDDTDRGTVYMEISGLRAADTAVYYCTRRGLCD..HCSKWTFEHWGQGTPVIVSS
gVRC-H3dc38  RIELHQSGSQVKKSGASVRISCETSGFKFMD.........SHLHWVRQVAGQRFEWMGWIFTSGGGVNY.ARQFQGRLR..LTR.......DVFSESVFMDLSGLNSGDTGVYYCVKGTGGN.....EWGFVWGQGSLVVVSP
gVRC-H4dc38  RINLDQSGSQVKKSGASVRISCETSGFKFMD.........SHLHWVRQVAGQPFEWMGWIFTSGGGVNY.ARQFQGRLT..LTR.......DVFSETVFMDLSGLNAGDTGVYFCVKGTGGN.....EWGFVWGQGTVVVVSP
gVRC-H5dc38  RINLHQSGSQVKRSGASVRISCETSGFKFMD.........SHLHWVRQVAGQGFEWMGWIFTSGGGVNY.ARQFQGRLT..LTR.......DVFTDTVFMDLSGVNVGDTGVYYCVKGTGGN.....EWGFSWGQGTVVVVSP
gVRC-H6dc38  QVSLVQSGNQLKKPGASVRISCETSGYNFLN.........HFIHWVRQVPGHGLEWLGWINPRGGGVNY.SRNFQGKVS..LTR.......NIDMETVYLDVRGLTPGDTAVYYCARGFGGS.....DWNFVWGQGTRITVSA
gVRC-H7dc38  QVRLVQSGNQVRKPGASVRISCEASGYKFID.........HFIHWVRQVPGHGLEWLGWINPRGGGVNY.SRGFQGKLSMTMTR.......DNFEETAYLDLSRLNPGDTAVYYCARGFAGY.....EWSFLWGQGTLVIVSS
gVRC-H8dc38  QVHLVQSGTQVKKPGASVRVSCETSGFKFLD.........SIIHWFRQAPGEGLFWMGWIKPYTGSVNY.VRRYQGRVS..MTR.......DVYSDTAYMDLSGLNSDDTAVYFCTYGAGDG......WNLVWGQGTLVIVSA
gVRC-H9dc38  RVHLVQSGTQVKKPGASVKVSCETSGFKFLD.........SLIHWVRQAPGQGLYWMGWIKPFRGSVNY.DGYFRGRVS..MTR.......DIYTDTAYMELSGLRSDDTAIYYCAFGAGDG......WDLVWGQGTLVIVSS
gVRC-H10dc38 RVHLVQSGTQVKKPGASVKVSCETSGFKFLD.........SLIHWVRQAPGQGLYWMGWIKPYRGSVNY.DGYFRGRVS..MTR.......DIYTDTAYLELSGLRSDDTAIYYCAFGAGDG......WDLVWGQGTLVIVSS
```

Fig. 4. Sequence comparison of C38 heavy chains identified by cross-donor phylogenetic analysis. (*A*) Sequence alignment of 12 expressed C38 heavy chains from amplifications using H1 and G1 primers, inferred germ-line gene (IGHV1-2*02), and the heavy chains of 13 template VRC01 class antibodies (VRC01, VRC02, VRC03, VRC-PG04, VRC-PG04b, VRC-CH30, VRC-CH31, VRC-CH32, 3BNC60, 3BNC117, 12A12, 12A21, and NIH45-46). Amino acids in V genes that differ from IGHV1-2*02 are highlighted in red. The CDR H3 regions are circled in blue dash line. The nomenclature gVRC-Hn$_{dC38}$ is used for neutralizing heavy chains, and the indexes of two nonneutralizing heavy chains, 304943 and 240171, are shown in italic. Contact residues between VRC01 V-gene segment and gp120 are labeled above the germ-line gene: ●, backbone and side chain contacts; ○, backbone contacts only; *, side chains contacts only. (*B*) Identity/divergence analysis of 10 neutralizing C38 heavy chains. Heavy chain sequences obtained from H1 (*Left*) and G1 (*Right*) primers are plotted as a function of maximal sequence identity to the heavy chains of 13 template VRC01 class antibodies and of sequence divergence from putative germ-line genes. The 10 neutralizing heavy chains are shown as black dots on the plots.

antibody light chains from the antibodyome. Four κ chains, but no λ chains, were found to meet the criteria. Given this result, we performed a second 454 pyrosequencing experiment to amplify only κ chains and obtained greater sequencing depth (*SI Appendix*, Table S3). Nine more κ chains with this sequence motif were identified from the total of 448,125 raw reads.

Sequence alignment of 13 light chain sequences (Fig. 5*A*) showed prevalent use of the IGKV3-20 germ-line gene, the same light chain variable gene used in VRC01, VRC03, and VRC-PG04. These 13 sequences, along with the two antibodyomes where they were identified, were compared with the same set of template VRC01 class antibodies used in the heavy chain analysis (Fig. 1*B*). None of the 13 light chains was found to be >80% identical to any of the template light chains from other donors (Fig. 5*B*), suggesting a lack of sequence homology in their variable genes. Furthermore, a large portion of the κ chain sequences in two antibodyomes, notably closer to germ-line genes on the identity/divergence plots (Fig. 5*B*), were found to be 80–90% identical to the template light chains. These results suggested that random sequences of low divergence, instead of the 13 light chains that resemble the CDR L1 signature of VRC01 class antibodies, would be selected by a pure homology-based method of identification. Collectively, we show with both heavy chains (Figs. 1*B* and 4*B*) and light chains (Fig. 5*B*) that sequence homology to known VRC01 class antibodies from other donors cannot identify such antibodies from donor C38.

De Novo Identification of Functional VRC01 Class Antibodies from Donor C38. Pairing 10 neutralizing heavy chains with 13 candidate light chains may reveal their optimal combinations but would require significant experimental effort. Here we adopted a two-step approach to this problem. In the first step, we used a VRC01 heavy chain to screen 13 light chains. Six of 13 light chains were able to be reconstituted as full antibodies with a VRC01 heavy chain, but only one light chain, named gVRC-L1$_{dC38}$, showed weak neutralization of two HIV-1 isolates on

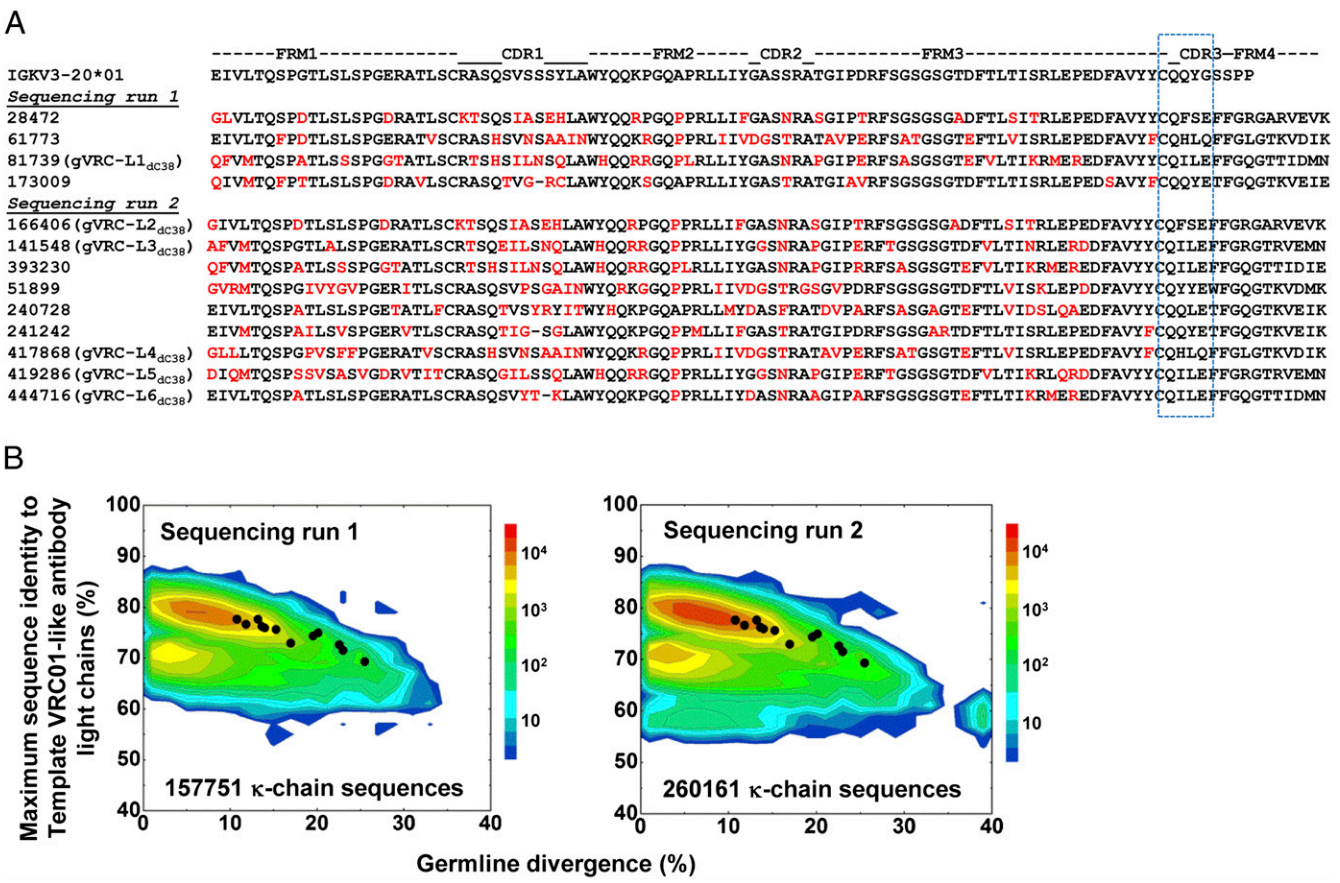


Fig. 5. Identification of VRC01 class antibody light chains from donor C38 antibodyome. (*A*) Thirteen VRC01-like light chains identified from two sequencing experiments are aligned to the putative germ-line gene allele IgKV3-20*01. All 13 light chains possess the same signature: CDR L3 loop of five amino acids and a glutamine or glutamate acid at position 96 (Kabat numbering). Amino acids mutated from the germ-line sequence are colored in red. Light chain sequences that show neutralization when reconstituted with VRC01 heavy chain and gVRC-H3$_{dC38}$ (Fig. 6) are named gVRC-L(n)$_{dC38}$. (*B*) Divergence/identity analysis of donor C38 κ-light chain sequences generated from 454 pyrosequencing with both κ and λ primers (*Left*) and κ primers only (*Right*), with the 13 VRC01-like light chains shown as black dots. Light chain sequences are plotted as a function of maximal sequence identity to the light chains of 13 template VRC01 class antibodies (VRC01, VRC02, VRC03, NIH45-46, VRC-PG04, VRC-PG04b, VRC-CH30, VRC-CH31, VRC-CH32, 3BNC60, 3BNC117, 12A12, and 12A21) and of sequence divergence from inferred germ-line alleles.

a seven-virus panel (Fig. 6*A*). In the second step, we used the most potent C38 heavy chain obtained from cross-donor analysis, gVRC-H3$_{dC38}$, and the only effective C38 light chain from previous screening, gVRC-L1$_{dC38}$, to search for their respective partner chains. With heavy chain gVRC-H3$_{dC38}$, 8 of 13 light chains were expressed as full antibodies (Fig. 6*B*, *Upper*) compared with 6 with a VRC01 heavy chain (Fig. 6*A*). When tested on seven HIV-1 isolates, six of eight reconstituted antibodies showed neutralization with various breadth and potency (Fig. 6*B*, *Upper*), as opposed to a single neutralizer when paired with a VRC01 heavy chain (Fig. 6*A*). Interestingly, gVRC-L1$_{dC38}$ remained the most effective of C38 light chains tested, suggesting that the initial screening with the VRC01 heavy chain was unbiased. With the light chain gVRC-L1$_{dC38}$, we observed broader neutralization from most C38 heavy chains tested, with the isolate ZM109.4 neutralized by three antibodies, and identified gVRC-H3$_{dC38}$/gVRC-L1$_{dC38}$ as the optimal pair of C38 heavy and light chains (Fig. 6*B*, *Lower*).

The marked difference of breadth and potency seen in the two screening experiments of C38 light chains (Fig. 6 *A* and *B*, *Upper*) suggested that some unfavorable interactions between C38 light chains and the VRC01 heavy chain might underlie deteriorated function when paired. The comparison of 13 light chain sequences (Fig. 5*A*) offered some clues to the nonexpression of five tested light chains, four of which possess a nonphenylalanine (F) residue at position 97—the last residue of the CDR L3 loop—and the fifth sequence, with an index of 393230, is 97% identical to gVRC-L1$_{dC38}$ but with a glutamate (E) mutated to arginine (R) in the CDR L2 loop. The best C38 heavy/light chain pair, gVRC-H7$_{dC38}$/gVRC-L1$_{dC38}$, showed an ~10-fold decrease in potency compared with gVRC-Hd$_{dC38}$/VRC01L (Fig. 6*B*, *Lower*), suggesting that the C38 light chain might be suboptimal for complementing VRC01 class antibody heavy chains, even those from the same donor. A possible explanation might involve the long CDR L1 loops. All 13 C38 light chains have one or zero deletions in the CDR L1 region (Fig. 5*A*), whereas VRC01 or other light chains of this class have two or more residues deleted or mutated to glycines in this region.

With the functional pairing of 10 heavy chains and 6 light chains from donor C38, we next sought to understand the evolutionary relationship of these sequences. We calculated maximum-likelihood phylogenetic trees rooted by their respective germ-line genes (Fig. 6*C*). Similar topology was observed for the heavy chain and light chain dendrograms, with the optimal pair, gVRC-H7$_{dC38}$/gVRC-L1$_{dC38}$, formed by sequences from two corresponding branches, suggesting that this chimera might resemble a native pair.

A

Neutralization (IC_{50}) of antibodies reconstituted by pairing C38 light chains with VRC01 heavy chain.

Antibody	Q842.d12	DU156.12	ZM109.4	UG037.8	Q23.17	JR-FL	Yu2	MuLV
VRC01 H/28472	>50	>50	>50	>50	>50	>50	>50	>50
VRC01 H/61773	>50	>50	>50	>50	>50	>50	>50	>50
VRC01 H/gVRC-L1$_{dC38}$	>50	>50	>50	19.3	>50	>50	12.1	>50
VRC01 H/gVRC-L2$_{dC38}$	>50	>50	>50	>50	>50	>50	>50	>50
VRC01 H/gVRC-L4$_{dC38}$	>50	>50	>50	>50	>50	>50	>50	>50
VRC01 H/gVRC-L6$_{dC38}$	>50	>50	>50	>50	>50	>50	>50	>50
VRC01 H/VRC01 L	0.019	0.088	0.127	0.109	0.068	0.014	0.136	>50

B

Screening of light-chain partners for gVRC-H3$_{dC38}$ and of heavy-chain partners for gVRC-L1$_{dC38}$ by neutralization (IC_{50}).

Antibody	Q842.d12	DU156.12	ZM109.4	UG037.8	Q23.17	JR-FL	YU2	MuLV
gVRC-H3$_{dC38}$/28472	>50	>50	>50	>50	>50	50.0	>50	>50
gVRC-H3$_{dC38}$/61773	>50	>50	>50	>50	>50	>50	>50	>50
gVRC-H3$_{dC38}$/gVRC-L1$_{dC38}$	0.775	0.907	>50	0.449	>50	0.101	0.583	>50
gVRC-H3$_{dC38}$/gVRC-L2$_{dC38}$	>50	>50	>50	>50	>50	9.33	>50	>50
gVRC-H3$_{dC38}$/gVRC-L3$_{dC38}$	0.641	1.06	>50	2.63	>50	0.032	0.213	>50
gVRC-H3$_{dC38}$/gVRC-L4$_{dC38}$	>50	>50	>50	22.9	>50	4.99	11.7	>50
gVRC-H3$_{dC38}$/gVRC-L5$_{dC38}$	10.7	>50	>50	>50	>50	0.026	0.136	>50
gVRC-H3$_{dC38}$/gVRC-L6$_{dC38}$	>50	>50	>50	>50	>50	0.062	0.72	>50
gVRC-H2$_{dC38}$/gVRC-L1$_{dC38}$	>50	>50	>50	>50	>50	>50	>50	>50
gVRC-H4$_{dC38}$/gVRC-L1$_{dC38}$	>50	>50	>50	>50	>50	1.95	7.6	>50
gVRC-H5$_{dC38}$/gVRC-L1$_{dC38}$	5.31	11.4	>50	3.54	36.2	0.669	2.21	>50
gVRC-H6$_{dC38}$/gVRC-L1$_{dC38}$	14.5	>50	>50	0.697	12	0.267	0.454	>50
gVRC-H7$_{dC38}$/gVRC-L1$_{dC38}$	0.278	0.473	0.232	0.152	0.541	0.026	0.326	>50
gVRC-H8$_{dC38}$/gVRC-L1$_{dC38}$	>50	>50	>50	5.43	>50	5.4	21.9	>50
gVRC-H9$_{dC38}$/gVRC-L1$_{dC38}$	0.051	3.51	4.48	0.385	0.235	0.018	0.095	>50
gVRC-H10$_{dC38}$/gVRC-L1$_{dC38}$	0.073	2.72	7.37	0.557	0.197	0.017	0.117	>50
VRC01 H/VRC01 L	0.019	0.088	0.127	0.109	0.068	0.014	0.136	>50

C

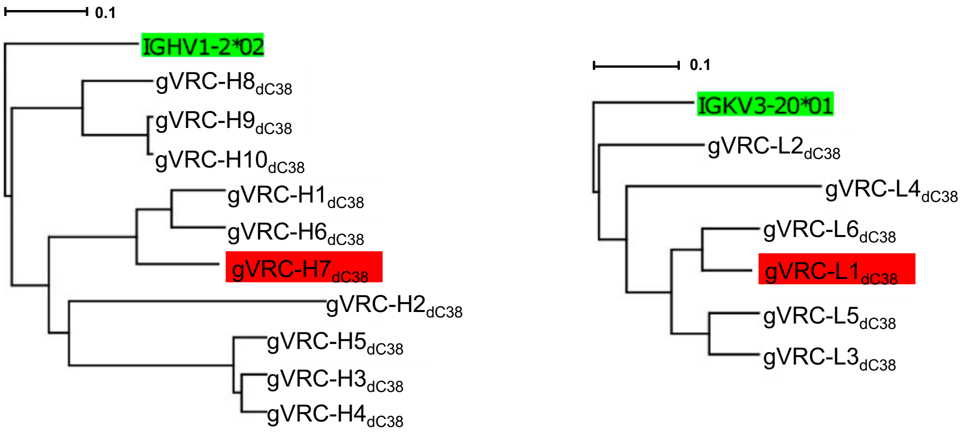


Fig. 6. Identification of unique VRC01 class antibodies by pairing heavy and light chains. (*A*) Neutralization of six reconstituted antibodies by pairing VRC01 heavy chain with C38 light chains. (*B*) Neutralization screening of functional partner chains for C38 heavy chain, gVRC-H3$_{dC38}$, and C38 light chain, gVRC-L1$_{dC38}$. IC_{50} values listed in *B* and *C* represent antibody concentration in μg/mL required to achieve 50% neutralization (IC_{50}), and color coding indicates the potency, with $IC_{50} < 1$ μg/mL in red shading, 1 μg/mL < IC_{50} <50 μg/mL in green shading, and no shading for IC_{50} >50 μg/mL. (*C*) Maximum-likelihood trees of functional C38 heavy chains and light chains rooted in their respective germ-line genes, with the most potent heavy/light chain pair highlighted in red shading. Bars representing 0.1 changes per nucleotide site are shown.

Discussion

Biological sequencing has progressed from analyzing single genes (22–24) to genomes (25–27) and, more recently, to the analysis of multiple genomes (28). Similarly, analysis of antibodies has progressed from single antibody chains to whole expressed repertoires and is now poised to analyze antibodyomes from multiple individuals. Previous studies of antibodyomes from HIV-1–infected individuals mainly focused on the fundamental questions related to antibody maturation (7, 11, 29–31) and somatic variation (21), whereas here we extend the scope of questions that can be addressed to a more practical domain: antibody identification. The high-throughput sieving method described here, cross-donor phylogenetic analysis for heavy chain and motif matching for light chain, can identify VRC01 class antibodies from a donor sample even if their frequencies are low, e.g., <0.0004% for donor C38, where ∼80% of the neutralizing activity is not depleted by RSC3. Identification of such antibodies would not be possible with homology-based sequence analysis, as the sequence identity to known the VRC01 class antibody is below the threshold of recognition for both heavy and light chains (*SI Appendix*, Table S13). Given the potential of VRC01 class antibodies as a vaccine template, our de novo approach should have significant implications for HIV-1 vaccine research related to this important antibody class. The ability to identify, and as a result, study the development of VRC01 class antibodies from donor samples—and potentially from vaccines—should help illuminate the appropriateness and feasibility of class-based elicitation strategies, such as germ-line activation, for obtaining an HIV-1 vaccine (32, 33).

It may be possible to apply the methods described here to de novo identification of antibodies of other types, although each case of antibody identification will depend on the bioinformatic or evolutionary signatures particular to the antibody of interest. The success of such de novo identification may depend on the similarity of the target antibodies to a template antibody, and in this regard, it is advantageous to study antibodies of a class, meaning they are derived from similar B-cell ontogenies and recognize similar epitope, despite being elicited in different individuals. With HIV-1, two types of antibodies form classes: the VH1-2–derived VRC01 class and the VH1-69–derived CD4-induced antibodies (31, 34); a third potential class may be formed by VH3-derived antibodies that target the first and second variable regions on HIV-1 gp120 (9, 35–37). Such class-derived antibodies are found against other pathogens, such as with influenza, where highly similar hemagglutinin stem-directed antibodies all derive from the VH1-69 germ-line gene and use the same mode of recognition (38, 39).

It is worth noting that a related phylogenetic method has been reported for repertoire analysis, intradonor phylogenetic analysis, which has been applied to the broadly HIV-1–neutralizing lineages that include antibodies PGT135-137, 10E8, and PGT141-145 (21, 40). Both the cross-donor and intradonor phylogenetic methods are capable of finding new antibodies, but the differences are (*i*) cross-donor analysis identifies evolutionarily similar sequences from a heterologous donor, whereas intradonor analysis identifies somatic variants of the template(s) from the same donor; and (*ii*) cross-donor analysis proved effective for the VRC01 class heavy chains, whereas intradonor analysis can be applied to any type of antibodies, heavy and light chains alike. The innovative use of phylogenetic analysis will likely have other applications in the analysis of antibody repertoire and maturation.

The technical advances presented here build on a detailed understanding of individual antibody sequences enabled by previously developed computational tools (41–44), which in our analysis are integrated to derive population-based metrics for antibodyomes. This type of analysis, antibodyomics, should have widespread utility for understanding the humoral immune response in natural infection (7, 29, 31) or by vaccination (6), as well as for de novo identification of functional antibodies from neutralizing sera. Meanwhile, the development of new antibody sequencing technologies, as exemplified by the recent advance in the paired sequencing of antibody heavy and light chains (45), will continue to improve the quality of data available and facilitate the antibodyomics analysis. Such new technologies should provide increasingly powerful tools to answer questions of antibody development and vaccine elicitation (46–48).

Materials and Methods

Experimental and computational methods used in this study are briefly summarized here, with details presented in *SI Appendix, SI Materials and Methods*.

Human Specimens. The sera and PBMCs were obtained from HIV-1–infected donors (*SI Appendix*, Table S1) enrolled in investigational review board–approved clinical protocols at the National Institute of Allergy and Infectious Diseases, including the VRC200 protocol. At the time of sampling, all HIV-1–infected donors were off antiretroviral treatment.

Antibody Expression, Purification, and Neutralization. Similar procedures and measurements were used as described in a previous study (7, 21).

454 Sample Preparation and 454 Pyrosequencing. The procedure was similar to ref. 7 but included new primers aimed to capture highly matured heavy chain sequences. For IGHV1 amplification, the H1 primers used in previous study extended from the end of the V-gene leader sequence into the start of the V region, whereas the G1 primers (10) used in the current study were more upstream and overlapped only with the V-gene leader sequence.

Computational Analysis of 454 Pyrosequencing Data. The tools used for analysis were implemented as a suite of PERL scripts, named Antibodyomics1.0. The analysis consisted of two steps. In the first step, sequences encompassing antibody heavy chain variable domains generated by 454 pyrosequencing were processed using a bioinformatics pipeline (7, 21) to derive biologically defined properties (such as germ-line divergence and sequence identity), as well as to improve sequence quality by correcting the most common sequencing errors. In pipeline processing, sequences were reformatted, assigned to germ-line gene families, error-corrected using a template-based algorithm, compared with a set of user-specified antibodies, and subjected to the determination and comparison of CDR H3 regions. Full-length heavy chain variable domain sequences with detailed annotation were outputted for subsequent analyses. In the second step, various systems-level analyses were carried out to mine the antibody population, including detailed analyses of antibody properties (e.g., CDR H3 lineages and their distribution), lineage rank, and cross-donor phylogenetic analysis.

Data Deposition and Software Distribution. The 10 functional heavy chain sequences ($\text{gVRC-H1-10}_{\text{dC38}}$) and 6 functional light chain sequences ($\text{gVRC-L1-6}_{\text{dC38}}$) identified from donor C38 antibodyomes have been deposited in GenBank under accession numbers KF306044–KF306069. The 454 pyrosequencing data sets have been deposited in the NCBI Short Reads Archives under accession number SRP026397. The software suite described here, Antibodyomics1.0, can be obtained by request from J.Z., P.D.K., or L.S.

ACKNOWLEDGMENTS. We thank J. Almeida and D. Douek for protocols of PBMC cDNA preparation and for helpful discussions; H. Coleman, S. Ho, M. Park, B. Schmidt, and A. Young for 454 pyrosequencing at the National Institutes of Health Intramural Sequencing Center; and J. Stuckey for assistance with figures. We also thank Rahul Kohli and members of the Structural Biology Section and Structural Bioinformatics Core, Vaccine Research Center, for discussions or comments on the manuscript. Support for this work was provided by the Intramural Research Program of the Vaccine Research Center, National Institute of Allergy and Infectious Diseases and the National Human Genome Research Institute, National Institutes of Health, and grants from the International AIDS Vaccine Initiative's Neutralizing Antibody Consortium.

1. Colman PM (1988) Structure of antibody-antigen complexes: Implications for immune recognition. *Adv Immunol* 43:99–132.
2. Padlan EA (1977) Structural basis for the specificity of antibody-antigen reactions and structural mechanisms for the diversification of antigen-binding specificities. *Q Rev Biophys* 10(1):35–65.
3. Wilson IA, Rini JM, Fremont DH, Fieser GG, Stura EA (1991) X-ray crystallographic analysis of free and antigen-complexed Fab fragments to investigate structural basis of immune recognition. *Methods Enzymol* 203:153–176.
4. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress in modeling of protein structures and interactions. *Science* 310(5748):638–642.
5. Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: Implications for the nature of 'fold space', and structure and function prediction. *Curr Opin Struct Biol* 16(3):393–398.
6. Reddy ST, et al. (2010) Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* 28(9):965–969.
7. Wu X, et al.; NISC Comparative Sequencing Program (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333(6049):1593–1602.
8. Zhou T, et al. (2010) Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* 329(5993):811–817.
9. Kwong PD, Mascola JR (2012) Human antibodies that neutralize HIV-1: Identification, structures, and B cell ontogenies. *Immunity* 37(3):412–425.
10. Scheid JF, et al. (2011) Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333(6049):1633–1637.
11. Klein F, et al. (2013) Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* 153(1):126–138.
12. West AP, Jr., Diskin R, Nussenzweig MC, Bjorkman PJ (2012) 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 USA* 109(30):E2083–2090.
13. Mardis ER (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9:387–402.
14. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24(3):133–141.
15. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: An integrative approach. *Nat Rev Genet* 11(7):476–486.
16. Fischer N (2011) Sequencing antibody repertoires: The next generation. *MAbs* 3(1): 17–20.
17. Prabakaran P, et al. (2012) Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics* 64(5):337–350.
18. Krause JC, et al. (2011) Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *J Immunol* 187(7):3704–3711.
19. Wu X, et al. (2010) Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* 329(5993):856–861.
20. Tiller T, et al. (2008) Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *J Immunol Methods* 329(1-2):112–124.
21. Zhu J, et al. (2012) Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Frontiers Microbiol* 3:315.
22. Min Jou W, Haegeman G, Ysebaert M, Fiers W (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* 237(5350):82–88.
23. Fiers W, et al. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature* 260(5551):500–507.
24. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94(3):441–448.
25. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496–512.
26. Lander ES, et al.; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
27. Venter JC, et al. (2001) The sequence of the human genome. *Science* 291(5507): 1304–1351.
28. Chen R, et al. (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148(6):1293–1307.
29. Liao HX, et al.; NISC Comparative Sequencing Program (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* 496(7446):469–476.
30. Briney BS, Willis JR, Crowe JE, Jr. (2012) Human peripheral blood antibodies with long HCDR3s are established primarily at original recombination using a limited subset of germline genes. *PLoS ONE* 7(5):e36750.
31. Zhou T, et al. (2013) Multidonor Analysis Reveals Structural Elements, Genetic Determinants, and Maturation Pathway for HIV-1 Neutralization by VRC01-Class Antibodies. *Immunity* 39:245–258.
32. Jardine J, et al. (2013) Rational HIV immunogen design to target specific germline B cell receptors. *Science* 340(6133):711–716.
33. McGuire AT, et al. (2013) Engineering HIV envelope protein to activate germline B cell receptors of broadly neutralizing anti-CD4 binding site antibodies. *J Exp Med* 210(4): 655–663.
34. Huang CC, et al. (2004) Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proc Natl Acad Sci USA* 101(9):2706–2711.
35. Walker LM, et al.; Protocol G Principal Investigators (2009) Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* 326(5950):285–289.
36. Walker LM, et al.; Protocol G Principal Investigators (2011) Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* 477(7365):466–470.
37. McLellan JS, et al. (2011) Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* 480(7377):336–343.
38. Kwong PD, Wilson IA (2009) HIV-1 and influenza antibodies: Seeing antigens in new ways. *Nat Immunol* 10(6):573–578.
39. Ekiert DC, et al. (2009) Antibody recognition of a highly conserved influenza virus epitope. *Science* 324(5924):246–251.
40. Zhu J, et al.; NISC Comparative Sequencing Program (2013) Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc Natl Acad Sci USA* 110(16):6470–6475.
41. Munshaw S, Kepler TB (2010) SoDA2: A Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* 26(7):867–872.
42. Gaëta BA, et al. (2007) iHMMune-align: Hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* 23(13):1580–1587.
43. Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36(Web Server issue):W503–W508.
44. Souto-Carneiro MM, Longo NS, Russ DE, Sun HW, Lipsky PE (2004) Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J Immunol* 172(11): 6790–6802.
45. DeKosky BJ, et al. (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* 31(2):166–169.
46. Hilleman MR (2002) Overview of the needs and realities for developing new and improved vaccines in the 21st century. *Intervirology* 45(4-6):199–211.
47. Karlsson Hedestam GB, et al. (2008) The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nat Rev Microbiol* 6(2):143–155.
48. Haynes BF, Kelsoe G, Harrison SC, Kepler TB (2012) B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat Biotechnol* 30(5):423–433.