Supplementary Materials and Methods for

Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection

Xueling Wu[1,2*], Zhenhai Zhang[3,4,5*], Chaim A. Schramm[3*],
M. Gordon Joyce[1*], Young Do Kwon[1*], Tongqing Zhou[1*], Zizhang Sheng[3*],
Baoshan Zhang[1], Sijy O'Dell[1], Krisha McKee[1], Ivelin S. Georgiev[1], Gwo-Yu Chuang[1],
Nancy S. Longo[1], Rebecca M. Lynch[1], Kevin O. Saunders[1],
Cinque Soto[1], Sanjay Srivatsan[1], Yongping Yang[1],
Robert T. Bailer[1], Mark K. Louder[1],
NISC Comparative Sequencing Program[6], James C. Mullikin[6],
Mark Connors[7],
Peter D. Kwong[1#], John R. Mascola[1#], and Lawrence Shapiro[1,3#]

RUNNING TITLE: Maturation of the VRC01-antibody lineage

[1] Vaccine Research Center, and [7] Laboratory of Immunoregulation, National Institute of Allergy and Infectious Diseases, and [6] NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
[2] Aaron Diamond AIDS Research Center, Rockefeller University, New York, NY 10016, USA.
[3] Department of Biochemistry and Molecular Biophysics and Department of Systems Biology, Columbia University, New York, NY 10032, USA.
[4] State Key Laboratory of Organ Failure Research and [5] National Clinical Research Center for Kidney Disease, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, 510515, China

* Equal contribution.

[#] To whom correspondence should be addressed:
E-mail: pdkwong@nih.gov (P.D.K.), jmascola@nih.gov (J.R.M.), lss8@columbia.edu (L.S.)

Extended Experimental Procedures

Human specimens. The sera and peripheral blood mononuclear cells (PBMC) described in this study were collected from an HIV-1 infected individual, donor 45, who enrolled in an Institutional Review Board approved clinical protocol at the National Institute of Allergy and Infectious Diseases. Donor 45 was infected with a clade B virus for more than 15 years and never received anti-retroviral treatment. He was a slow progressor with CD4 T-cell counts over 500 cells/μl and plasma HIV-1 RNA values fluctuating around 10,000 copies/ml (Table S2A).

Expression and purification of gp120 probes. The gp120 core-derived probes RSC3 and its Δ371I mutant ΔRSC3 (Wu et al., 2010) and the gp120 outer domain (OD)-derived probes HG3.2 and its D368R mutant ΔHG3.2 (Joyce et al., 2013) were expressed by transient transfection of 293F cells as described. Briefly, genes encoding the proteins of interest were each synthesized with a C-terminal His tag (GeneArt, Regensburg, Germany), and cloned into a mamalian CMV/R expression vector (Barouch et al., 2005). Proteins were produced by transient tranfection using 293fectin (Invitrogen, Carlsbad, CA) in 293F cells (Invitrogen) maintained in serum-free free-style medium (Invitrogen). Culture supernatants were harvested 5 - 6 days after transfection, filtered through a 0.45 μm filter, and concentrated with buffer-exchange into 500 mM NaCl, 50 mM Tris (pH 8.0). Proteins were purified by Co-NTA (cobalt-nitrilotriacetic acid) chromatography method using a HiTrap IMAC HP column (GE Healthcare, Piscataway, NJ). The peak fractions were collected, and further purified by gel-filtration using a HiLoad 16/60 Superdex 200 pg column (GE Healthcare). The fractions containing monomers of each protein were combined, concentrated and flash frozen at -80℃.

Antibody expression and purification. As previously described (Li et al., 2012; Scheid et al., 2011; Wu et al., 2010), the anti-CD4bs mAbs VRC01, VRC02, VRC03, VRC06, VRC06b, NIH45-177, NIH45-243 and NIH45-46 were isolated and cloned from Donor 45. Other cloned mAbs were expressed by co-transfection of 293F cells with equal amount of the paired heavy and light chain plasmids and purified using a

recombinant protein-A column (GE Healthcare). Antibody sequences derived from deep-sequencing were synthesized and cloned into the CMV/R expression vector containing the constant regions of IgG1. The full-length IgGs were then expressed and purified similarly.

HIV-1 neutralization assay. Neutralization was measured using HIV-1 Env-pseudoviruses to infect TZM-bl cells as described (Li et al., 2005; Seaman et al., 2010; Wu et al., 2009). Neutralization curves were fit by nonlinear regression using a 5-parameter hill slope equation as described (Seaman et al., 2010). The 50% and 80% inhibitory concentrations (IC_{50} and IC_{80}) were reported as the antibody concentrations required to inhibit infection by 50% and 80% respectively.

Probed single B-cell sorting and mAb cloning. As described previously (Wu et al., 2010), the Avi-tagged probes were biotinylated using the biotin ligase Bir A (Avidity, Denver, CO) and then conjugated with streptavidin-allophycocyanin (SA-APC) (Invitrogen) for RSC3 and HG3.2 and streptavidin-phycoerythrin (SA-PE) (Sigma) for ΔRSC3 and ΔHG3.2. About 20 million donor 45 PBMCs were stained with the probes as indicated and with an antibody cocktail to select for IgG+ B cells. The stained PBMCs were sorted using a modified 3-laser FACSAria cell sorter using the FACSDiva software (BD Biosciences). Single cells with the phenotype of CD3-, CD8-, aqua blue-, CD14-, CD19+, CD20+, IgG+, IgM-, RSC3+ and ΔRSC3- or HG3.2+ and ΔHG3.2- were sorted into 96-well PCR plates containing 20 µl of lysis buffer per well. The lysis buffer contained 0.5 µl of RNase Out (Invitrogen), 5 µl of 5x first strand buffer (Invitrogen), 1.25 µl of 0.1M DTT (Invitrogen) and 0.0625 µl of Igepal (Sigma). The PCR plates with sorted cells were stored at -80˚C. For reverse transcription, the frozen plates were thawed at room temperature and added into each well 3 µl of random hexamers (Gene Link, Hawthorne, NY) at 150 ng/µl, 2 µl of dNTP mix, each at 10 mM, and 1 µl of SuperScript III (Invitrogen), followed by the thermocycle at 42^{o}C for 10 min, 25^{o}C for 10 min, 50^{o}C for 60 min and 94^{o}C for 5 min. The cDNA plates were stored at -20^{o}C, and the IgH, Igκ and Igλ variable region genes were independently amplified by nested PCR as described (Scheid et al., 2011; Tiller et al., 2008; Wu et al., 2010). PCR products

were sequenced and selected for re-amplification with custom primers for subsequently cloning into the corresponding Igγ1, Igκ and Igλ expression vectors as previously described (Tiller et al., 2008; Wu et al., 2010).

IgG gene family analysis. The IgG heavy and light chain nucleotide sequences of the variable region were analyzed with JoinSolver® (http://Joinsolver.niaid.nih.gov) (Souto-Carneiro et al., 2004) and IMGT/V-Quest (http:/www.imgt.org/IMGT_vquest/share/textes/) (Brochet et al., 2008). The mAb V_K gene use was determined by homology to germline genes in the major 2p11.2 IGK locus (Malcolm et al., 1982). The mAb D gene use was determined by homology to genes in the major 14q32.33 IGH locus. A combination of consecutive matching length with a +1/-2.02 scoring algorithm in the context of the V to J distance was applied for determining IGHD alignments and VD and DJ junctions in mutated sequences. Immunoglobulin rearrangements were grouped into clades based upon the VDJ gene use, similarity of replacement and silent mutations and the CDR3 identity.

Construction of the HIV-1 envelope sequence phylogenetic trees. The HIV-1 gp160 protein sequences of isolates used in the neutralization assays were aligned using MUSCLE, for multiple sequence comparison by log-expectation (Edgar, 2004a, b). The protein distance matrix was calculated by “protdist” using the Jones-Taylor-Thornton model (Jones et al., 1992), and the dendrogram was constructed using the neighbor-joining method (Kuhner and Felsenstein, 1994) by “Neighbor”. The analysis was performed at the NIAID Biocluster (https://niaid-biocluster.niaid.nih.gov/). The trees were displayed with Dendroscope (Huson et al., 2007).

454 pyrosequencing. As described (Wu et al., 2011), mRNA was extracted from 10-20 million PBMC into 200 µl of elution buffer (Oligotex Direct mRNA Mini Kit, Qiagen), then concentrated to 10-30 µl by centrifuging the buffer through a 30 kD micron filter (Millipore). The reverse transcription was performed in one or multiple 35 µl reactions, each composed of 13 µl of mRNA, 3 µl of oligo$(dT)_{12-18}$ at 0.5 µg/µl (Invitrogen), 7 µl of 5x first strand buffer (Invitrogen), 3 µl of RNase Out (Invitrogen), 3 µl of 0.1M DTT

(Invitrogen), 3 µl of dNTP mix (each at 10 mM), and 3 µl of SuperScript II (Invitrogen). The reactions were incubated at 42^{o}C for 2 hours. The cDNAs from each sample were combined, purified and eluted in 10-20 µl of elution buffer (NucleoSpin Extract II kit, Clontech). Each 1 µl of the resultant cDNA was roughly equivalent of the transcripts from 1 million PBMCs. The immunoglobulin gene-specific PCRs were set up in a total volume of 50 µl, using 3-5 µl of the cDNA as template (equivalent of transcripts from 3-5 million PBMCs) and the Phusion High-Fidelity DNA Polymerase system (Finnzymes). The reaction mix was composed of water, 10 µl of 5x buffer, 2 µl of dNTP mix (each at 10 mM), 1-2 µl of primers or primer mixes (below) at 25-50 µM, and 1 µl of Phusion DNA polymerase. For IGHV1 amplification, the forward primers are a pool of primers (G1) of equal parts of the seven VH1 primers published by Scheid et al (Scheid et al., 2011); the reverse primers are a mix of equal 3'CγCH1#2, 5'GGGGAAGACCGATGGGCCCTTGGT3' and 3'CµCH1, 5'GGGAATTCTCACAGGAGACGA3'. For IGKV3 amplification, the forward primer is 5'L-VK3, 5'CTCTTCCTCCTGCTACTCTGGCTCCCAG3', and the reverse primer is 3'CK1, 5'CAGCAGGCACACAACAGAGGCAGTTCC3'. The primers each contained the appropriate adaptor sequences (XLR-A or XLR-B) for subsequent 454 pyrosequencing. The PCRs were initiated at 98^{o}C for 30 sec, followed by 25 cycles of 98^{o}C for 10 sec, 58^{o}C for 30 sec, and 72^{o}C for 30 sec, followed by 72^{o}C for 10 min. The PCR products at the expected size (~500bp) were gel extracted (Qiagen) and quantified using Qubit (Life Technologies, Carlsbad, CA). Library concentrations were determined using the KAPA Biosystems qPCR system (Woburn, MA) with 454 standards provided in the KAPA system. 454 pyrosequencing of the PCR products was performed on a GS FLX sequencing instrument (Roche-454 Life Sciences, Bradford, CT) using the manufacturer's suggested methods and reagents. The quality of each run was assessed by analysis of internal control sequences included in the 454 sequencing reagents. Reports were generated for each region of the PicoTiterPlate (PTP) for both the internal controls and the samples.

Bioinformatics analysis of NGS datasets. The sequencing reads for both heavy and light chains were processed through an Antibodyomics pipeline implemented in Python

(Figure S2A, B). This pipeline is currently available upon request from the authors and will be described more fully and made publically available in a subsequent publication. Briefly, (1) sequences were filtered by read length, retaining only transcripts with a length of 300 to 600 nucleotides. (2) Transcripts were assigned sequential serial numbers starting at 000001 for each sample; the serial number is preceded by a letter indicating sample collection time. (3) Each read was assigned to a germline variable (V) gene using an in-house implementation of IgBLAST (http://www.ncbi.nlm.nih.gov/igblast/). (4) The divergence from the assigned V gene and sequence identity to each probe-identified VRC01 lineage antibody was calculated via ClustalW2 (Larkin et al., 2007). To visualize the data, germline VH gene origins of all sequences were assigned using BLAST, and identity-divergence (I-D) plots were produced to show the behavior of bulk transcripts with respect to their divergence from germline and identity to probe-identified antibodies from each clade.

Inspection of I-D plots using antibodies from different VRC01 lineage clades as the identity referents (Y-axis) reveals several important features of the lineage (Figure 2, left panels). A subpopulation of sequences in the same CDR H3 or VL group as the identity referent (i.e. VRC08) or one of the other probe-identified antibodies from the same clade have ≥ 80% (heavy) or 90% (light) identity (yellow and purple dots) to the referent antibody. For example the purple dots in the larger left panel (Figure 2A) show curated heavy chain sequences from the VRC08 CDR H3 group, and these sequences show > 80% identity to the full VRC08 heavy chain sequence. In contrast, antibodies from other clades of the lineage (labeled red X's) and bulk cross-donor positive sequences (blue dots) tend to overlap with unrelated background sequences (gray contours) with identity up to 70% (heavy chains) or 80% (light chains). Thus, this sequential sieve analysis and I-D plot visualization identifies heavy and light chain sequences that are highly related to reference sequences from each of the major antibody clades.

These visualizations also track the development of each VRC01-clade longitudinally (Figure 2, right panels, Figures S2C-F). Clusters of curated sequences in the same CDR H3 or VL groups as the probe-isolated antibodies from clades 01+07, 03+06, and 08 (purple dots) provide insight into the evolution of the clades over time. For the 08 clade, clusters of sequences separated from the main body of background transcripts

were clearly visible at ~85% identity for the heavy chain (Figure 2A) and ~90% for the light chain (Figure 2B) from 2001 forward. Between 2001 and 2009, multiple clusters of related sequences are generally visible, with the more closely related heavy chain cluster reaching nearly 100% identity to VRC08 in 2001 before declining to around 90% identity. (VRC08 was isolated from a 2002 sample.) For clade 08 light chains (Figure 2B), the related sequences reach maximal identity in 2006. Similar patterns were observed on I-D plots for clades 01+07 and 03+06 as well (Figures S2C-F).

Extraction of potential VRC01-class heavy chain sequences. To identify potential VRC01-class heavy chain sequencing reads, multiple iterations of a modified cross-donor phylogenetic analysis similar to that in (Wu et al., 2011; Zhu et al., 2013) were performed (Figure S2A). Primarily, all reads matching the length criteria and having any VH gene match were included in the analysis, without requiring assignment to VH1-2. This results in the inclusion of a much higher fraction of VH1-2 transcripts being defined as “cross-donor positive,” as well as the inclusion of a number of transcripts using other VH genes (~26% of all cross-donor positive sequences). In addition, many additional VRC01 class antibodies have been discovered in the interim, and the use of a larger number of exogenous templates captures a larger portion of the phylogenetic tree, including many low-divergence reads (~5% of cross-donor positive sequences were < 10% mutated from germline) that were ultimately determined not to be part of the VRC01 lineage.

In each iteration, a FASTA-format file containing 3,000 donor 45 NGS reads, the IGHV1-2*02 germline gene sequence, and the heavy chain sequences of 19 known VRC01-class sequences (VRC01, VRC02, VRC03, NIH45-46, VRC-PG04, VRC-PG04B, VRC-CH30, VRC-CH31, VRC-CH32, VRC-CH33, VRC-CH34, 3BNC60, 3BNC117, 12A12, 12A21, 1NC9, 1B2530, 8ANC131, and 8ANC134) from a variety of donors was created. A neighbor-joining tree was constructed from the file using ClustalW2 and rooted on the IGHV1-2*02 germline sequence. The minimum sub-tree containing all known VRC01-class antibodies was selected and all NGS reads within this minimum sub-tree were extracted and passed to the next iteration of cross-donor analysis. When the analysis converged (more than 95% of input reads were retained in the minimum

sub-tree in the last cycle), the remaining reads were considered “cross-donor positive” and carried forward for further analysis.

The modifications described above were made to increase to sensitivity of the cross-donor analysis, with the goal of describing the VRC01 lineage as completely as possible. As expected, this led to a corresponding decrease in specificity. We compensated for this effect by clustering the selected sequences (“Construction of CDR H3 groups” and “Construction of light chain variable region (VL) groups”, below) and keeping only the largest clusters of closely related sequences. In addition, representative sequences were synthesized from each retained cluster (“Selection of representative reads for expression and neutralization assays”, below), and only functionally validated groups were considered to be part of the VRC01 lineage. Finally, sequences from the accepted neutralizing groups were subjected to rigorous quality-control procedures (“Curation of high-quality sequences for analyzing the VRC01 lineage”, below).

Construction of CDR H3 groups. The CDR H3s of cross-donor positive sequences from all time points were extracted using the conserved flanking CxR/K and WGxG motifs. CDR H3s (including these surrounding conserved motifs) were assembled into a single FASTA-format file, together with CDR H3s from the 39 probe-identified antibodies, and clustered using BLASTClust (http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html) with thresholds of 90% coverage and 90% identity. While individual sequences may reflect PCR error or sequencing error, these are expected to average out in a group of closely related transcripts. CDR H3 groups are therefore more likely to be biologically relevant than the raw reads. While average full-length sequence diversity ranged up to 25% within in a single clade, the defined CDR H3 groups averaged only 5-15% full-length sequence diversity. For example, VRC01 and VRC07b fall into different CDR H3 groups, despite being members of the same phylogenetic clade (Figure S4). The CDR H3 groups were sorted based on size, and those with at least 300 raw sequences were selected for further analysis.

Identification of potential VRC01-class light chain sequences. Due to their short length and lower maturation/mutation rate, cross-donor analysis fails to reliably identify VRC01-class light chains. Structural studies, however, have demonstrated that VRC01-class light chains have a CDR L3 length restricted to exactly 5 amino acids (West et al., 2012; Zhou et al., 2013), which is expected to occur in only ~1% of background sequences (West et al., 2012). Thus, we expect this sieve to result in a set of sequences highly enriched for VRC01-class antibodies. We selected light chain NGS reads with a 15 nucleotide CDR L3 and no stop codons as potential VRC01-like light chain sequencing reads.

Construction of light chain variable region (VL) groups. We clustered light chain sequences to identify large groups of sequences that are likely to be biologically relevant; however, because CDR L3 was short and functionally constrained, clustering was performed using the entire light chain variable region (Figure S2B, bottom). Light chain reads from all time points were assembled into a single FASTA-format file together with the light chain sequences of the 39 probe-identified antibodies and clustered using BLASTClust with thresholds of 90% coverage and 96% identity. These groups were then sorted based on size and those with at least 75 raw reads were selected for further analysis.

Selection of representative reads for expression and neutralization assays. Two reads from each CDR H3 and VL group were selected as follows. First, using BLASTClust, each group was clustered into sub-groups at 99% sequence identity, and a read within the largest group was selected as the most over-represented read. Second, a single read from each sub-group was extracted and the consensus of these was calculated from the ClustalW2 alignment. Then the representative read from each sub-group was scored based on its similarity to the consensus sequence, and the one with the highest score was selected as closest-to-consensus. If the two methods selected the same read, then the second-highest scoring read was selected as closest-to-consensus. This procedure takes into account the fact that many transcripts are likely to contain PCR and sequencing error and provides results that are expected to be

biologically reliable. The chosen sequences were reconstituted in IgG expression vectors with the complementary heavy or light chain from VRC01 or VRC03, and the chimeric antibodies were assessed for HIV-1 neutralization on a 10-virus panel.

For heavy chains, representative sequences from 19 of the CDR H3 groups neutralized HIV-1 when reconstituted with VRC01 or VRC03 light chain (Figure 3A and Table S3). Six of these groups included at least one probe-identified antibody, while the other 13 (including the most populous group, H.A, with 34,600 raw/158 curated sequences) were novel, containing CDR H3 regions more than 10% divergent from any of the 39 probe-identified antibodies. Overall, pairing with VRC01 light chain gave greater functionality than the VRC03 pairing, even within the 03+06 clade. While the CDR H3 groups containing known antibodies generally showed the highest potency and breadth, antibodies reconstituted from groups H.E, H.H, and H.S neutralized all 10 isolates (Figure 3A). All but two of the 19 CDR H3 groups had shared the same origin J gene, suggesting they were members of the VRC01 lineage. However, two weakly neutralizing groups (H.I and H.N) had non-matching JH germline genes from J2 and J4, respectively, despite features indicating them to be VRC01-class members.

For light chains, 18 VL groups neutralized HIV-1 when reconstituted with VRC01 or VRC03 heavy chain (Figure 3B and Table S3). In general, the most strongly neutralizing VL region groups showed greater functionality when paired with VRC01 heavy chain than with VRC03 (Figure 3B). Three of the groups included known antibodies, while the remaining 15 were novel. One group, L.C, contained a mix of unrelated antibodies from several different Vκ and Jκ genes, but the synthesized “representative” was nonetheless modestly neutralizing. Interestingly, three groups (L.F, L.K, and L.P) do not conform to the expected VRC01 class signature, despite having Vκ and Jκ genes that matched that of the VRC01 lineage. These groups had either Arg or Glu at the third position of CDR L3 instead of a hydrophobic residue, indicating a less stringent VRC01 signature for the light chain CDR L3 than previously proposed (West et al., 2012; Zhou et al., 2013).

From the 11 neutralizing CDR H3 groups with matching VH and JH gene assignments that did not contain probe-isolated antibodies, 7 were members of clade 03+06 or clade 08; the remaining 3 groups defined 3 additional clades (Figure S4), labeled H3, H4, and H5 based on total number of curated sequences (Figure 3A, center

column; Figure 4A). For light chains, 10 of the 14 new neutralizing VL region groups which did not include probe-identified antibodies are nonetheless members of the 03+06 and 08 clades; the remaining 4 neutralizing VL groups defined 2 new light chain clades, L3 and L5 (Figure 3B, center column; Figure 4B).

Curation of high-quality sequences for analyzing the VRC01 lineage. Beginning with the raw sequences from the neutralizing CDR H3 and VL region groups described above, several filters were applied to ensure the quality and biological relevance of sequences used to analyze the lineage. Importantly, every sequence is thus closely related (≥90% CDR H3 identity or ≥96% VL region identity) to a functionally verified antibody. This resulted in 124,814 heavy chain and 28,500 light chain raw transcripts across all time points. These are expected to have an RMS miscall error of 1.38%, as we have previously reported, though individual sequences may contain as much as 25% error (Zhu et al., 2012).

From these groups, we removed all reads for which the J gene could not be assigned, the junction was out of frame, or a stop codon occurred anywhere in the variable region. The remaining 73,254 heavy chain and 19,987 light chain reads were then clustered across all time points at 97.25% sequence identity. This causes sequences with up to twice the RMS miscall error to be collapsed on to the true biological parent sequence, resulting in 1,862 unique heavy chain and 533 unique light chain sequences. These sequences were used in certain cases for comparison to bulk data (e.g. temporal prevalence in Figure 4C). Finally, heavy chain sequences assigned to V genes other than VH1-2 were removed and all sequences were manually inspected and those containing potential PCR chimeras, likely homopolymer indels, or outlier sequences were removed, leaving 1,041 heavy chain and 492 light chain sequences across all time points for analysis of the lineage. Of these, 162 curated heavy chain sequences and 119 curated light chains were identified at two or more times. In addition, we separately processed two heavy chain sequencing replicates from the 1995 time point (2 half chips which had been pooled prior to the main analysis, see Figure 2). When considered at the 100% identity level, overlap between the two replicates was poor (9.5% of unique sequences were observed in both datasets). This improved

considerably when using the 97.25% threshold to define uniqueness (21.8% overlap), and was still further enhanced by our other quality control procedures (37.7% overlap for curated sequences).

A third tier of sequence quality comprises the synthesized representative sequences. As described in the previous section, these representatives were the most over-represented and the closest-to-consensus in each group. Previous work has shown taking consensus sequences of small groups can produce biologically reliable output (Zhu et al., 2012). In addition, we subjected 454 data from a plasmid control sample containing 10 VRC01-class heavy chain sequences (Zhu et al., 2012) to the quality control procedures used here. The sequences chosen for hypothetical synthesis from that data had an RMS miscall error of 1.35%, and none were greater than 3% divergent from the original inputs. Two of the ten input sequences were recovered exactly, and 5 additional input sequences were recovered at >99% identity. Furthermore, in previous work, unrelated sequences with high predicted structural compatibility with VRC01, VRC03, and VRC-PG04 failed to neutralize HIV-1 in 10 out of 10 cases, even when paired with the native VRC01 light chain (Wu et al., 2011). Thus the successful neutralization of HIV-1 by 63 out of 72 expressed sequences (Figure 3 and Table S3A,B) provides further evidence that we are successfully accounting for error.

Average divergence and temporal prevalence. To estimate the fraction of B cells belonging to the VRC01 lineage at each time point, we compared the number of sequences in each clade to the total number at each time point. Because it was not possible to manually curate all of the NGS sequences obtained, prevalence was calculated as the number of unique sequences with in-frame junctions and no stop codons (but without manual curation) in a particular clade divided by the total number of equivalent bulk reads at that time point. Average divergence for each clade and time point was calculated from fully curated sequences, as described above.

Heavy and light chain “birthday” phylogenetic trees. Curated sequences from all clades and time points were combined; in cases where a sequence appeared at multiple time points, a read from the earliest time point (the apparent “birthday”) was

selected as the representative. DNAML was then used to construct maximum likelihood trees from these representatives. Trees were visualizes as radial phylograms in Dendroscope (Huson and Scornavacca, 2012) and colored based on the source time point of each leaf (Figure 6B).

Analysis of J chain. The JH regions of all 39 probe-isolated antibodies were extracted and aligned to the JH1 germline gene (Figure S1). The substitutions observed in this alignment were used to calculate an empirical substitution matrix and to estimate the Karlin-Altschul parameters K and λ (Karlin and Altschul, 1990). These values were then used to calculate a P-value for each NGS read to be assigned to each of the six possible human JH germline genes. The same procedure was used for light chain Jκ genes, as well.

Binding studies using biolayer interferometry. A fortéBio Octet Red384 instrument was used to measure binding kinetics of HIV-1 gp120 extended core and germline-reverted antibodies. All the assays were performed with agitation set to 1,000 rpm in phosphate-buffered saline (PBS) buffer supplemented with 1% bovine serum albumin (BSA) in order to minimize nonspecific interactions. The final volume for all the solutions was 100 µl/well. Assays were performed at 30℃ in solid blac k 96-well plates (Geiger Bio-One). Germline-reverted antibodies (40 µg/ml) in PBS buffer was used to load anti-human IgG probes for 300 s. Typical capture levels were between 0.7 and 1 nm, and variability within a row of eight tips did not exceed 0.1 nm. Biosensor tips were then equilibrated for 300 s in PBS/1% BSA buffer prior to binding measurements of the HIV-1 gp120 molecules in solution (10 to 0.25 µM) for 300 s; binding was then allowed to dissociate for 120 s. Dissociation wells were used only once to prevent contamination. Parallel correction to subtract systematic baseline drift was carried out by subtracting the measurements recorded for a sensor loaded with germline-reverted antibodies incubated in PBS/1% BSA. To remove nonspecific binding responses, the Influenza antibody CR6261 was loaded onto the anti-human IgG probes and incubated with HIV-1 gp120 molecule, and the nonspecific responses were subtracted from the RUA antibody: HIV-1 gp120 response data. Data analysis and curve fitting were carried out using Octet

software, version 7.0. Experimental data were fitted with the binding equations describing a 1:1 interaction. Global analyses of the complete data sets assuming binding was reversible (full dissociation) were carried out using nonlinear least-squares fitting allowing a single set of binding parameters to be obtained simultaneously for all concentrations used in each experiment.

Evolutionary rate estimation for antibody heavy and light chains. 5-30 NGS reads were selected from each relevant group at each time point. We chose representative sequences for each group by clustering at a threshold chosen to result in between 5 and 30 clusters, and choosing one sequence from each cluster using CD-HIT (Fu et al., 2012; Li and Godzik, 2006). An equal number of representative sequences was chosen for each time point, and each dataset was constrained to have fewer than 300 total sequences. Each set of representative sequences was aligned using Muscle (Edgar, 2004b) with manual adjustment. The GTR+G+I substitution model was selected using MEGA5 (Tamura et al., 2011) as the best overall model for fitting to sequence data for all sequence groups. Five categories were used when modeling rate heterogeneity by Γ distribution. For groups H.I, H.N and L.C, the HKY+Γ model was used. Evolutionary rates were then estimated using Beast v1.8 (Drummond and Rambaut, 2007) with the selected models.

To examine which molecular clock model (restricted or relaxed) fit the data better, we first ran a test simulation for each dataset with a lognormal relaxed molecular clock model and a coalescent Bayesian skyline plot as the tree prior. We then analyzed the posterior distribution of the coefficient of variation (COV) statistic, which reflects the degree of clock-likeness of each dataset (Drummond and Rambaut, 2007). If the distribution of COV did not include 0, the lognormal relaxed model was used for later analysis. Otherwise, the restricted molecular clock model was used.

We then tested two coalescent models for the phylogenetic tree prior: either a constant population size or a Bayesian skyline plot (Drummond et al., 2005). The optimal coalescent model was chosen based on the Bayes factor (BF), with $2\ln(BF)>3$ used as the criterion for determining that one model was better than the other (Drummond and Rambaut, 2007; Suchard et al., 2001).

We then estimated the evolutionary rate for each dataset using the optimal molecular clock and coalescent models. All simulations were run for more than $5x10^7$ Bayesian Markov chain Monte Carlo steps. The simulation was terminated when it had converged and the effective sample size was larger than 200 (http://beast.bio.ed.ac.uk/). The first 10% of simulation steps were discarded as burn-in, and simulation results were analyzed using Tracer v1.5 (Drummond and Rambaut, 2007).

HIV Env gene evolutionary rate. Env sequences from donor 45 and CH505 were retrieved from Genbank (Liao et al., 2013; Wu et al., 2012). Sequences having >50 nt deletions or a frameshift were removed. Potential recombination among Env sequences was assessed using RDP (Martin et al., 2010). Sequences having recombination signals with p value<0.05 were removed. For both Env gene datasets analyzed, the HKY+I substitution model was selected using MEGA5. Selection of optimal models, simulations, and analysis were performed as described above for antibody datasets.

Construction, expression and purification of HIV-1 extended core gp120 proteins. The extended core (coreE) gp120 proteins from different HIV-1 stains were constructed as reported previously (Diskin et al., 2011; Wu et al., 2011; Zhou et al., 2010). The gp120 proteins were produced by transient transfection in $GNTI^-$ cells as described previously (Zhou et al., 2010). Culture supernatants were harvested 5 - 6 days after transfection, filtered through a 0.45 µm filter and then passed through an antibody 17b-affinity column. After extensive washes with 1xPBS, the gp120 proteins were eluted with IgG Elution buffer (Pierce). The eluate fractions containing gp120 protein were brought to pH 7.4 by 1 M Tris/Cl^-, pH 8.5, concentrated, flash-frozen in liquid nitrogen and stored at -80℃ until use.

Crystallization of gp120:VRC01-class antibody complexes. Complexes of gp120 and Fab of donor 45 antibodies were formed by mixing deglycosylated gp120 and antibody Fabs at 1:1.2 molar ratio at room temperature and purified by size exclusion chromatography (Hiload 26/60 Superdex S200 prep grade, GE Healthcare) with buffer containing 0.35 M NaCl, 2.5 mM Tris-HCl pH 7.4, 0.02% NaN_3. Fractions with gp120:antibody complexes were concentrated to appropriate concentrations, flash frozen with liquid nitrogen and stored at -80^{o}C before crystallization screening experiments.

Initial crystallizations were carried out using a mosquito® Crystal robot (TTP Labtech, UK) and commercially available Hampton (Hampton Research), Precipitant Synergy (Emerald Biosystems) and Wizard (Emerald Biosystems) crystallization screens. Droplets were allowed to equilibrate at 20º C and imaged at scheduled times with RockImager (Formulatrix.). Robotic crystal hits were optimized manually using hanging drop vapor-diffusion method. For the clade A/E 93TH057 gp120: 45-VRC01.H03+06.D-001739 complex, crystals of diffraction quality were obtained in 10% PEG 8000, 0.1M Tris-HCl, pH 8.5. For the 93TH057 gp120: 45-VRC01.H08.F-117225 complex, crystals of diffraction quality were obtained in 10% PEG 4000, 0.2M NaAcetate, 0.1M Tris-HCl, pH 8.5. For the 93TH057 gp120: 45-VRC01.H5.F-185917 complex, crystals of diffraction quality were obtained in 12% PEG 8000, 5% isopropanol, 0.1M HEPES, pH 7.5. For the 93TH057 gp120:VRC6b complex, crystals of diffraction quality were obtained in 7% PEG 3350, 2.5% isopropanol, 0.1M Tris-HCl, pH 8.5. For the clade A Q842.d12 gp120:VRC08 complex, crystals of diffraction quality were obtained in 10.5% PEG 1500, 1% MPD, 0.1M Tris-HCl 8.5. For the Q842.d12 gp120:VRC08c complex, crystals of diffraction quality were obtained in 13% PEG 1500, 2% MPD, 0.1M Tris-HCl, pH 8.5 for Q842.d12 gp120:VRC08c complex. For the Q842.d12 gp120:VRC08c complex, crystals of diffraction quality were obtained in 13% PEG 1500, 2% MPD, 0.1M Tris-HCl, pH 8.5 for Q842.d12 gp120:VRC08c complex. For the d45-01G5 gp120: 45-VRC01.H01+07.O-863513/45-VRC01.L01+07.O-110653 complex, crystals of diffraction quality were obtained in 0.1 M Na Cacodylate pH 6.5, 0.2 M NaAcetate, 6 % PEG 8000. For the 93TH057 gp120:45-VRC01.H01+07.O-

863513/45-VRC01.L01+07.O-110653 complex, crystals of diffraction quality were obtained in 0.1 M Na Cacodylate pH 6.5, 12% PEG 8000, 0.16 M NaAcetate.

X-ray data collection, structure determination and refinement for the gp120:VRC01-class antibody complexes. X-ray diffraction data were collected at the synchrotron beam line SER-CAT BM22 and ID22 (Advanced Photon Source, Argonne National Laboratory) using 1.0000 Å radiation under cryogenic conditions. Optimal cryo-protectant conditions were screened as previously described (Zhou et al., 2010). Supplemented with respective reservoir solutions, optimal cryo-protectants were 30% ethylene glycol the 93TH057 gp120: 45-VRC01.H03+06.D-001739 complex, 15% ethylene glycol, 15%glycerol, and 7.5% 2R,3R-butanediol for both 93TH057 gp120: 45-VRC01.H08.F-117225 and 93TH057 gp120: 45-VRC01.H5.F-185917 complexes, 40% glycerol for the 93TH057 gp120:VRC06b complex, 15% 2R,3R-butanediol for both Q842.d12 gp120:VRC08 and Q842.d12 gp120:VRC08c complexes, and 12% 2R,3R-butanediol for the d45-01dG5 gp120:45-VRC01.H01+07.O-863513/45-VRC01.L01+07.O-110653 and 93TH057 gp120:45-VRC01.H01+07.O-863513/45-VRC01.L01+07.O-110653 complexes.

Data were processed using HKL2000 (Otwinowski and Minor, 1997) and structures were determined by molecular replacement using Phaser (McCoy et al., 2007) with VRC06-bound gp120 (PDB ID:4JB9) as the search model for VRC06b-, VRC08-, or VRC08c-bound gp120 structures, VRC03-bound gp120 (PDB ID: 3SE8) as the search model for 45-VRC01.H03+06.D-001739 -bound gp120 complex, VRC01-bound gp120 structure (PDB ID:3NGB) as the search model for 45-VRC01.H08.F-117225 - or 45-VRC01.H5.F-185917-bound gp120 complexes.

Refinements were carried out with PHENIX (Adams et al., 2002). Starting with torsion-angle simulated annealing with slow cooling, iterative manual model building was carried out on COOT (Emsley and Cowtan, 2004) with maps generated from combinations of standard positional, individual *B*-factor, TLS refinement algorithms. Ordered solvents were added during each macro cycle. Throughout the refinement processes, a cross validation (R_{free}) test set consisting of 5% of the data was used and hydrogens were included as riding model. Structure validations were performed

periodically during the model building/refinement process with MolProbity (Davis et al., 2007). Final refinement statistics are summarized in Table S5.

Neutralization fingerprinting. Antibody neutralization fingerprinting analysis was performed as described previously (Georgiev et al., 2013). Briefly, the relative potency pattern with which a given antibody neutralizes a set of diverse HIV-1 strains was defined as the antibody neutralization fingerprint. Neutralization fingerprints are epitope-specific, with similar fingerprints for antibodies targeting similar epitopes, and vice versa (Georgiev et al., 2013). Fingerprint-based hierarchical clustering of the antibodies was then used to define epitope-specific clusters of antibodies. A panel of 34 diverse HIV-1 strains was used in the analysis (strain, clade): (6101.10, B), (7165.18, B), (57128.vrc15, D), (Bal.01, B), (BG1168.01, B), (BL01.DG, B), (CAAN.A2, B), (CAP210.E8, C), (DU156.12, C), (DU172.17, C), (DU422.01, C), (JRCSF.JB, B), (JRFL.JB, B), (KER2008.12, A), (KER2018.11, A), (PVO.04, B), (Q168.a2, AD), (Q23.17, A), (Q769.h5, A), (Q842.d12, A), (RW020.2, A), (SO18.18, C), (THRO.18, B), (TRJO.58, B), (TRO.11, B), (TV1.29, C), (TZA125.17, C), (UG037.8, A), (YU2.DG, B), (ZA012.29, C), (ZM106.9, C), (ZM109.4, C), (ZM176.66, C), (ZM55.28a, C).

Figure S1. Immunoglobulin gene analysis of 39 probe-identified VRC01-lineage antibodies, Related to Figure 1.
CDR H3 lengths, and VH and Vκ mutations for the 39 probe identified antibodies (left columns). All were assigned to VH1-2*02, JH1*01, Vκ3-20*01, and Jκ2*01. The JH alignments are also shown (right columns). Red line delineates the assigned recombination point. Positions that differ between JH1*01 and JH2*01 are highlighted. JH1 is the best assignment for all probe-isolated antibodies, including VRC03. The 03+06 clade appears to have a deletion within the J region, but this would not change the assignment. The alignments shown here were used to construct a substitution matrix which was in turn used to calculate P-values for the germline J assignment of each 454 sequence (see Extended Experimental Procedures).

Figure S2. Identification of VRC01-class sequences for heavy and light chain from NGS-derived B cell transcripts, Related to Figure 2.
(A) Antibody-heavy chain pipeline for variable region analyses. Raw 454 sequences (left) were filtered for read length. Germ line V gene was assigned to each sequence and to construct identity-divergence (I-D) plots (middle, top). Cross-donor phylogenetic analysis (middle, bottom) was performed by splitting sequences into multiple FASTA files and building neighbor-joining trees with ClustalW2, rooted on IGHV1-2*02. In each tree, the smallest sub-tree containing all of the exogenously added known VRC01-class antibodies was extracted and the NGS-reads in this sub-tree were passed to the next iteration of neighbor-joining trees until convergence was reached. Converged sequences were defined as being cross-donor positive. BLASTClust (far right) was used to cluster CDR H3 from cross-donor positive sequences. For the most populated CDR H3 groups, one over-represented sequence and the sequence closest to consensus were synthesized, reconstituted with VRC01 or VRC03 light chain, and tested for neutralization (Fig. 3A).
(B) Antibody-light chain pipeline for variable region analyses. Raw 454 sequences (left) were filtered for read length. Germ line V gene was assigned to each sequence and to construct identity-divergence (I-D) plots (middle, top). A 5-amino acid CDR L3 signature (middle, bottom) was used to identify potential VRC01-class sequences. BLASTClust (far right) was used to cluster the entire light chain-variable region from sequences with the CDR L3 signature. For the most populated groups, one over-represented sequence and the sequence closest to consensus were synthesized, reconstituted with VRC01 or VRC03 heavy chain, and tested for neutralization (Fig. 3B).
(C)-(F) Clade-specific identity-divergence plots from longitudinal samples for heavy (C, E) and light (D,F) chains from clade 01+07 (C, D) and clade 03+06 (E, F). Sequence divergence from the assigned germ line V gene (x axis) and sequence identity to the antibody variable domain indicated (y axis). The top row shows a heat map for positions of all 454 sequences. The total number of sequences is indicated at the right borders. The middle row shows the distribution of cross-donor positive sequences, with blue dots indicating cross-donor positives (C, E) or sequences with a 5 aa CDR L3 (D, F) and blue numbers indicating the number of such sequences. Gray contours indicate raw sequences. The bottom row shows the distribution of sequences in the same CDR H3 group (C, E) or VL group (D, F) as any probe-isolated antibody from that clade as yellow dots, with purple dots indicating the subset of those which survived quality filtering.

Yellow and purple numbers, respectively, display the number of each type of sequence in the relevant groups. Blue dots from the middle row are shown here as contours for comparison.

Figure S3. Curation of VRC01-lineage related sequences and comparison to non-lineage sequences, Related to Figure 3.
(A) Identity-divergence plots of accepted groups and curated sequences from all ten time points. Yellow dots represent all sequences in the accepted neutralizing groups shown in Figure 3. Purple dots represent those sequences which survived multiple quality filters (see Extended Experimental Procedures) to be included in the final set of curated sequences. For comparison, all 39 probe-identified antibodies (Figure 1) are shown as black squares and the synthesized representatives (Figure 3 and Table S3) are shown as red diamonds. Synthesized neutralizers from excluded groups H.I, H.N, and L.C are shown as green stars.
(B) Analysis of the presence of the Cys98/99 signature in VRC01-lineage member sequences and non-member sequences. Cys98 or Cys99 is present in almost all VRC01-lineage member sequences (99.4%) as compared to non-members (3.7%) ($P<0.0001$), suggesting that there is a high likelihood that the VRC01-lineage member sequences from the diverse clades are indeed derived from a single lineage, the VRC01 lineage. Curated VRC01-lineage sequences are compared to unique background transcripts with assigned V and J genes, continuous ORF, and in-frame junctions which appear more than once.

Figure S4. Hierarchical classification of the VRC01 lineage, Related to Figure 4.
A phylogenetic tree of all curated heavy chain NGS sequences from the VRC01 lineage demonstrates the hierarchical system used to classify sequences. This is the same tree as shown in Figure 6B (left panel), but in a rectangular layout and with sequences colored by clade. All lineage sequences (colored tips, right side) evolved from the single unmutated common ancestor (UCA) B cell (root of tree, left side). Clades are major sub-lineage units that are defined by the geometry of phylogenetic tree (boxes, labels indicate the branch point at which the clade diverges from others). They represent evolutionarily close groups of sequences that are evolutionarily distant from other clades. We have defined 6 heavy chain clades in the VRC01 lineage (Figure 4C). As a smaller, sub-lineage and sub-clade unit of relationship, we have defined CDR H3 groups (ovals) for heavy chains and VL groups for light chains (not shown). These are defined by clustering at specific sequence identity levels (Figure S2A,B), so that all sequences within a CDR H3 group have at least 90% CDR H3 sequence identity to each other, but less than 90% CDR H3 sequence identity to reads in any other group, even within the same clade. Because these two sub-lineage levels (clade and group) are defined by different criteria, some clades only contain a single group, while others contain multiple groups. In addition, some groups may contain a larger number of sequences than an entire other clade.

Figure S5. Conservation of VRC01+07 lineage structure and recognition, Related to Figure 5.

(A) Structure of VRC01-lineage antibodies from clade 01+07. Left, 45-VRC01.H01+07.O-863513/45-VRC01.L01+07.O-110653 (green) and right, NIH45-46 (red) shown in cartoon representation with contact residues shown in stick representation (colored gold).
B) Root mean squared deviation and common contact surface area overlaps of antibody: HIV-1 gp120 structures.
(C) Top, amino acid sequences of germline-reverted VRC01 antibody compared to mature VRC01 and the germline sequences of VH1-2*02 or VK3-20*01. Bottom, binding data generated with an Octet biosensor for germline-reverted VRC01, the 1995 clade 01+07 antibody VRC01.H01+07.O-863513/45-VRC01.L01+07.O-110653, and the 2008 antibody VRC07 for gp120s from 3 different autologous viruses. The germline-reverted VRC01 binds to the early virus 01dG5, but not to the later virus, while the mature antibodies from 1995 and 2008 bind gp120s from all three viruses with sub-nanomolar affinity.

Figure S6. Germline divergence over time and implied dating of common ancestors, related to Figure 6.
(A) Donor 45 VRC01-class lineage for germline VH (left) and germline Vκ (right). Average germline divergence for sequences without manual curation is plotted for each clade of the VRC01 lineage (grey dotted lines), as well as for deposited sequences from the CH103 lineage from Donor CHAVI505 (orange dashed lines) and the VRC26 lineage from donor CAP256 (green solid lines).
(B) Inferred evolutionary time line for the VRC01 lineage heavy (left) and light (right) chains from Donor 45, using curated sequences. Branches are colored by clade and branch length shows estimated divergence time. The estimated date of the most recent common heavy chain ancestor is 1971, with the 95% confidence interval (CI) (black bar) extending from 1955 to 1979. For the light chain, the estimated date of the most recent common ancestor is 1979, with the 95% CI extending from 1970 to 1986. These dates are implausibly early, given the known history of the AIDS epidemic, suggesting that the evolutionary rate of the VRC01 lineage was once faster than observed during the study period.
(C, D) Inferred evolutionary time line for the CAP256-VRC26 lineage (C) and CHAVI505-CH103 lineage (D) heavy and light chains (left and right, respectively). Branch length shows estimated divergence time. The estimated dates of the most recent common ancestors are indicated by the black bars. These dates are before the known date of the original lineage recombination (red dashed line), suggesting that the evolutionary rate was once faster than observed during the study period.

Figure S7. Structural and functional characteristics of selected VRC01-lineage antibodies, Related to Figure 7.
(A) The extended framework region 3s (FR3) of clade 03+06 (cyan) were spatially proximal to the V1V2 stem of gp120.This resulted in different V1V2 stem conformation in VRC03, VRC06, or 45-VRC01.H03+06.D-001739-bound gp120.
(B) Docking of the FR3s to the BG505 SOSIP trimer allows potential contacts between the region with helix-α0 (residue 66), V1V2 (residues 203-207), V3 (residues 304-318), and β21(residues 437-440) of the neighboring gp120 protomer (dark gray).

(C) Potential interactions between the CDR H3 (cyan) of VRC08c and helix-α0 (red) of the neighboring gp120 protomer in trimer. For clarity, only FR3 and CDR H3 regions of the antibody are shown.
(D) Neutralization fingerprinting analysis for VRC01-class antibodies from donor 45. Analysis was performed using data for 34 diverse HIV-1 strains. Soluble CD4 (sCD4) is shown as a control (left panel). Antibodies generally clustered in a similar pattern to the clade pattern observed in the phylogenetic tree of probe-identified heavy chains (right panel), although the differences between clades 01+07 and 08 were within the variation observed for multiple repeats of the neutralization experiment for the same antibody (shown as a numbered superscript after the antibody name). As in the phylogenetic tree, the 03+06 clade clustered farther apart from the other two clades, indicating more substantial functional differences.
(E) A phylogenetic tree of all VRC01-lineage heavy chains showing the relationship among native neutralizing antibodies (blue), functionally validated neutralizing sequences (red), and curated (but unvalidated) sequences (black). Importantly, the untested curated NGS sequences do not change or add to the overall architecture of the lineage, and the maximum interclade CDR H3 sequence diversity does not change significantly. Rather, they serve to fill in details of each clade.

Table S1. Neutralization IC_{50} and IC_{80} titers (μg/ml) of the antibody VRC08 against 195 HIV-1 Env-pesudoviruses, Related to Figure 1.

Virus	Clade	IC_{50}	IC_{80}
0260.v5.c36	A	0.328	0.849
0330.v4.c3	A	0.147	0.346
0439.v5.c1	A	0.779	2.44
3365.v2.c20	A	0.065	0.215
3415.v1.c1	A	0.146	0.469
3718.v3.c11	A	>50	>50
398-F1_F6_20	A	0.150	1.01
BB201.B42	A	0.340	0.757
BB539.2B13	A	0.072	0.172
BI369.9A	A	0.096	0.299
BS208.B1	A	0.027	0.100
KER2008.12	A	0.680	2.60
KER2018.11	A	1.29	3.78
KNH1209.18	A	0.096	0.340
MB201.A1	A	0.280	0.625
MB539.2B7	A	0.162	0.478
MI369.A5	A	0.149	0.526
MS208.A1	A	0.479	2.35
Q23.17	A	0.085	0.260
Q259.17	A	0.047	0.129
Q769.d22	A	0.044	0.125
Q842.d12	A	0.033	0.076
QH209.14M.A2	A	0.045	0.116
RW020.2	A	0.289	0.760
UG037.8	A	0.042	0.131
3301.V1.C24	AC	0.062	0.158
3589.V1.C4	AC	0.093	0.295
6540.v4.c1	AC	>50	>50
6545.V4.C1	AC	>50	>50
0815.V3.C3	ACD	0.071	0.199
6095.V1.C10	ACD	0.422	2.80
3468.V1.C12	AD	1.84	>50
Q168.a2	AD	0.119	0.400
Q461.e2	AD	0.351	0.823
620345.c1	AE	>50	>50
BJOX009000.02.4	AE	0.743	2.07
BJOX010000.06.2	AE	0.622	2.62
BJOX025000.01.1	AE	0.114	0.275
BJOX028000.10.3	AE	0.021	0.091
C1080.c3	AE	0.243	1.30
C2101.c1	AE	0.108	0.352
C3347.c11	AE	0.024	0.080
C4118.09	AE	0.057	0.153
CM244.ec1	AE	0.017	0.059
CNE3	AE	0.115	0.328
CNE5	AE	0.292	0.930
CNE55	AE	0.102	0.296
CNE56	AE	0.194	0.841
CNE59	AE	0.123	0.536
CNE8	AE	0.155	0.397
M02138	AE	0.335	1.02
R1166.c1	AE	0.826	1.73
R2184.c4	AE	0.071	0.175
R3265.c6	AE	0.085	0.274
TH023.6	AE	0.024	0.087
TH966.8	AE	0.050	0.172
TH976.17	AE	0.114	0.314
235-47	AG	0.061	0.256
242-14	AG	>50	>50
263-8	AG	0.103	0.412
269-12	AG	0.516	1.42
271-11	AG	0.071	0.138
928-28	AG	0.164	0.675
DJ263.8	AG	0.014	0.068
T250-4	AG	>50	>50
T251-18	AG	1.31	4.44
T253-11	AG	0.577	2.27
T255-34	AG	22.5	>50
T257-31	AG	1.25	4.99
T266-60	AG	0.596	2.15
T278-50	AG	>50	>50
T280-5	AG	0.027	0.081
T33-7	AG	0.020	0.045
3988.25	B	0.130	0.314
5768.04	B	0.159	0.548
6101.10	B	0.020	0.060
6535.3	B	0.088	0.328
7165.18	B	13.6	41.8
45_01dG5	B	0.004	0.012
89.6	B	0.094	0.335
AC10.29	B	0.570	2.13
ADA	B	0.037	0.215
BaL.01	B	0.031	0.081
BaL.26	B	0.012	0.032
BG1168.01	B	0.175	0.371
BL01	B	>50	>50
BR07	B	0.127	0.450
BX08.16	B	0.079	0.216
CAAN.A2	B	0.333	1.13
CNE10	B	0.099	0.295
CNE12	B	0.077	0.245
CNE14	B	0.035	0.112
CNE4	B	0.153	0.592
CNE57	B	0.074	0.202
HO86.8	B	>50	>50
HT593.1	B	0.052	0.227
HXB2	B	0.012	0.035
JR-CSF	B	0.105	0.302
JR-FL	B	0.006	0.018
MN.3	B	0.012	0.047
PVO.04	B	0.041	0.124
QH0515.01	B	0.453	1.64
QH0692.42	B	0.350	1.16
REJO.67	B	0.050	0.137
RHPA.7	B	0.065	0.176
SC422.8	B	0.057	0.150
SF162	B	0.033	0.097
SS1196.01	B	0.063	0.181
THRO.18	B	1.13	7.85
TRJO.58	B	0.149	0.373
TRO.11	B	0.088	0.267
WITO.33	B	0.060	0.178
Yu2	B	0.034	0.080
CH038.12	BC	0.242	0.651
CH070.1	BC	>50	>50
CH117.4	BC	0.037	0.116
CH181.12	BC	0.050	0.154
CNE15	BC	0.078	0.324
CNE19	BC	0.023	0.076
CNE20	BC	0.040	0.098
CNE21	BC	0.073	0.209
CNE40	BC	0.032	0.130
CNE7	BC	0.040	0.125
286.36	C	0.271	0.744
288.38	C	0.166	0.530
0013095-2.11	C	0.015	0.078
001428-2.42	C	0.009	0.022
0077_V1.C16	C	>50	>50
00836-2.5	C	0.015	0.037
0921.V2.C14	C	0.152	0.401
16055-2.3	C	0.041	0.111
16845-2.22	C	0.352	2.33
16936-2.21	C	0.078	0.207
25710-2.43	C	0.073	0.197
25711-2.4	C	0.185	0.501
25925-2.22	C	0.058	0.169
26191-2.48	C	0.132	0.352
3168.V4.C10	C	0.218	0.582
3637.V5.C3	C	0.091	0.289
3873.V1.C24	C	1.15	4.83
6322.V4.C1	C	>50	>50
6471.V1.C16	C	>50	>50
6631.V3.C10	C	0.964	5.59
6644.V2.C33	C	0.012	0.054
6785.V5.C14	C	0.112	0.344
6838.V1.C35	C	0.148	0.540
96ZM651.02	C	0.120	0.331
BR025.9	C	5.59	43.9
CAP210.E8	C	>50	>50
CAP244.D3	C	0.186	0.566
CAP45.G3	C	0.190	0.550
CNE30	C	0.237	0.662
CNE31	C	0.142	0.510
CNE53	C	0.051	0.182
CNE58	C	0.038	0.108
Du123.06	C	0.417	1.590
Du151.02	C	0.169	0.529
Du156.12	C	0.062	0.179
Du172.17	C	>50	>50
Du422.01	C	16.8	>50
MW965.26	C	0.070	0.427
SO18.18	C	0.022	0.081
TV1.29	C	>50	>50
TZA125.17	C	>50	>50
TZBD.02	C	0.010	0.030
ZA012.29	C	0.080	0.199
ZM106.9	C	0.058	0.136
ZM109.4	C	0.069	0.216
ZM135.10a	C	0.127	0.509
ZM176.66	C	0.400	>50
ZM197.7	C	0.207	0.792
ZM214.15	C	0.397	1.27
ZM215.8	C	0.050	0.177
ZM249.1	C	0.081	0.281
ZM53.12	C	0.211	1.01
ZM55.28a	C	0.048	0.143
3326.V4.C3	CD	>50	>50
3337.V2.C6	CD	0.048	0.122
3817.v2.c59	CD	>50	>50
191821.E6.1	D	0.151	0.601
231965.c1	D	17.7	>50
247-23	D	0.066	0.215
3016.v5.c45	D	0.231	0.821
57128.vrc15	D	>50	>50
6405.v4.c34	D	0.485	1.22
A03349M1.vrc4a	D	1.24	3.79
A07412M1.vrc12	D	0.091	0.351
NKU3006.ec1	D	0.587	1.57
UG024.2	D	0.604	2.75
P0402.c2.11	G	0.068	0.253
P1981.C5.3	G	0.117	0.368
X1193.c1	G	0.063	0.178
X1254.c3	G	0.249	0.715
X1632.S2.B10	G	2.88	>50
X2088.c9	G	>50	>50

Table S2. Clinical and sequencing parameters of donor 45 over 15 years of HIV-1 infection, Related to Figure 2.

A

Blood drawn date	3/20/1995	7/30/2001	4/16/2002	7/14/2006	1/9/2007	7/12/2007	1/17/2008	8/19/2008	6/2/2009	12/31/2009
CD4 count (cells/µl)	767	656	791	638	601	647	629	530	686	583
Plasma viral load (copies/ml)	11,000	9,858	9,607	9,129	17,796	12,709	16,920	8,588	5,153	5,158

B

	3/1995	**7/2001**	**4/2002**	**7/2006**	**1/2007**	**7/2007**	**1/2008**	**8/2008**	**6/2009**	**12/2009**	**TOTAL**
NGS heading	O	A	B	C	D	E	F	G	H	I	
raw heavy	932,697	304,968	280,114	529,389	319,684	377,019	220,926	476,045	267,953	506,209	4,215,004
total heavy with length and V	661,038	304,944	280,088	529,152	319,671	376,939	214,951	475,999	267,902	506,154	3,936,838
unique* heavy with length and V	*107,736*	*86,136*	*111,661*	*38,005*	*93,590*	*104,091*	*67,434*	*108,459*	*79,417*	*44,761*	*841,290*
total cross-donor positives (XD+)	56,855	7,494	16,654	43,652	7,322	17,555	21,143	11,946	27,797	49,835	260,253
unique* XD+	*5,760*	*766*	*4,210*	*1,847*	*2,175*	*4,537*	*5,013*	*1,800*	*4,534*	*1,813*	*32,455*
total in accepted groups ("AG")	6,875	4,645	8,085	26,545	2,757	7,744	8,778	9,204	16,547	33,654	124,834
AG with J, in frame, and ORF ("ORF")	4,842	3,005	2,931	17,194	1,645	3,153	4,297	4,898	8,519	23,040	73,524
ORF unique*	147	150	135	101	106	147	328	206	360	206	1,886
final curated	91	92	68	54	70	73	188	114	193	98	1,041
raw light	189,850	377,954	456,794	685,467	364,358	509,461	463,929	644,062	363,276	361,162	4,416,313
total light with length and V	189,615	377,919	455,930	685,405	364,340	509,441	463,862	641,637	362,433	361,134	4,411,716
unique* light with length and V	*78,402*	*52,043*	*95,404*	*23,380*	*48,211*	*64,796*	*70,487*	*284,903*	*55,439*	*16,023*	*789,088*
total 5aa CDR L3	6,363	561	4,329	8,279	1,113	8,853	4,947	8,323	5,155	6,475	54,398
unique* 5 AA CDR L3	*2,145*	*275*	*1,067*	*324*	*329*	*788*	*710*	*2,987*	*875*	*310*	*9,810*
total in accepted groups ("AG")	177	208	1,322	7,814	640	3,994	3,072	3,136	3,245	4,892	28,500
AG with J, in frame, and ORF ("ORF")	108	147	738	5,761	273	2,903	2,335	1,797	2,139	3,786	19,987
ORF unique*	11	32	42	55	22	70	101	87	75	38	533
final curated	10	31	39	53	22	66	92	79	68	32	492

* Uniqueness was defined by clustering at 97.25% identity to account for sequencing error. The pipeline used for this work only takes account of uniqueness in the penultimate filtering step, for open-reading frames among the accepted groups. Unique counts after V assignment, cross-donor analysis, and filtering on CDR L3 length (italics) are shown for reference only.

Table S3. Additional synthesized sequences, Related to Figure 3. Representative sequences from the most prevalent CDR H3 (A) or VL (B) groups were synthesized, reconstituted with VRC01 and VRC03 light or heavy chains, and tested for neutralization. Only one representative from neutralizing groups are shown in Figure 3; the second representative and representatives from non-neutralizing groups are shown here. Total sequences in each group, probe-identified representative (if any), assigned V and J genes, and the most neutralizing representative and its CDR H3 or CDR L1 sequence (left columns) are shown. Neutralization breadth and potency for both VRC01 and VRC03 light or heavy chain pairings are provided against selected HIV-1 viruses from clades A, B, and C (right columns). Total number of sequences at each time point and the number of those also observed in at least one other time point for heavy (C) and light (D) chain sequences.

ATTACHED AS AN EXCEL FILE

Table S4. Crystallographic data collection and refinement statistics, Related to Figure 5.

	d45-01dG5: 45-VRC01.H01+ 07.O-863513 / 45-VRC01.L01+07.O-110653	93TH057: 45-VRC01.H01+ 07.O-863513 / 45-VRC01.L01+07.O-110653	93TH057: 45-VRC01.H03+06.D-001739	93TH057: 45-VRC01.H08.F-117225	93TH057: 45-VRC01.H5.F-185917	93TH057: VRC06b	Q842.d12: VRC08	Q842.d12: VRC08c
PDB accession code	4XVS	4XVT	4S1Q	4S1R	4S1S	4XNZ	4XMP	4XNY
Data collection								
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P4_3$
Cell constants								
a, *b*, *c* (Å)	67.6, 69.3, 200.6	64.6, 67.8, 199.2	66.1, 78.9, 194.2	65.6, 68.1, 204.1	55.3, 67.6, 266.7	68.3,189.4, 219.9	69.7, 82.5, 163.1	121.5, 121.5, 68.7
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Wavelength (Å)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Resolution (Å)	50.0-1.90 (1.97-1.90)*	50.0-1.70 (1.76-1.70)	50.00-2.40 (2.44-2.40)	50.0-3.20 (3.26-3.20)	50.00-3.40 (3.46-3.40)*	39.90-3.39 (3.51-3.39)	35.50-1.78 (1.85-1.78)	33.69-2.30 (2.38-2.30)
R_{merge}	15.2 (90.0)	13.7 (99.0)	11.3 (51.5)	15.6 (53.0)	9.7 (52.0)	9.3 (56.6)	8.8 (64.3)	7.8 (55.6)
R_{pim}	6.7 (56.1)	5.3 (53.2)	7.6 (63.4)	7.4 (38.8)	4.8 (35.4)			
I / σI	9.5 (1.1)	13.0 (1.08)	12.4 (1.6)	8.8 (1.3)	14.4 (1.6)	21.4 (1.8)	25.1 (1.9)	16.8 (1.9)
Completeness (%)	92.8 (88.6)	99.9 (99.9)	96.3 (80.4)	93.6 (53.7)	97.2 (82.8)	86.0 (50.0)	97.9 (89.7)	99.1 (94.3)
Redundancy	7.0 (5.3)	9.6 (6.2)	3.7 (1.8)	5.0 (1.7)	4.7 (2.4)	5.8 (3.4)	5.9 (4.2)	3.7 (3.0)
Refinement								
Resolution (Å)	1.90	1.70	2.40	3.20	3.40	3.40	1.78	2.30
Unique reflections	69,542	97,827	36,988	14,552	14,188	34,907	88,131	44,409
R_{work} / R_{free} (%)	19.4/21.4	18.7/22.5	21.9/25.8	26.0/28.6	26.1/29.9	24.2/28.9	16.1/19.3	18.2/21.7
No. atoms								
Protein	6045	6,038	6,048	6,019	6,032	17,958	6,021	6,027
Ligand/ion	729	772	154	154	155	366	166	154
Water	631	632	126	0	0	0	613	206
B-factors (Å^2)								
Protein	37.6	42.9	56.1	102.4	110.5	135.3	45.0	55.9
gp120-only	41.1	43.0	57.5	81.4	105.7			
Fv-only	37.3	41.9	52.2	92.0	90.2			
F_{CH1-CL}-only	31.0	37.5	76.1	150.9	134.6			
Ligand/ion	73.2	73.9	63.5	100.1	127.9	178.30	58.2	70.6
Water	46.4	49.3	48.7	-	-	-	47.4	50.0
R.m.s. deviations								
Bond lengths (Å)	0.007	0.010	0.003	0.004	0.003	0.007	0.009	0.003
Bond angles (°)	1.02	1.19	0.737	0.917	0.879	1.170	1.160	0.833
Ramachandran								
Favored regions (%)	96.5	96.5	95.7	90.5	90.9	90.5	98.0	97.0
Allowed regions (%)	3.4	3.4	4.3	8.6	7.4	7.6	1.9	2.7
Disallowed regions (%)	0.1	0.1	0	0.9	1.7	1.9	0.1	0.3

*Values in parentheses are for highest-resolution shell

Table S5. **Structural comparison of antibody heavy and light chains and binding surfaces, Related to Figure 7.**

(A) Cα-RMSDs (Å) were calculated after aligning framework regions of VRC01 heavy chain (residues 3-26, 36-49, 65-76, 78-94, and 103-113, first row of each cell) and light chain (residues 3-23, 35-49, 56-88, and 98-102, italicized bottom row of each cell) with the corresponding regions of antibodies of donor 45. RMSDs within clades were colored in red, blue, and purple for 01+07 clade, 03+06 clade, and 08 clade, respectively.
(B) Cα-RMSDs (Å) were calculated after aligning framework regions of VRC01 heavy chain (residues 3-26, 36-49, 65-76, 78-94, and 103-113, first row of each cell) and light chain (residues 3-23, 35-49, 56-88, and 98-102, italicized bottom row of each cell) with the corresponding regions of antibodies from donor 45 and other donors. RMSDs within clades were colored in red, blue, and purple for 01+07 clade, 03+06 clade, and 08 clade, respectively. RMSDs within donor 45 antibodies were colored in red. 45-VRC01.H03+06.D-001739, 45-VRC01.H08.F-117225, and 45-VRC01.H5.F-185917 were heavy chain NGS reads that were expressed and crystallized together with the VRC01 light chain.
(C) Binding surfaces on gp120 by donor 45 antibodies. 45-VRC01.H03+06.D-001739, 45-VRC01.H08.F-117225, and 45-VRC01.H5.F-185917 are heavy chain NGS reads that were expressed and crystallized together with the VRC01 light chain.

A	VRC01	NIH45-46	VRC03	VRC06	VRC06b	45-VRC01.H03+06.D-001739	VRC08	VRC08c	45-VRC01.H08.F-117225
NIH45-46	0.459 *0.363*								
VRC03	1.421 *0.641*	1.481 *0.681*							
VRC06	1.481 *0.953*	1.457 *1.017*	0.609 *0.83*						
VRC06b	0.695 *1.653*	0.73 *1.659*	1.067 *1.58*	1.193 *1.239*					
45-VRC01.H03+06.D-001739	2.408 *1.056*	2.364 *1.046*	2.287 *1.26*	2.401 *1.097*	1.871 *0.888*				
VRC08	1.028 *1.041*	0.968 *1.038*	0.943 *0.837*	1.017 *1.074*	1.086 *0.969*	2.579 *1.126*			
VRC08c	0.923 *0.895*	0.877 *0.911*	0.892 *0.692*	1.529 *1.023*	0.98 *0.679*	2.539 *0.978*	0.319 *0.671*		
45-VRC01.H08.F-117225	0.598 *0.461*	0.663 *0.387*	1.469 *0.734*	1.44 *1.01*	0.813 *0.828*	1.762 *0.522*	0.738 *1.006*	0.618 *0.888*	
45-VRC01.H5.F-185917	0.664 *0.473*	0.672 *0.378*	1.486 *0.699*	1.463 *1.023*	0.832 *0.861*	1.76 *0.531*	0.765 *1*	0.674 *0.935*	0.505 *0.407*

B	VRC01	VRC-PG04	3BNC117	12A12	VRC-CH31	VRC-PG20	VRC23
VRC-PG04	1.192 *0.723*						
3BNC117	1.812 *0.959*	1.868 *0.908*					
12A12	0.688 *1.064*	0.938 *1.047*	1.716 *1.126*				
VRC-CH31	1.313 *0.967*	1.047 *1.008*	2.009 *1.026*	1.223 *0.504*			
VRC-PG20	1.003 *1.442*	1.065 *1.358*	1.783 *1.426*	0.765 *1.192*	1.301 *0.504*		
VRC23	0.585 *1.056*	1.049 *1.032*	1.718 *1.117*	0.52 *1.126*	1.237 *0.596*	0.799 *0.44*	
VRC03	1.421 *0.641*	1.756 *0.848*	2.218 *1.06*	1.431 *0.787*	1.891 *0.775*	1.674 *1.255*	1.46 *0.755*
VRC06	1.481 *0.953*	1.719 *1.007*	2.171 *1.032*	1.425 *0.958*	1.937 *0.847*	1.573 *1.344*	1.417 *0.963*
VRC08	1.028 *1.041*	0.989 *0.792*	1.875 *1.143*	0.688 *0.96*	1.297 *0.945*	0.677 *1.286*	0.833 *1.008*
45-VRC01.H5.F-185917	0.664 *0.473*	1.086 *0.705*	1.769 *0.986*	0.687 *1.122*	1.307 *1.029*	0.917 *1.464*	0.658 *1.131*

C	CDR H1	CDR H2	CDR H3	FR 3	Light Chain	Total
VRC01	34.5	563	153.7	140.7	369	1260.9
NIH45-46	57.1	669.5	360.9	166.4	192.4	1446.3
Average	**45.8**	**616.3**	**257.3**	**153.6**	**280.7**	**1353.6**
VRC03	83.6	659.4	143.1	176.5	336.9	1399.5
VRC06	68	630.7	148.9	176.5	335.2	1359.3
45-VRC01.H03+06.D-001739	69	665.2	103.2	270.7	332.5	1440.6
Average	**73.5**	**651.7**	**131.7**	**207.9**	**334.8**	**1399.8**
VRC08	61.2	501.2	456.7	120.5	425.5	1565.1
VRC08c	57.1	553.6	416.1	68.4	370.4	1465.6
45-VRC01.H08.F-117225	33.9	558.4	483.9	59.8	229.8	1365.8
Average	**50.7**	**537.7**	**452.2**	**82.9**	**341.9**	**1465.5**
45-VRC01.H5.F-185917	35.8	608.9	198.9	44	328.6	1216.2

Supplemental References

Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K., and Terwilliger, T.C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta crystallographica Section D, Biological crystallography *58*, 1948-1954.

Barouch, D.H., Yang, Z.Y., Kong, W.P., Korioth-Schmitz, B., Sumida, S.M., Truitt, D.M., Kishko, M.G., Arthur, J.C., Miura, A., Mascola, J.R., *et al.* (2005). A human T-cell leukemia virus type 1 regulatory element enhances the immunogenicity of human immunodeficiency virus type 1 DNA vaccines in mice and nonhuman primates. Journal of virology *79*, 8828-8834.

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

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., 3rd, Snoeyink, J., Richardson, J.S., *et al.* (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res *35*, W375-383.

Drummond, A.J., Rambaut, A., Shapiro, B., and Pybus, O.G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular biology and evolution *22*, 1185-1192.

Edgar, R.C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics *5*, 113.

Edgar, R.C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research *32*, 1792-1797.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta crystallographica Section D, Biological crystallography *60*, 2126-2132.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150-3152.

Georgiev, I.S., Doria-Rose, N.A., Zhou, T., Kwon, Y.D., Staupe, R.P., Moquin, S., Chuang, G.Y., Louder, M.K., Schmidt, S.D., Altae-Tran, H.R., *et al.* (2013). Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. Science *340*, 751-756.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. BMC bioinformatics *8*, 460.

Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic biology *61*, 1061-1067.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci *8*, 275-282.

Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences of the United States of America *87*, 2264-2268.

Kuhner, M.K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular biology and evolution *11*, 459-468.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Malcolm, S., Barton, P., Murphy, C., Ferguson-Smith, M.A., Bentley, D.L., and Rabbitts, T.H. (1982). Localization of human immunoglobulin kappa light chain variable region genes to the short arm of chromosome 2 by in situ hybridization. Proceedings of the National Academy of Sciences of the United States of America *79*, 4957-4961.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics *26*, 2462-2463.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*, 658-674.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. In Methods in Enzymology, C.W. Carter, Jr., and R.M. Sweet, eds. (New York: Academic Press), pp. 307-326.

Seaman, M.S., Janes, H., Hawkins, N., Grandpre, L.E., Devoy, C., Giri, A., Coffey, R.T., Harris, L., Wood, B., Daniels, M.G., *et al.* (2010). Tiered categorization of a diverse panel of HIV-1 Env pseudoviruses for neutralizing antibody assessment. Journal of virology *84*, 1439-1452.

Souto-Carneiro, M.M., Longo, N.S., Russ, D.E., Sun, H.W., and Lipsky, P.E. (2004). Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. Journal of immunology *172*, 6790-6802.

Suchard, M.A., Weiss, R.E., and Sinsheimer, J.S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. Molecular biology and evolution *18*, 1001-1013.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular biology and evolution *28*, 2731-2739.

Tiller, T., Meffre, E., Yurasov, S., Tsuiji, M., Nussenzweig, M.C., and Wardemann, H. (2008). Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. J Immunol Methods *329*, 112-124.

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