(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 October 2011 (13.10.2011)

PCT

(10) International Publication Number
WO 2011/126576 A2

(51) International Patent Classification:
C07K 14/155 (2006.01) *C07K 16/10* (2006.01)
C12N 15/49 (2006.01) *A61P 31/18* (2006.01)
A61K 39/12 (2006.01)

(21) International Application Number:
PCT/US2011/000642

(22) International Filing Date:
11 April 2011 (11.04.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/322,663 9 April 2010 (09.04.2010) US

(71) Applicants *(for all designated States except US)*: **DUKE UNIVERSITY** [US/US]; 2812 Erwin Road, Durham, NC 27705 (US). **LOS ALAMOS NATIONAL SECURITY, LLC** [US/US]; Los Alamos Mational Laboratory, LC/IP, Ms A187, Los Alamos, NM 87545 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **HAYNES, Barton, F.** [US/US]; c/o Duke University, 2812 Erwin Road, Durham, NC 27705 (US). **MONTEFIORI, David, C.** [US/US]; c/o Duke University, 2812 Erwin Road, Durham, NC 27705 (US). **LIAO, Hua-Xin** [US/US]; c/o Duke University, 2812 Erwin Road, Durham, NC 27705 (US). **GAO, Feng** [US/US]; c/o Duke University, 2812 Erwin Road, Durham, NC 27705 (US). **KORBER, Bette, K.** [US/US]; c/o Los Alamos National Security, LLC, Los Alamos National Laboratory, LC/IP, MS A187, Los Alamos, NM 87545 (US). **GNANAKARAN, S.** [LK/US]; c/o Los Alamos National Security, LLC, Los Alamos National Laboratory, LC/IP, MS A187, Los Alamos, NM 87545 (US). **DANIELS, Marcus, G.** [US/US]; c/o Los Alamos National Security, LLC, Los Alamos National Laboratory, LC/IP, MS A187, Los Alamos, NM 87545 (US). **BHATTACHARYA, Tanmoy** [IN/US]; c/o Los Alamos National Security, LLC, Los Alamos National Laboratory, LC/IP, Ms A187, Los Alamos, NM 87545 (US). **LAPEDES, Alan, S.** [US/US]; c/o Los Alamos National Security, LLC, Los Alamos National Laboratory, LC/IP, MS A187, Los Alamos, NM 87545 (US).

(74) Agent: **WILSON, Mary, J.**; Nixon & Vanderhye P.C., 901 North Glebe Road, 11th Floor, Arlington, VA 22203-1808 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

WO 2011/126576 A2

(54) Title: GENETIC SIGNATURES IN THE ENVELOPE GLYCOPROTEIN OF HIV-1

(57) Abstract: The present invention relates, in general, to HIV-1 and, in particular, to immunogens that elicit broadly neutralizing antibodies against HIV-1, and compositions comprising same. The invention further relates to methods of inducing the production of such antibodies in a subject.

GENETIC SIGNATURES IN THE ENVELOPE GLYCOPROTEIN OF HIV-1

This application claims priority from U.S. Provisional Application No. 61/322,663, filed April 9, 2010, the entire content of which is incorporated herein by reference.

This invention was made with government support under Grant No. AI067854 awarded by the National Institutes of Health. The government has certain rights in the invention.

TECHNICAL FIELD

The present invention relates, in general, to HIV-1 and, in particular, to immunogens that elicit broadly neutralizing antibodies against HIV-1, and compositions comprising same. The invention further relates to methods of inducing the production of such antibodies in a subject.

BACKGROUND

Elicitation of broadly cross-reactive neutralizing antibody (NAb) responses is a high priority for HIV-1 vaccines [1-4]. Many candidate immunogens elicit strong NAb responses against highly neutralization-sensitive strains of HIV-1; however, these vaccine-elicited antibodies neutralize very few circulating strains [5-7] and have not afforded protection in past human efficacy trials [8-10]. A recently completed efficacy trial in Thailand (RV144), in which a modest reduction in the rate of HIV-1 infection was observed [11], provides hope that with further improvements a more acceptable level of efficacy is obtainable. It is too soon to know whether NAbs might have contributed to the observed efficacy in RV144. Based on immunogenicity data from earlier phase I and II clinical trials of this and related vaccines [4,12], improved NAb responses may be one way to achieve greater protection. Such improvements are likely to require novel vaccine designs.

Most current efforts to design NAb-based HIV-1 vaccine immunogens are guided in part by knowledge of the molecular structure of the viral Envelope (Env) glycoproteins that serve as the sole targets for NAbs [13-16]. These Env glycoproteins consist of a surface gp120 and transmembrane gp41 that associate non-covalently and assemble into a trimeric complex of gp120-gp41 heterodimers on the virus surface, where the mature Env trimer spike mediates virus entry into host cells [17-19]. Entry is mediated by successive binding of gp120 to its cellular CD4 receptor and an obligatory coreceptor, most often the chemokine receptor CCR5, triggering conformational changes that permit gp41 to induce membrane fusion [18-20]. Env trimers and their individual constituents are genetically variable, conformationally flexible and heavily glycosylated, making them difficult targets for NAbs [1,2,19,21]. Because fitness constraints do not permit the virus to evolve to become completely resistant to neutralization [22,23], certain NAb epitopes remain vulnerable that are of particular interest for vaccine development. Some of these epitopes are well studied, whereas others remain unknown or only partially characterized [2,4,24].

The structural complexity of Env requires sophisticated methods for the analysis of NAb epitopes. X-ray crystallography and cryo-electron tomography, together with data from mutagenesis and biophysical studies, have been used to illuminate several vulnerable regions in great detail. Examples of how this information is used for novel immunogen designs include the optimization and stabilization of epitopes in the receptor and coreceptor binding regions of gp120 [25-27]. Other examples include innovative structural variants of gp41 [28-30] and optimal mimics of gp120 and gp41 epitopes recognized by broadly neutralizing monoclonal antibodies (mAbs) [31-35]. Although these new design efforts are in early stages of testing, none so far have yielded substantial improvements.

Many new concepts for NAb-inducing vaccines based on HIV-1 Env are being explored. These concepts are complicated by inconsistencies between the antigenic and immunogenic properties of key epitopes. For example, Env antigens that possess high affinity epitopes for broadly neutralizing mAbs fail to

elicit these types of antibodies [36-39, 28-30]. Also, gp120 antigens similar to those that performed poorly as early vaccine candidates contain epitopes that are capable of absorbing-out a substantial fraction of broadly NAbs in sera from a subset of HIV-1-infected individuals [40-43]. Some B cell responses might be down regulated by self-tolerance mechanisms, as has been suggested for epitopes in the membrane proximal external region (MPER) of gp41 [44,45]. Other B cell responses might be down regulated by immunosuppressive properties of gp120 [46-48]. Although it remains unclear why some of the most attractive Env epitopes are poor immunogens, the potent neutralizing activities of a subset of human mAbs [49,50] and sera from HIV-1-infected individuals [51] suggest it might be possible to design better vaccine immunogens.

A greater understanding of the antigenic and immunogenic properties of Env should facilitate the discovery of an effective HIV-1 vaccine. New insights are being gained from the use of computational analyses of large neutralization datasets derived from assays with HIV-1-positive sera and molecularly cloned Env-pseudotyped viruses. Statistically significant associations are sought between the neutralization susceptibility of a virus and its Env amino acid sequence. Previously, several amino acid signatures in gp120 and gp41 were identified that strongly associate with the antigenic determinants of NAbs in sera from HIV-1-infected subjects [52,53]. Such signatures could either be a consequence of direct contacts for NAbs, or reflect conformational requirements/constraints that regulate Ab access.

Because of the distinctive lineages in HIV evolution, found at multiple levels, it is critical to correct for the phylogenetic associations among sequences when defining signatures. Not accounting for phylogeny can lead to spurious positive signals that result from lineage effects and a reduced sensitivity, as was seen when associations were sought between host HLA and amino acid substitutions at the population level [54, 55].

The present invention results, at least in part, from the use of computational strategies to identify amino acid mutational patterns that correlate with NAb profiles independently of founder effects. Three distinct

phylogenetically-corrected statistical approaches have been used. The first included modifications of the approach taken by Bhattacharya et al. [54]; new modifications enabled looking at combinations of sites and amino acids within sites. Two novel statistical strategies for defining signatures were added, conditional mutual information and a modified decision forest approach. The computational signature identification methods were tested by accurately identifying a subset of the known determinants of the epitope for broadly neutralizing mAb b12. These methods were then applied in a reciprocal fashion to determine whether amino acid signatures in the Env proteins from HIV-1 infected individuals with particularly broad NAb responses could be identified relative to individuals who do not elicit broad responses. The findings suggest that broadly NAb responses are determined in part by features in the CD4-induced (CD4i) co-receptor binding site (CoRbs) of gp120.

SUMMARY OF THE INVENTION

In general, the present invention relates to HIV-1. More specifically, the invention relates to immunogens that elicit broadly neutralizing antibodies against HIV-1, and to compositions comprising same. The invention further relates to methods of inducing the production of broadly neutralizing antibodies in a subject.

Objects and advantages of the present invention will be clear from the description that follows.

BRIEF DESCRIPTION OF THE DRAWINGS

The patent application file contains at least one drawing executed in color. Copies of this patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

Figure 1. Maximum Likelihood tree of the Env sequences showing ancestral states and amino acid in the end taxa for position 185. This tree illustrates the distribution of b12 sensitive (magenta) and resistant (gray) Envs

among the different subtypes and recombinant lineages, and their genetic relationships. Envs (n = 319) are included in the tree, of which 251 are matched to phenotypes. Among the 68 without a phenotype is the blinded test set; b12 sensitivity values were obtained for 56 of these in the test set. There are many recombinant sequences included in the tree (as with essentially all HIV population trees), limiting the accuracy of the reconstruction of the evolutionary history. The phylogenetic corrections utilized for signature analysis are, however, dependent on the local region of the tree and the ancestral states near the tips of the branches, reducing the impact of inter-subtype recombination on the analyses. An Aspartic acid (D) at position 185 is strongly associated with b12 susceptibility (D versus not D [written as !D]) in the top contingency table. The phylogenetically corrected signature analysis supports this association, indicating it is not merely an artifact of one or more clades within the tree being more or less susceptible to b12. Thus, when an Env has mutated towards a D in this position (!D to D), the Env tends to be susceptible to b12, but when it moves away from it, (D to !D), the Env tends to be resistant. This tree is displayed with midpoint rooting.

Figures 2A and 2B. Correlates of b12 sensitivity. Fig. 2A) Counts of b12 sensitive and resistant viruses grouped by subtype, intersubtype recombinant and circulating recombinant forms (CRFs). Subtypes D (1 Env) and G (n = 5), CRF14 (n = 3), and recombinant Envs including A/CRF02 (n = 1), A/C/D/ (n = 3) and A/D (n = 4), were grouped into the "other" category. The only 2 subtype categories with greater numbers of b12 sensitive than resistant Envs were subtype B and B/C recombinants (in these two cases the green bars are higher than purple). A Kruskal-Wallace non-parametric comparison of all groups indicated that at least one subtype was distinctive (p = 0.033). A comparison of the B subtype Envs versus all others indicated that they were far more likely to be sensitive to neutralization by b12 (Fisher's exact p-value 3.7×10^{-5}, odds ratio 3.6, 95 % confidence interval: 1.9, 6.9). Fig. 2B) sCD4 susceptibility is greater among b12 sensitive viruses. The amount of sCD4 required for 50% neutralization was greater among b12 resistant viruses (Wilcoxon rank test, p= 0.0001, median and

interquartile range shown to the right of each distribution, median 9.8 μg/ml among b12 resistant viruses, median 5.1 μg/ml among sensitive viruses). Furthermore, among just the b12 sensitive viruses, the levels required for 50% neutralization by sCD4 and b12 were correlated (linear regression of log values, p = 0.003, R = 0.29, data not shown).

Figure 3. Alignment of b12 signature positions from each sequence with particular amino acids associated with b12 resistance and sensitivity. This alignment includes the 7 non-contiguous positions found using the contingency table approach with defined resistant/susceptibility patterns: positions 173, 185, 268, 364, 369, 461, and 651. The 5 signature sites identified by the decision forest strategy are a subset of these 7 sites. The positions are aligned to the consensus of the susceptible viruses, which in each case is an amino acid that was associated with b12 susceptibility, as shown in dark green at the top of each column. The 7 positions were extracted from each sequence. If the amino acid was the same as the b12 sensitive consensus at the top of the column, a space is left in the row. Spaces are indicative of the consensus susceptible form, where differences in sequence stand out more sharply. If the amino acid differed from the susceptible consensus, but was another amino acid associated with susceptibility, it is shown as light green. Amino acids associated with resistance are shown in red. Amino acids that were different but not associated with either resistance or susceptibility are shown in black. The susceptible viruses are ordered from the top left column through the second column, from the most sensitive (top left) to the least sensitive (bottom right) in terms of the concentration required for 50% neutralization. The least sensitive Envs (those require concentrations of 25-50 μg/ml of b12) were grouped with sensitive viruses and are boxed at the bottom of the second column.

Figures 4A-4D. Structural mapping of b12 signature sites in gp120. Fig. 4A. Locations of 8 b12 signature sites in a three-dimensional structure of gp120 (PDB code: 1RZK) with V1, V2 and V3 loops modeled for visualization as

described previously [128]. Yellow balls mark the C-alpha positions of signature residues. Fig. 4B. Locations of 3 signature sites that occur at the b12 binding face of gp120. The b12 (magenta) bound structure of gp120 (blue), corresponding to PDB code:2NY7. The red region in gp120 is less than 6.5 Å from the bound b12 antibody. Fig. 4C. Isosurface of the gp120 molecule showing the difference in electrostatic potential (+0.3 kT/e) due to mutation E268R in gp120 that results in a net positive electrostatic potential (blue) at the b12-gp120 interface region. Isosurface (+/-1 kT/e) of the b12 molecule showing the positive (blue) and negative (red) electrostatic potentials indicating b12 is highly electropositive (overall charge of +12). Fig. 4D. An illustration to capture how position 651 could impact binding to b12 through an allosteric pathway involving the gp120-gp41 interface. X-ray structure of b12 (marked in magenta) bound to a liganded gp120 core protein (PDB code: 2NY7) with a monomer gp41 that was homology-modeled based on the NMR structure of SIV gp41 [129] that is more representative of a post-fusion conformation. The region of gp120 that is in contact with b12 is marked in red. The disulfide bridged loop region of the gp41 molecule that is expected to interact with gp120 was placed in close proximity to the region where the N- and C-termini of gp120 come in close contact. This model is useful for illustration but does not represent the actual gp120-gp41 interaction, which is not yet resolved. A yellow ball indicates position 651 in gp41. Green balls are used to show the covarying sites at positions 84, 169, 429 and 432 in gp120, and position 602 in gp41. Silver balls in the model capture sites in gp120 and gp41 that have been shown through past experimental studies to influence gp120-gp41 assembly.

Figures 5A and 5B. Clustered heatmaps of sera and the test panel. Figs. 5A. K-means clustering of serum samples and virus isolates in the test panel, k=3. A 90% threshold for stability was used as a minimum criterion for defining robust clusters in the sera, given re-sampling noise due to experimental variation and bootstrap re-sampling of the test panel of Envs. 75% was used for the clustering the Envs in the figure, and these clusters were not subsequently used

for analysis. The color keys on the top and on the left indicate the clusters and their statistical robustness: red, blue and yellow correspond to the three clusters, with each robust cluster boxed. Blends of the three primary colors indicate how often in the re-sampling tests for a given serum or Env the sample falls in different clusters. The intensity of the color indicates how frequently each falls in its primary cluster. In the heat map, darker red indicates that the serum neutralized the virus potently, progressively lighter colors through yellow indicate increasing resistance, and cream color is completely resistant. Fig. 5B. K-means clustering of serum samples and virus isolates in the test panel, k= 2; again a 90% threshold for stability was used for the sera, 75% for the viral Envs.

Figure 6. Maximum Likelihood tree of the Env sequences showing ancestral states and amino acid in the end taxa for position 185. This tree illustrates the distribution Envs organized into a phylogeny as sampled from potent (magenta) or weak (gray) sera. The Envs used in the test panel of pseudoviruses were included along with the Envs from the serum samples in the tree; the taxa without a magenta or dark gray mark are from the test panel. The evolution of the signature site 419 K with respect to the phylogenetic tree is highlighted. Arg (R) is the most common amino acid in this position, and K is very rarely an ancestral state.

Figure 7. Alignment of signature sites that were associated with potent NAb responses. Unlike the alignment in Figure 3 that only included signature positions, this alignment captures short contiguous regions of Env near the CoRbs. Signature amino acids are highlighted using the same color scheme and organization as the heatmap in Figure 5A. Red highlights are amino acids that associate with potent sera; yellow highlights are amino acids that associate with weak sera. A vaccine strain selected on the basis of Envs in potent neutralizing sera from HIV-1-infected individuals might ideally capture as many of the red positions and as few of the yellow as possible (e.g., CH0219.e4 and CH080510.e.p2). CH0219.e4 (see Figs. 13 and 14) might be particularly

promising because it also has short variable loops (data not shown). Position 186 was identified using CMI and thus does not have specific amino acids associated with the serological behavior; however, both E and N seem particularly enriched in the group with the highest cross reactivity (cluster III).

Figure 8. The four signature sites in the CCR5 CoR region shown in a crystallographic three-dimensional structure of gp120 complexed with CD4 and the CD4i-specific mAb 17b (PDB code: 1RZK). The yellow balls mark the C-alpha positions of the signature residues. Three regions in gp120 are indicated: the inner domain in light blue, the outer domain, dark blue; and the bridging sheet, brown. Definitions for these regions are based on the X-ray study of Kwong et al. [127]. CD4 is marked in green. The light and heavy chains of 17b are marked in light and dark magenta, respectively.

Figure 9. CMI sites: HXB2 positive 163,182,665.

Figure 10. Signature sensitivity score correlates with b12 sensitivity.

Figure 11. Genetic signatures of broadly NAb responses. Shown is how sera from HIV-infected individuals were tested for neutralizing activity against genetically diverse strains HIV. Results among the serum samples were used to construct a "heat-map" to identify common patterns of reactivity. Novel computational analyses were used to compare these patterns to Env sequences in the serum samples in an effort to identify genetic signatures that associated with potent neutralizing antibody responses. Five signatures were identified.

Figure 12. Signatures on an X-ray-crystal structure of gp120. Shown is the location of four of the signatures identified in Fig. 11 on a crystal structure of ligated gp120 (the fifth signature is not shown because it is in a region of gp120 that is not present in the crystal structure). All 5 signatures reside in the CD4i

region of gp120 that is reconized by monoclonal antibody 17b (17b is the pink ribbob structure in the figure).

Figure 13. Sequence information for 0219 Env, including gp160 encoding sequence with start and stop codons shown, gp160 amino acid sequence, gp160 codon optimized encoding sequence, gp140 amino acid sequence and gp140 codon optimized encoding sequences.

Figure 14. SDS-PAGE and Western Blot of purified HV13341 (CH0219_e4ENV gp140C).

DETAILED DESCRIPTION OF THE INVENTION

An increase in knowledge of the molecular and antigenic structure of the HIV-1 envelope glycoproteins (Env) has yielded important new insights for vaccine design but translating this information to an immunogen that elicits broadly neutralizing antibodies has been difficult. The present invention is based, at least in part, on the use of phylogenetically-corrected statistical methods to identify amino acid signature patterns in Env that are associated with the neutralizing potency of the serum from which they were derived. The utility of methods for defining signature amino acid mutation patterns that correlate with neutralization phenotype was examined by analyzing Env sequences from 251 clonal viruses that were differentially sensitive to neutralization by the well-characterized gp120-specific monoclonal antibody, b12. Ten b12-neutralization signatures sites were identified, including key variable amino acid positions that occur in the b12-binding surface of gp120, and positions in the V2 region, known to impact b12 sensitivity. Other signatures were identified in gp120 and gp41 that may reflect an impact of quaternary structure on the b12 epitope. A simple algorithm based on the b12 signature pattern was predictive of b12 sensitivity/resistance in an additional blinded panel of 57 viruses.

As described in the Example that follows, these computational methods were applied to defining signature patterns in Env proteins based on the magnitude and breadth of neutralizing antibody responses in HIV-1-infected individuals. An analysis was made of a checkerboard-style neutralization dataset comprising a multi-subtype panel of 25 clonal Env-pseudotyped viruses tested with sera from 69 HIV-1-infected individuals from whom a serum gp160 sequence was derived by single genome amplification. Distinct clusters of sera with high and low neutralization potencies were identified (see Example below). Mutational patterns in six amino acid signature positions in serum Env sequences were strongly associated with either the high or low potency cluster. Five were in the CD4-induced coreceptor binding site of gp120, suggesting an important role for this region in the elicitation of broadly neutralizing antibody responses against HIV-1.

An Env that retains the full amino acid signature associated with potent antibody responses (see Fig. 7) represents a preferred vaccine antigen. As will be clear from the Example, CH0129.e4 and CH080510.ep2 are strains that retain such signature positions. The gp160 and gp140 sequences (including codon optimized DNA encoding sequences) for CH0219.e4 are set forth in Fig. 13. CH0219.e4 is a particularly preferred vaccine antigen because it has short variable loops.

The present invention thus relates to HIV Envs that retain the signature (preferably, the full amino acid signature) associated with potent antibody responses (e.g., the CH0219.e4 Env) and methods of using same as vaccine immunogens. The invention further relates to such Envs for use as diagnostic targets in diagnostic tests. The invention further relates to the use of wildtype (WT) virus sequences (e.g., CH0219.e4 sequences) in the preparation of a polyvalent HIV-1 vaccine (U.S. Provisional Application No. 61/282,526, filed February 25, 2010). Sequences that can be included in such a polyvalent vaccine for B cell response include env and for T helper and cytotoxic T cell response include gag, pol, nef and tat sequences (U.S. Application No. 11/990, 222, filed Aug. 23, 2006).

The vaccine antigens (immunogens) of the invention (e.g. Envs sequences that retain the signature associated with potent antibody responses) can be chemically synthesized and purified using methods well known in the art. The immunogens can also be synthesized by well-known recombinant DNA techniques. Nucleic acids encoding the immunogens of the invention can be used as components of, for example, a DNA vaccine wherein the encoding sequence is administered as naked DNA or, for example, a minigene encoding the immunogen can be present in a viral vector. The encoding sequence can be present, for example, in a replicating or non-replicating adenoviral vector, an adeno-associated virus vector, an attenuated mycobacterium tuberculosis vector, a Bacillus Calmette Guerin (BCG) vector, a vaccinia or Modified Vaccinia Ankara (MVA) vector, another pox virus vector, recombinant polio and other enteric virus vector, Salmonella species bacterial vector, Shigella species bacterial vector, Venezuelean Equine Encephalitis Virus (VEE) vector, a Semliki Forest Virus vector, or a Tobacco Mosaic Virus vector. The encoding sequence, can also be expressed as a DNA plasmid with, for example, an active promoter such as a CMV promoter. Other live vectors can also be used to express the sequences of the invention. Expression of the immunogen of the invention can be induced in a patient's own cells, by introduction into those cells of nucleic acids that encode the immunogen, preferably, using codons and promoters that optimize expression in human cells. Examples of methods of making and using DNA vaccines are disclosed in, for example, U.S. Pat. Nos. 5,580,859, 5,589,466, and 5,703,055.

The invention includes compositions comprising an immunologically effective amount of the immunogen of the invention (e.g., the gp160 or gp140 sequence set forth in Fig. 13) or fragment thereof (e.g., gp41, gp120, either alone or associated with lipids, or fragments of gp120), or nucleic acid sequence encoding same, in a pharmaceutically acceptable delivery system. The compositions can be used for prevention and/or treatment of immunodeficiency virus infection 9e.g., in a human). The compositions of the invention can be formulated using adjuvants (e.g., alum, AS021 (from GSK), oligo CpGs, MF59 or Emulsigen), emulsifiers, pharmaceutically-acceptable carriers or other ingredients

routinely provided in vaccine compositions. Optimum formulations can be readily designed by one of ordinary skill in the art and can include formulations for immediate release and/or for sustained release, and for induction of systemic immunity and/or induction of localized mucosal immunity (e.g, the formulation can be designed for intranasal administration). The present compositions can be administered by any convenient route including subcutaneous, intranasal, intrarectal, intravaginal, oral, intramuscular, or other parenteral or enteral route, or combinations thereof. The immunogens can be administered in an amount sufficient to induce an immune response, e.g., as a single dose or multiple doses. Optimum immunization schedules can be readily determined by the ordinarily skilled artisan and can vary with the patient, the composition and the effect sought.

Examples of compositions and administration regimens of the invention include consensus or mosaic gag genes and consensus or mosaic nef genes and consensus or mosaic pol genes and consensus Env with an Env that retains the above-described signature or mosaic Env with an Env that retains the above-described signature, expressed as, for example, a DNA prime recombinant *Vesicular stomatitis* virus boost and a recombinant Env protein boost for antibody, a poxvirus prime such as NYVAC and a protein Env oligomer boost, or fragment thereof, or DNA prime recombinant adenovirus boost and Env protein boost, or, for just antibody induction, only the recombinant envelope gp120 or gp140 as a protein in an adjuvant. (See U.S. Application No. 10/572,638, PCT/US2006/032907, U.S. Application nos. 11/990,222 and 12/192,015.)

The invention contemplates the direct use of both the immunogen of the invention and/or nucleic acid encoding same and/or the immunogen expressed as a minigene in the vectors indicated above. For example, a minigene encoding the immunogen can be used as a prime and/or boost.

It will be appreciated from a reading of this disclosure that the whole Envelope gene can be used or portions thereof (i.e., as minigenes). In the case of expressed proteins, protein subunits can be used.

As pointed out above, the invention also relates to diagnostic targets and diagnostic tests. For example, a signature-retaining Env of the invention can be expressed by transient or stable transfection of mammalian cells (or they can be expressed, for example, as recombinant Vaccinia virus proteins). The protein can be used in ELISA, Luminex bead test, or other diagnostic tests to detect antibodies to the transmitted/founder virus in a biological sample from a patient at the earliest stage of HIV infection.

The present invention also relates to antibodies specific for signature-retaining Envs of the invention, and fragments of such antibodies, and to methods of using same to inhibit infection of cells of a subject by HIV-1. The method comprises administering to the subject (e.g., a human subject) the HIV-1 specific antibody, or fragment thereof, in an amount and under conditions such that the antibody, or fragment thereof, inhibits infection.

In accordance with the invention, the antibodies can be administered prior to contact of the subject or the subject's immune system/cells with HIV-1 or after infection of vulnerable cells. Administration prior to contact or shortly thereafter can maximize inhibition of infection of vulnerable cells of the subject (e.g., T-cells).

As indicated above, either the intact antibody or fragment (e.g., antigen binding fragment) thereof can be used in the method of the present invention. Exemplary functional fragments (regions) include scFv, Fv, Fab', Fab and $F(ab')_2$ fragments. Single chain antibodies can also be used. Techniques for preparing suitable fragments and single chain antibodies are well known in the art. (See, for example, USPs 5,855,866; 5,877,289; 5,965,132; 6,093,399; 6,261,535; 6,004,555; 7,417,125 and 7,078,491 and WO 98/45331.)

The antibodies, and fragments thereof, described above can be formulated as a composition (e.g., a pharmaceutical composition). Suitable compositions can comprise the antibody (or antibody fragment) dissolved or dispersed in a pharmaceutically acceptable carrier (e.g., an aqueous medium). The compositions can be sterile and can in an injectable form. The antibodies (and fragments thereof) can also be formulated as a composition appropriate for topical

administration to the skin or mucosa. Such compositions can take the form of liquids, ointments, creams, gels, pastes or aerosols. Standard formulation techniques can be used in preparing suitable compositions. The antibodies can be formulated so as to be administered as a post-coital douche or with a condom.

The antibodies and antibody fragments of the invention show their utility for prophylaxis in, for example, the following settings:

i) in the setting of anticipated known exposure to HIV-1 infection, the antibodies described herein (or binding fragments thereof) can be administered prophylactically (e.g., IV or topically) as a microbiocide,

ii) in the setting of known or suspected exposure, such as occurs in the setting of rape victims, or commercial sex workers, or in any sexual transmission with out condom protection, the antibodies described herein (or fragments thereof) can be administered as post-exposure prophylaxis, e.g., IV or topically, and

iii) in the setting of Acute HIV infection (AHI), antibodies described herein (or binding fragments thereof) can be administered as a treatment for AHI to control the initial viral load and preserve the CD4+ T cell pool and prevent CD4+ T cell destruction.

Suitable dose ranges can depend, for example, on the antibody and on the nature of the formulation and route of administration. Optimum doses can be determined by one skilled in the art without undue experimentation. Doses of antibodies in the range of 10ng to 20 μg/ml can be suitable.

The present invention also includes nucleic acid sequences encoding the antibodies, or fragments thereof, described herein. The nucleic acid sequences can be present in an expression vector operably linked to a promoter. The invention further relates to isolated cells comprising such a vector and to a method of making the antibodies, or fragments thereof, comprising culturing such cells under conditions such that the nucleic acid sequence is expressed and the antibody, or fragment, is produced.

Certain aspects of the invention can be described in greater detail in the non-limiting Example that follows. (U.S. Application entitled "Methods For The

Generation Of Monoclonal Antibodies Derived From Human B Cells", filed April 9, 2010, Atty. Docket 01579-1561, is incorporated herein by reference.)

EXAMPLE

Experimental Details

Viruses, serum samples and mAb b12. All viruses were used as molecularly cloned Env-pseudotyped viruses that expressed the entire gp160 of the designated strain. The multisubtype panel of viruses used for analysis of b12 neutralization is described in Tables 7 and 8. The 25 viruses used to assess the neutralizing activity of HIV-1-positive serum samples were isolated from sexually acquired infections and were sampled early in infection to closely resembled transmitted/founder viruses. Among these, isolates 6535.3, QH0692.42, SC422661.8, PVO.4, AC10.0.29 and RHPA4259.7 belong to a recommended panel of subtype B reference strains [110]. Isolates Du156.12, Du172.17, Du422.1, ZM197M.PB7 and ZM214M.PL15 belong to a recommended panel of subtype C reference strains [111]. Isolates Q23.17, Q842.d12, Q168.a2, Q259.d2.17, Q461.e2 and Q769.d22 are subtype A reference strains [112]. Isolates BB1006-11.C3.1601, BB1054-07.TC4.1499, 700010040.C9.4520 and WEAU-d15.410.787 are subtype B clones that were confirmed by single genome amplification (SGA) and sequencing analysis to be true transmitted/early founder Envs [56], as were C subtype isolates Ce1086_B2, Ce0393_C3, Ce1176_A3 and Ce2010_F5 [113]. These latter 25 viruses utilized CCR5 as their major coreceptor and were considered to possess a tier 2 neutralization phenotype [114].

Serum samples were obtained from HIV-1-infected subjects who were enrolled in clinical protocols of the Center for HIV/AIDS Vaccine Immunology (CHAVI). All subjects were chronically infected at the time of enrollment. The precise length of time of infection was not known. The mAb b12 was provided by Quality Biologicals, Inc. (Gaithersburg, MD) as a complete IgG molecule.

SGA amplification and sequencing of gp160 genes. The SGA methods used here were described previously [115] and result in sequences that are not

corrupted by recombination during amplification. Viral RNA was prepared from 400 µl of patient plasma and eluted into 60 µl of elution buffer using EZ1Virus Mini Kit V2.0 (Qiagen, Valencia, CA). Viral cDNA was prepared with 20 µl of vRNA and 80 pmol of primer 1.R3.B3R (5'-ACTACTTGAAGCACTCAAGGCAAGCTTTATTG-3') in a 50 µl volume using Superscript III (Invitrogen; Carlsbad, CA). SGA of the cDNA was performed using nested PCR to obtain the *rev/env* cassette and to avoid artificial recombination and resampling of the viral genomes [116]. The cDNA was diluted 1:3, 1:9 and 1:27 (8 reactions per dilution) to determine a dilution with a positive rate of 20% or less. Each diluted cDNA (1 µl) was used for the first round amplification with primers 07For7 (5'CAAATTAYAAAAATTCAAAATTTTCGGGTTTATTACAG-3') and 2.R3.B6R (5'-TGAAGCACTCAAGGCAAGCTTTATTGAGGC-3'). First round PCR was carried out with 1 unit of Platinum Taq Polymerase High Fidelity (Invitrogen; Carlsbad, CA) and 10 pmol of each primer in a 20 µl volume. First round PCR products (2 µl) were used for a second round of PCR with primers VIF1 (5'-GGGTTTATTACAGGGACAGCAGAG-3') and Low2c (5'-TGAGGCTTAAGCAGTGGGTTCC-3'). The second round PCR used 2.5 units of Platinum Taq Polymerase High Fidelity and 20 pmol of each primer in a 50 µl volume. PCR thermocycling conditions were as follows for both rounds of PCR: one cycle at 94°C for 2 minutes; 35 cycles of denaturing step at 94°C for 15 seconds, an annealing step at 60°C for 30 seconds, an extension step at 68°C for 4 minutes, and one cycle at 68°C for 10 minutes. PCR products were visualized on a 1% agarose gel and purified with the QiaQuick PCR Purification kit (Qiagen; Valencia, CA). Sequence analysis of *env* PCR products was performed on both DNA strands by cycle-sequencing and dye terminator methods using an ABI 3730xl genetic analyzer (Applied Biosystems; Foster City, CA). Individual overlapping sequence fragments for each *env* SGA were assembled and edited using the Sequencher program 4.7 (Gene Codes, Ann Arbor, MI). The newly obtained *env* sequences were aligned with standard sequences for each subtype and circulating recombinant form (CRF) from the LANL database

(http://www.hiv.lanl.gov/content/index) using CLUSTAL W [117]. Manual adjustment for optimal alignment was performed by using MASE [118]. Alignments were used for the initial subtyping analysis using the SIMPLOT software [119]. Bootscan analysis was also performed to confirm the breakpoints of identified recombinant sequences using SIMPLOT. All sequences were further validated with RIP and HIV Blast (www.hiv.lanl.gov). Subtyping and recombination discrepancies between the methods were carefully considered and resolved.

Neutralization assay. Neutralization was measured as reductions in luciferase (Luc) reporter gene expression after a single round of infection with Env-pseudotyped viruses as described [110]. Briefly, 200 TCID50 of virus was incubated with serial 3-fold dilutions of test sample in duplicate in a total volume of 150 μl for 1 hr at 37°C in 96-well flat-bottom culture plates. Freshly trypsinized TZM-bl cells (10,000 cells in 100 μl of growth medium containing 37.5 μg/ml DEAE dextran) were added to each well. One set of control wells received cells plus virus (virus control) and another set received cells only (background control). After a 48-hour incubation, 100 μl of cells was transferred to a 96-well black solid plates (Costar) for measurements of luminescence using the Britelite Luminescence Reporter Gene Assay System (PerkinElmer Life Sciences). Neutralization titers are either the 50% inhibitory dilution (ID50, serum samples) or 50% inhibitory concentration (IC50, mAb b12) at which relative luminescence units (RLU) were reduced by 50% compared to virus control wells after subtraction of background RLUs. Assay stocks of molecularly cloned Env-pseudotyped viruses were prepared by cotransfecting 293T/17 cells with an Env-expressing plasmid and an env-minus backbone plasmid (pSG3Δenv) as described [110].

Definitions of neutralization sensitivity. To conduct Env sequence signature analyses with the goal of identifying mutational patterns that correlate with neutralization phenotypes, neutralization phenotypes needed first to be defined. For mAb, b12 the Envs were initially defined based on whether or not a 50% reduction in RLU could be achieved at the highest concentration of b12

used; if not, the Env was considered b12 resistant. This provided a Boolean neutralization sensitive/resistant phenotype to use as a basis for comparing the 251 Envs tested with b12. Later, a comparison was made of the levels of neutralization-sensitivity with the patterns in the b12 signature sites by using IC50 values.

Defining a serological phenotype based on a profile of potency of neutralization against a panel of viruses was more complex. It was first necessary to group HIV-1-positive serum samples that exhibited similar neutralization profiles against a panel of 25 viruses. To achieve this, a k-means clustering strategy with added features was used to assess the robustness of the clusters, that factors in the uncertainty that results from limited sampling and inter-assay variability (the impact of experimental noise was explored using a smooth bootstrap). Sampling limitations were explored by re-sampling either by rows or columns 1000 times, using a random-with-replacement bootstrap strategy. The impact of inter-assay variation was explored by a smooth bootstrap, re-sampling from a Gaussian model of noise centered on zero and based on a limited number of repeat data values. Noise was adding back to the original scores based on the model. The k-means clusters were then re-estimated 1000 times with noise added back [120]. Using these two strategies it was found that no more than k=3 distinctive clusters of sera were statistically justified, in that 2 or more sera were assigned to each of the three clusters with 90% confidence. Defining more than k=3 clusters was not justified using this criteria. Sera that were not assigned to a cluster 90% of the time were considered indeterminate; clustering patterns were generally more sensitive to sampling than inter-assay variability. To describe the NAb reactivity pattern of the 3 sera clusters in a Boolean framework (there are two categories, high versus low) for signature pattern analyses, a comparison was made of Envs that were members of each of the robust serological clusters to all other Envs in the study. For example, the Env sequences associated with the strongest sera (cluster III) were compared to the remaining Envs by combining those that were in clusters I and II and those that were poorly resolved. In a

second analysis, k was set to k=2 and just the statistically robust high and low clusters were compared, excluding the intermediate values from the comparison.

Computational methods: alignments, phylogenetic and signature analyses. Alignments used for signature analysis were generated with GeneCutter (www.hiv.lanl.gov) to provide codon-aligned DNA for phylogenetic analysis. Phylogenetically corrected methods were used to identify all signature sites; the contingency table method illustrated in Fig. 1 and Fig. 6 was described in detail in [54]. The reason phylogenetic corrections are critical is that observed patterns in data can result either from correlations imposed by the initial historical emergence of a lineage of viruses (founder effects), or in the case of HIV-1, a consequence of recent biological interactions. Not accounting for founder effects can lead to erroneous statistical conclusions [52]. The sequence of the virus depends on its full evolutionary history, while causal correlations are manifest in correlations with recent changes (Fig. 1 and Fig. 6). The separation of the two effects, i.e. a phylogenetic correction, is needed to estimate the impact of recent changes on phenotype, requiring statistical reconstruction the genealogical relationships between the viruses and a maximum likelihood estimate of recent ancestral forms of the viruses. This is implemented through maximum likelihood phylogenies. A large sample size is essential to power explorations of associations between phenotype and mutational patterns. Thus phylogenetic reconstruction becomes technically challenging because the number of possible relationships grows factorially with the number of sequences sampled, where even heuristic searches fail to find reasonable models without extensive computation. To improve the maximum likelihood tree reconstructions, the phylogenic code was adapted to new high performance computing platforms (http://www.lanl.gov/roadrunner/).

Felsenstein first developed the method of phylogenetically independent contrasts [121,122] to address similar problems, i.e., obtaining phylogenetic corrections when looking for correlations of mutational patterns with quantitative data. This method was applied to look at whether variable loop length and the number of PNLGs correlated with potent NAb responses. Because these quantities do not diffuse randomly through the phylogeny, the application of this

method is an approximation. Moreover, because hypervariable loop lengths and the number of PNLGs vary rapidly within infected individuals, a phylogenetic correction at the population level is less essential in this framework. Simple Spearman correlation tests were performed to explore these quantitative measures.

Conditional Mutual Analysis (CMI) based Signatures: Conditional mutual information (CMI) was used as a second computational method to identify positions that exhibit an association between mutation and phenotype (neutralization sensitivity) that is independent of phylogenetic lineage. CMI [123] generalizes the conventional mutual information measure [123] that quantifies the association between two objects, e.g., mutation and phenotype. CMI also quantifies the association between two objects but it conditions the association on a third object, in this case the ancestral state. CMI sums over the associations conditioned on different ancestral states, and so is potentially more sensitive for detecting associations than the contingency table analysis that involves one ancestor state at a time. On the other hand, if the biological signal exists only for some ancestral states and not others, the extra noise added may reduce the power of the test. The statistical significance of a CMI value at any given position was assessed by fixing the ancestral state to each candidate ancestor state in turn, and permuting the relation between mutation and phenotype 1000 times in order to break any potential association. The distribution of CMI values for such permuted data was used to determine p-values, whereas q-values were obtained from these using the method of Storey and Tibshirani [124]. To be inclusive, a cutoff of $q < 0.2$ was used to identify statistically interesting sites, such that a 20% false positive rate was expected among the identified signatures.

Ensemble Learning Technique Using Classification Trees. To model sequence changes across sites, an ensemble learning technique using classification trees was employed [125]. As with the CMI and contingency table approaches, a sensitive/resistant neutralization category was compared to phylogenetic signals. This neutralization quantity indicates when a virus is neutralized by a fixed amount of b12 antibody. A change observed between an observed amino acid and the corresponding position in the inferred parent sequence provides one

phylogenetic signal. Changes toward or away from each observed or inferred amino acid across all of the envelope protein sequences served as the set of phylogenetic signals. Signals are conditioned on the ancestor amino acid; thus, any given position can be an instance of the signal, not an instance of the signal, or not applicable for the signal.

To form a decision tree, a signal was first identified that best separated sequences into resistant and sensitive neutralization sets. Each set was then partitioned into two more sets using further signals that best track the neutralization phenotype. This refinement procedure was repeated until no additional signals improved the classification. It was necessary that the classification tree handle the absence of signals as well as their Boolean state in order to avoid phylogenetic artifacts. Prediction was performed by taking a tree and following a main signal, secondary signal, tertiary signal, and so on, according to signal values derived from new data. Even in the absence of mutational signals, a decision tree would still provide a prediction on the basis of whether resistant or sensitive viruses were more common in a training set.

It is conceivable that coordinated mutations or reversions could occur in a universal way across viruses (case 1). Alternatively, the interplay of viruses and hosts could result in different patterns of coordinated sequence change (case 2). To address the possibility that there can be context dependence on unmeasured quantities (i.e., virus behavior groups formed by some unknown process), a subset of the full training data (75%) was randomly sampled when building decision trees, performing 140 interactions of decision tree building with different training set samples. 75% of the data was chosen as a trade-off between statistical power (ability to see any group behavior) and diversity (ability to see several groups). One hundred forty iterations were chosen for computational feasibility. Evaluation of the performance of the decision tree models needed to be separate from the construction of the training data. Thus, before iterating the training set sampling and tree building, 5 sensitive and 5 resistant viruses were reserved for testing purposes. Good models from the 140 decision tree builds were defined as those models that perform better than 60% (instead of the expectation of 50% for

random guesses) on this reserve dataset.

Any one of the 140 training samples and resulting decision trees could represent either case 1 or case 2, as described above. Therefore, the full process of reserving a random test set and generating 140 models to 'hit' each test set was iterated 32 times. For each test set, on average 10 of the 140 models were obtained that were predictive to at least 60% accuracy. A majority vote of these model predictions was noted for each test set. A "majority vote" was conducted across the 32 test sets to provide the final neutralization prediction. Next, mutational patterns were identified that recurred most often at the top-level splits in the subset of good models across all runs. These provided another strategy for defining amino acid signatures of that correlate with neutralization phenotype (Table 1).

Unlike other decision forest or bootstrap aggregation approaches (a.k.a. bagging) [126], cross-validation within the training set was effected and the trees were pruned back before using them. This may limit overall accuracy, but it has the advantage that any decision tree model could be interpreted without overtly over-fitting a particular training data set.

Exploring whether the ability to find b12-related patterns in the signature data was specific for the b12 signatures. Positions 655 and 651 exhibit high levels of co-variation with sites in gp120 that either directly interact with b12, or may be important for gp41/gp120 interactions. To test whether sites 655 and 651 were not being over-interpreted and that it would not be possible to find b12-related patterns in virtually any random set of covarying sites in Env if a hard enough examination of the literature were made, three positions with comparable Shannon entropy to sites 651 and 655 were examined that were *not* associated with b12 by the analysis. The question then asked was whether the sets of sites that co-varied with these 3 random sites have the potential to offer reasonable hypotheses for the b12 sensitivity. Unlike 655 and 651, these covariation sets did not suggest any direct interpretation in terms the b12 binding surface on gp120 [35], alanine scanning for sites relevant to b12 [63], or the regions implicated in gp120-gp41 contact and stability, thereby increasing confidence that the

biological interpretations of the covariation results for gp41 positions 655 and 651 are meaningful (data not shown).

Structural mapping of signature positions. For structural mapping in gp120, three different structures were used. Use was made of a structure with loops modeled when residue positions in loops needed to be shown. In this structure, the core of gp120 corresponded to the X-ray structure of CD4-bound YU2 gp120 [127], with variable loops V1V2 and V3 modeled for visualization purpose as described previously [128]. For signature positions in the b12 binding surface of gp120, the X-ray structure corresponding to the PDB code 2NY7 [35] was used. Finally, for spatial mapping of the signature positions in the CD4i region, the X-ray structure with a PDB code, 1RZK, [127] that was solved with a CD4-17b complex was used. In one instance a three-dimensional structure of gp41 was used to suggest the possibility of allosteric effects within the gp120-gp41 complex. This latter gp41 structure was homology-modeled based on the NMR structure of SIV-1 gp41 structure [129]. Signature positions were mapped onto this structure based on the alignment of sequences with respect to HXB2. The positional numbering refers to HXB2. Three-dimensional images were generated using VMD [130].

Validation of b12 signatures. A holdout set of 56 pseudotyped Envs, for which the b12 sensitivity was known but withheld from the analysis team, was kept aside as a blinded test set to determine if it were possible to predict the b12 phenotype of Env-pseudotyped viruses based on either signature amino acid positions or the ensemble learning strategy. The training and test set of Envs are included in the phylogenetic tree shown in Fig. 1; viruses known to be b12-sensitive are magenta, those known to be b12-resistant are dark grey, and those used as a blinded test set are light gray. Several strategies to predict phenotype were employed, including the simple requirement of at least 4 sensitive and no more than 1 resistant amino acid in the 7 signature sites, a logistic regression based on the 7 signature sites, and the ensemble learning strategy based on the full Env alignment. A prediction of b12 sensitivity or resistance was made based on

all three strategies (Tables 3 and 4, Table 8) for each of the 251 original training sequences and 56 test sequences.

Results

Identification of signature sites and mutational patterns associated with b12 susceptibility. Neutralization data and Env sequences relating to the b12 epitope that overlaps the CD4 binding site (CD4bs) of gp120 [35] were analyzed as a means to partially validate the computational methods. The mAb b12 was chosen for methodological validation purposes because many details regarding its epitope are known, and because it is an epitope of great interest for vaccine design. The analyses utilized genetic sequences and b12 sensitivities of 251 clonal Env-pseudotyped viruses representing many HIV subtypes, recombinant lineages and disease stages (Fig. 1, Table 7). IC50 values were determined from neutralization curves where the highest dose of b12 tested was either 25 μg/ml or 50 μg/ml, depending on the experiment. Viruses not neutralized at the highest dose tested are referred to here as being resistant; that is not to say, however, that some of the viruses would not have been neutralized by higher b12 concentrations. Among the 251 viruses tested, 88 (35%) were sensitive at varying levels (Table 8), and the other 163 were resistant at the highest concentration tested.

First, potential correlates of b12 sensitivity were examined, including viral genetic subtype, sensitivity to soluble CD4 (sCD4), and the disease stage of the donor at the time of virus isolation. Multiple subtypes were included in the study (Fig. 1). Envs that were B subtype exhibited the highest frequency of b12 neutralization susceptibility (Fisher's exact test $p = 3.6 \times 10^{-4}$, comparing B subtype to all others, Fig. 2A). In situations like this, in which there is a strong clade structure in the evolutionary tree and an enrichment of the phenotype of interest in a particular clade, it is critical to employ strategies that include phylogenetic correction to avoid spurious positives when seeking amino acid signatures. In particular, amino acids that are enriched in the B subtype because

of lineage effects will have an inherent bias that can make them appear to be associated with b12 sensitivity.

Envs of the target viruses were obtained and sequenced at different stages of infection. The Fiebig stage [56] for most subjects at the time the Env was sampled was experimentally determined as an indicator of stage of infection (Table 7). When the Fiebig stage was not experimentally determined, the subjects were generally noted to be in a "chronic" or "early/acute" stage at the time the sample was obtained (Table 7). When the subjects were broken into categories of "chronic" (grouping those in Fiebig stages VI or V/VI, with those noted to be in chronic infection) and "early" (grouping Fiebig stages I-V, with those noted to be in acute or early infection) there was no difference between b12 sensitivity or resistance, nor was there any correlation between b12 sensitivity and the series of Fiebig stages (data not shown). Thus the results from this cross-sectional examination of b12 resistance at different stages of infection suggests that the emergence of b12 resistance over time that was observed in a longitudinal study in a small number of subjects [57] may not be a common pattern. Finally, consistent with previous findings [58], Envs that were susceptible to b12 neutralization were more sensitive to neutralization by sCD4 ($p = 0.00013$, Wilcoxon rank sum test, Fig. 2B). Among just the b12 sensitive viruses, there was a weak correlation between the neutralizing potencies of b12 and sCD4 (Kendall tau Rank Correlation: $p= 0.0015$, $tau = 0.23$, data not shown).

The signature analyses strategies identified ten b12 amino acid signatures in Env. Associations with a q value (false positive rate) < 0.2 are presented in Tables 1 and 2. Seven signatures (6 in gp120 and 1 in gp41) were identified by phylogenetically corrected contingency table analysis [54]. Specific amino acid mutational patterns in each position formed the basis of contingency table analysis; these are noted in Tables 1 and 2. An example of a single amino acid contingency analysis through the maximum likelihood tree, Aspartic Acid (D) at position 185, is illustrated in Fig. 1. The simple uncorrected Fisher's exact p value for this amino acid ($p < 10^{-8}$) indicated that a D in position 185 is highly associated with b12 sensitivity. The low p-values for the patterns of change and

stability relative to the most recent ancestral state as estimated through the maximum likelihood tree, showed that mutation away from D in resistant viruses ($p = 0.0005$), and towards D in sensitive viruses ($p= 0.0004$) were also associated with b12 sensitivity, providing assurance that the profound association with 185D and b12 sensitivity was not simply an artifact of shared lineages (Fig. 1, Tables 3 and 4). The low q values (Table 2, $q = 0.06$ and $q = 0.04$, respectively) indicate that these low p-values are not expected by chance alone, despite the very large number of tests performed (i.e., every amino acid found in every position in Env, and all combinations of up to three amino acids in every position). An analysis was also made of all potential N-linked glycosylation sites (the amino acid pattern NX[ST]) for associations with b12 activity. None had a q-value < 0.2, and the only one that showed borderline significance was found at position 149 (noted in Table 2). Finally, the b12/gp120 interface was explored more deeply, including all combinations of amino acids in pairs of sites in this region. Single sites accounted for most of the statistically significant signatures (Table 2). (A listing of these sites is included in Table 9).

Of these 7 sites defined by phylogenetically corrected contingency analyses, 5 were also identified as b12 signatures by an ensemble learning technique using classification trees, while 3 were also identified by conditional mutual information (CMI) analysis (Tables 1). The best predictors from the ensemble learning approach included a subset of the most significant amino acids in the contingency table (Tables 1 and 2). An additional 3 signature sites were uniquely identified by CMI analysis: 2 in gp120 and 1 in gp41 (Tables 1 and 2). The CMI approach was used with the intent of increasing the sensitivity to capture additional sites of interest. The contingency table analyses restrict each comparison at each site to a particular amino acid or combinations of amino acids in the ancestral state, using a subset of the available data. In contrast, CMI utilizes information across all ancestral states, but does not identify particular amino acids at the site of interest, just the sites that had mutational patterns associated with resistance or susceptibility. An alignment of the three additional sites that were identified by the CMI method is provided in supplement Fig. 9.

Each of these positions was relatively conserved; examining these alignments suggests the consensus amino acids 163T, 182V, and 655K are well tolerated among viruses with b12 sensitivity, but that mutations 163A, 182E and mutations away from 655K, were enriched among resistant viruses.

It is important to remember that while these associations are statistically supported (Tables 1 and 2), any mutation in isolation may not be able to alter the phenotype of a virus in the context of a given natural strain. For example, although a change away from D at position 185 was most significantly associated with b12 resistance, and was most predictive of the phenotype, the Env carrying the mutation remained b12 sensitive in 13/48 (27%) natural occurrences of this pattern. Thus, the signatures identified point to the biological relevance of mutational patterns among a population of circulating viruses but are not necessarily predictive in isolation in a single strain. Despite this, the pattern among the signature sites that was evident in their alignment (Fig. 3) was associated with phenotype. For example, higher frequencies of amino acid substitutions associated with a b12 resistant phenotype, and loss of substitutions associated with a b12 sensitive phenotype, summed over all 7 signature sites, were strongly associated with resistance. This indicates that effects at the positions identified were cumulative. Notably, the signature sites were identified based on a simple Boolean resistant/sensitive phenotype, yet resistance-associated amino acids accumulated across these sites in viruses with diminishing b12 sensitivity. Specifically, the left hand box in Fig. 3 includes all b12 sensitive pseudoviruses tested, and is ordered by diminishing sensitivity. Combinations of more resistant and fewer sensitive amino acids are evident among the least sensitive viruses nearing the end of the columns. This cumulative effect was the basis for a regression analysis used later in an attempt to predict b12 phenotype among a set of 56 hold out viruses (see below).

Structural and biological interpretation of the b12 signature sites b12 contact surface signatures in gp120. Figure 4A shows the locations of the 8 gp120 signature sites found in a three-dimensional structure of gp120 [35, 59-61]. Three

b12 signatures (positions 364, 369 and 461) occurred in (364 and 369) or near (461) the b12 contact surface of gp120 [35,58]. These three sites are shown in the context of a b12-bound gp120 structure in Figure 4B. Sites 364 and 369 are located in the CD4 binding loop in the outer domain of gp120, where both sites directly contact residues in the heavy chain of b12 in a crystallographic structure of b12 Fab complexed with a stabilized gp120 core molecule [35], and mutations at these positions have been shown to alter the b12 susceptibility of multiple HIV-1 viruses [58,62,63]. Alanine scanning showed that an N to A substitution at position 461 could diminish b12 binding affinity more than 10-fold [63]. Because site 461contacts CD4 and lies adjacent to residues that directly contact b12 in the gp120-b12 crystal structure [35], it may affect epitope exposure.

Wu et al. identified 3 amino acid substitution patterns (S364H, P369L/T/Q and T373M) that were predicted to impact b12 binding because of potential clashes in side chain rotomers at the b12 contact surface [58]; two of these were among the signature sites (364 and 369). They showed that an S to H substitution at position 364 substantially increased b12 binding and neutralization susceptibility in several natural viruses. In contrast, the relevant substitutions at positions 369 and 373 impacted b12 binding to gp120, but did not restore neutralization to several resistant natural viruses, suggesting the epitope was shielded in the functional Env trimer in these strains [58]. The analyses indicated that a P or S at position 364 was associated with susceptibility, whereas an A or H at this position was associated with resistance. In addition, an A or P at position 369 was associated with susceptibility, whereas an I, L, or Q was associated with resistance (Table 1). The T at position 369 that was predicted by Wu et al. [58] to interfere with binding is rare and was found only once in the data and that single occurrence was in a susceptible virus (Fig. 3). The third site identified by Wu et al., mutation T373M, was not found among the signature sites. In Wu et al., 373M was enriched among subtype B resistant viruses in conjunction with other mutations. In the present study, 34% of the b12-sensitive viruses overall carried an M at position 373, whereas 28% of the resistant viruses carried an M, and no significant association was found between an M at position 373 and resistance, in

fact M was slightly more common among sensitive viruses. There was a trend suggesting mutations away from T at position 373 were more common among resistant viruses ($p = 0.057$); however, this did not approach significance ($q = 1$).

V2 region b12 signatures. Four additional signatures (sites 163, 173, 182 and 185) occur near the C-terminus of the V2 region of gp120 (Fig. 4A). Some regions of V2 contain frequent insertions and deletions, making them difficult to align, and such regions were not included in the analyses. The signature sites identified in V2 were embedded in parts of the alignment that were conserved enough to be meaningful. Because no X-ray crystal structures of gp120 are available with an intact V2 loop, the positions on the loop are shown on a modeled loop for visualization (Fig. 4A, see Experimental Details). Based on the crystal structure of the V1/V2 stem, positions near the C-terminal end of the V2 loop are predicted to impact the b12 epitope [60,61]. Indeed, results from Alanine scanning mutagenesis confirm the critical importance of the V2 region for b12 binding. For example, a D to A substitution at the signature position 185 was previously found to diminish b12 binding affinity greater than 10-fold [63]. Moreover, a mutation in this position resulted in escape from b12 neutralization [62]. It was also found that significantly reduced V2 loop lengths, and a reduced number of potential N-linked glycosylation sites in the V5 loop, were associated with b12 neutralization (Table 2). A complete scan of the gain or loss of individual PNLGs throughout Env did not reveal an association with any one particular glycosylation site in b12 binding at the statistical threshold of $q < 0.2$.

The b12 signature at site 268. Site 268 is not believed to have been previously investigated for an effect on b12 binding and neutralizing activity. This site is spatially distant from the interface of b12 and gp120, located approximately 30 Å away [35] (Fig. 4A). Intriguingly, this signature involved a charge reversal from an acidic residue to a basic residue resulting in a +2 change at this site. Such a change could potentially have a long-range electrostatic effect, thereby impacting b12 binding, particularly since b12 is highly positively charged. Therefore, electrostatic potential calculations were carried out using the Adaptive Poisson-Boltzmann Solver (APBS) to quantify the change in electrostatic

contributions to the b12 binding arising from the substitution of a negative with a positive charge at this position. APBS solves the Poisson-Boltzmann equation, a continuum model for describing electrostatic interactions numerically [64]. The recent X-ray structure of b12-bound to the JRFL gp120 was used for these calculations [35], and the appropriate site-mutations were modeled in the backbone of JRFL. The overall structure was not relaxed and only the side-chain rotomer of the replaced residue was positioned in an energetically feasible position. It was found that a change from 268E to either 268R or K results in an estimated decrease of b12 binding by 1.4 Kcal/mol. In Fig. 4C, the isosurface surrounding gp120 shows the difference in electrostatic potential (+0.3 kT/e) due to the mutation E268R on gp120; interestingly the isosurface is close to the b12-gp120 interface region. This figure also shows that b12 is highly electropositive (isosurface of +/-1 kT/e) due to the charged nature of b12 (overall charge of +12), explaining the large decrease in binding energy upon E268R mutation. This is consistent with the phenotypic directionality captured by the signature analysis.

The finding of a b12 signature at site 268 that underwent a charge reversal prompted an exploration as to whether there are additional acidic residues in gp120 that could undergo similar charge changes. Obviously not all charged residues are in a position to reverse their charge state to escape immune pressure. Some are highly conserved due to functional constraints. Other acidic residues may take part in critical electrostatic interactions that stabilize the structure. Often charged residues are involved in salt-bridge interactions. In this latter case it is possible that co-varying charge changes could occur simultaneously at the salt bridge forming partners (i.e.: K/R---E/D salt bridge pair becomes E/D---K/R pair); a simple continuum electrostatics model would then predict no significant effect on electrostatic binding energy. To address these possibilities, a systematic examination was made of all of the acidic residues in the gp120 in the gp120-b12 bound X-ray structure. Details of these sites are provided in Table 10. Except for positions 106 and 268, all other positions had dependencies that prevent a negative to positive change, either due to salt-bridge interactions or sequence conservation. Positions 106 and 268 are the only acidic residues in the gp120

core that are not conserved and do not take part in a salt bridge interaction. Thus, site 268 provides a rare opportunity for a charge reversal pathway that would allow the virus to become resistant to neutralization by b12 or other positively charged antibodies.

b12 signatures in gp41. Two statistically significant signatures were identified in gp41. Both sites (positions 651 and 655) are in the C-heptad repeat that is expected to lie proximal to the N-heptad repeat targeted by the HIV-1 fusion inhibitor T-20 in the post-fusion conformation [65]. The C-heptad repeat also contributes to the formation of a six-helix bundle that mediates viral fusion with the cellular membrane [66]. Finding b12 signatures in gp41 is not unexpected, as mutations in gp41 are known to affect NAb epitopes in the CD4bs [67-75], including the b12 epitope [58,68]. These mutations include amino acids at positions 569, 577, 582, 668 and 675 in gp41 that affect CD4bs epitopes; and mutations at positions 569 and 675 affect the b12 epitope directly [58,68]. While positions 651 and 655 have not been directly implicated in b12 binding in previous studies, those studies were based on escape mutations in single virus strains (IIIB, MN, JR-CSF, Q461, Q769, YU-2). In contrast, this study was based on systematically identifying significant associations among 251 genetically diverse viruses. This broader scope of analysis may have led to the identification of sites in gp41 that more generally affect the b12 epitope among global variants.

To explore the question of how sites in the gp41 C-heptad repeat that are distant from the gp120-b12 binding interface could influence the b12 epitope, an identification was made of all sites within Env that significantly co-vary (hence potentially interact) with positions 655 and 651. To do this, the phylogenetically corrected contingency table approach was used to identify the sites that covaried with signature sites in Envelope. The resulting co-variation patterns for all 10 of the b12 signature sites, including the two gp41 signature sites, are summarized in Table 11. Position 655 was found to significantly co-vary with a single position, site 185, which was also the most significant signature site in gp120. As noted above, this site is located in the V2 region of gp120 and has been shown to be a critical residue for b12 binding affinity [63]. Thus, the association between

mutational patterns in position 655 and b12 neutralization could be a consequence of quaternary structural interactions, giving rise directly to the correlation between mutational patterns of position 655 and b12 sensitivity. Alternatively, the 185-655 interactions could be driven by a relationship that is independent of the b12 epitope. In this latter case, the statistical association between site 655 and b12 neutralization may be due to a correlation that is one step removed, i.e. an ancillary consequence of the direct interactions of site 185 and b12. 655K is the most common amino acid in this position, where both K and E appear to be associated with b12 neutralization sensitivity in the signature analysis. As an aside, O'Rouke et al. [76] studied in detail the impact of substitutions on neutralization in a site they call 655, but because they did not use standard HXB2 numbering, their site 655 is actually 653 in HXB2 and is not the signature site identified here.

Covariation patterns were more complex for site 651, which was found to have 9 covarying sites (Table 11), 4 of which are captured in a schematic molecular diagram in Figure 4D. Site 80 and site 169 are in a region of the V2 loop for which no crystal structure is available and therefore were excluded from gp120 in this diagram. Similarly, 3 sites were in the cytoplasmic tail and thus were not included here (sites 798, 817, and 822). Based on crystallographic data, covarying sites 429 and 432 (though not statistically supported b12 signatures) are spatially close to the CD4 binding loop in a region that contacts b12 [35]. A K432A substitution diminished b12 binding affinity > 10-fold [63]. The presence of this complex chain of covarying sites in gp41 and gp120 suggests allosteric effects, where site 651 is part of a set of spatially distant residues that modulate the gp120-gp41 interface and thereby influence the exposure of the b12 epitope in the quaternary configuration of Env. Receptor and coreceptor binding induce structural re-arrangements at the gp120-gp41 interface as a requisite step for membrane fusion [18,20]. In principle, genetic changes that influence the gp120-gp41 contact surface could have reciprocal allosteric effects on the CD4bs of gp120. Consistent with this hypothesis, two of the 651 covarying sites (position 84 in the N-terminal C1 region of gp120; position 602 in the gp41 disulfide loop)

occur in regions implicated directly in gp120-gp41 contact and stability [77-84] (Fig. 4D). Alternatively, the mutations in site 651 that correlate with b12 susceptibility might influence a different allosteric pathway that relies on quaternary interactions with the CD4 binding loop region (sites 429 and 432) or possibly V2 (site 169) in the context of a trimer.

Signature-based predictions of b12 neutralization. Three computational approaches (described above) were used to determine if it were possible to predict b12 neutralization phenotypes based on sequence information. Prediction strategies were developing based on the "training" set of 251 sequences used to define the original signature pattern. The three strategies were tested by predicting the b12 phenotype of a blinded set of 56 pseudotyped Envs sequences. The first strategy applied a simple rule based on inspection of the alignment of the seven signature sites with defined amino acids shown in Figure 3. If the sequences contained at least 4 "sensitive" amino acids, and no more than 1 resistant amino acid in these seven sites, it was classified as sensitive. In the second approach, logistic regression was used to formalize the contribution of change at each site in an attempt to refine the predictive ability of the signature. The third approach was to apply an ensemble learning technique using classification trees to the amino acid changes in the full alignment, with the thought that this method could be used both for prediction of b12 phenotype based on the full Env sequence, and for defining the particular signature positions and amino acids which contributed most to the b12 phenotype (Table 1). When applying the three methods to the original training set 251 viruses, it was found that the simple rule based on the alignment was less predictive than the logistic regression, and the ensemble learning method was the most predictive (Table 3). When the three methods were applied to the blinded test set, the order reversed, and in this case the first simple method was the most predictive (p-value = 0.007, Table 4). The predictive power of this simple signature based strategy further supports the relevance of the b12 signature sites and amino acids associations. The other two methods had higher rates of false negatives and were not significantly predictive (Table 4). Reason for this inadequate power are not clear

but could be due to differences in the sampling of the 251 viruses and the 56 viruses that limited the predictive power of the two computational prediction methods. The full set of predictions based on the three methods and the b12 experimental data are provided in Table 8.

As discussed previously, the signature sites were originally defined based on a simple classification of b12 sensitive or resistant phenotype. Thus, as seen in the left hand panel in Figure 3, the cumulative number of sensitive amino acids in the 7 positions tends to decrease as b12 sensitivity diminishes (green amino acids and agreement with the most common sensitive form), whereas resistant amino acids tend to accumulate (red amino acids). To formally test whether level of b12 sensitivity among the sensitive viruses was correlated with the signature pattern, the signature pattern was first reduced to a single sensitivity score. This was done by subtracting the number of resistant amino acids from sensitive amino acids (red from green, in Figure 3). The signature sensitivity score was correlated with b12 sensitivity ($p=0.0006$, Spearman's rho = -034, Fig. 10). Thus signature amino acids can be used to predict, with significant accuracy, both the initial sensitive and resistant classification and the level of sensitivity among b12 sensitive viruses. Because these sites were identified after correcting for founder effects in the training set, it can be assumed that the the correlation observed is causal.

Signature analysis of Envs that elicit potent NAb responses in HIV-1-infected individuals.

Clustering sera according to cross-reactivity and potency. A determination was next made as to whether the signature analyses methods could be used to identify amino acids that associate with broadly cross-reactive NAb responses in HIV-1-infected individuals based on heatmap clusters (Fig.5). Env sequences and neutralizing activities in sera from 69 chronically infected individuals were used for analyses. The serum samples were obtained from individuals in the United States, Malawi, South Africa, Tanzania and England and consisted of, 1 CF recombinant, 1 CRF01_AE, 1 A/G recombinant, 5 subtype A, 24 subtype B, and 37 subtype C HIV-1 infections (Fig. 6, and Table 12). These

69 serum samples were chosen from among 360 sera that were assayed against a panel of twelve viruses (6535.3, QH0692.42, SC422661.8, PVO.4, AC10.0.29, RHPA4259.7, Du156.12, Du172.17, Du422.1, ZM197M.PB7, ZM214M.PL15, CAP45.2.00.G3). The 69 selected samples represented a wide spectrum of neutralization potencies against these 12 viruses. For increased statistical power in terms of robust assignments of potent versus weakly cross-neutralizing sera, they were assayed against an additional multi-subtype panel of viruses, such that the total number of pseudoviruses assayed was 25 (6 subtype A, 10 subtype B, 8 C and 1 BC recombinant, all isolated early in infection, see Table 13). The final checkerboard-style results (Fig. 5) confirmed a wide spectrum of neutralization potencies, including a subset of samples that contained high titers of NAbs against a majority of viruses tested, and for contrast, a subset of comparable size that was poorly cross-neutralizing.

The combined neutralization results were clustered according to the ability of individual serum samples to neutralize the panel of 25 viruses, using a k-means strategy that factors in the robustness of the clusters according to the uncertainty that results from limiting sampling (bootstrap) and assay-to-assay variability (noise) (Fig. 5). To assess the impact of assay variability, error estimates were factored in based on a limited number of repeat experiments. To do this, error (drawn from a log-normal distribution based on the repeat data) was added to the real data, and created 1000 reconstructed data sets. This made it possible to resolve clusters that should be robust relative to inter-assay variation (Fig. 5, noise). Next, a re-sampling was made from among the 25 Envs used in the neutralization assays 1000 times to see if the clusters would be robust if a different test panel of Envs with similar, but less diverse, composition had been selected. Figure 5A shows 3 distinct clusters (k=3) that turned out to include sera with high, medium and low neutralization potencies, respectively. k=3 was the maximum number of clusters that could be meaningfully assigned, given the constraint that each cluster must contain at least 2 members, and that the members must meet the stability criteria of being associated with the assigned cluster in >90% of each of the two re-samplings experiments (i.e., "bootstrap" and "noise").

Standard k-means strategy was used to assign each serum to a single k-cluster; however, if based on re-sampling statistics described above, some sera could not be assigned to any of the k=3 clusters, they are shown as intermediate values. To make use of all 69 data points in a Boolean framework for signature analysis, including these intermediate values, three sets of signature analyses were conducted for the k=3 clusters, comparing each one of the 3 robust clusters to all other data points. A k=2 clustering was also performed that enabled a robust extreme "high" and "low" 2-cluster comparison that captured most of the data (Figure 5B). This latter signature analysis did not resolve new signature site, but did sometimes improve the statistical confidence in a given site (Table 6). Phylogenetically corrected methods similar to those described for the b12 sensitivity signatures were used to identify associations between serum Env sequences and distinct neutralization clusters.

Defining signature patterns in serum-derived Envs. Envs sequences from all 69 sera were scanned for patterns of mutations that correlated with particularly weak or strong neutralizing capacities. The analysis compared all single sites and all pairs of adjacent sites for signatures of either 1 amino acid or combinations of amino acids at each site. In this complete Env scan, a single signature was found in the CoRbs. This signature consisted of a pair of amino acids in which the combination of either G or S at position 412, together with N at 413, was found to be enriched in Envs from potent neutralizing sera. An examination was made for signatures in potential N-linked glycosylation sites (PNLGs) throughout Env and a single signature pattern with borderline significance was again found that was also located at position 413 in the CoRbs; in this case the PNLG was preserved in Envs from individuals with potent sera. Using the CMI approach to scan the full Env protein, an additional signature was identified at position 186 in the V2 loop.

A more in-depth exploration of regions in the receptor and coreceptor binding sites of gp120 and in the MPER of gp41 was next performed. The sets of positions used for these analyses, and the references from which they were drawn, are listed in Table 9. These three regions were selected because antibodies against each one have each been identified in a subset of HIV infected people who

possess potent cross-reactive NAb responses [85-88]. An examination was made of combinations of multiple amino acids at multiple positions in these regions of interest. This sort of in-depth exploration was neither computationally feasible with the full Env, nor was it desirable because multiple test issues would have limited the power to find weak signatures if the full Env was explored so intensively. The deeper focused analysis revealed additional signatures but only in the CoRbs (Tables 5 and 6). No correlations were found in either the CD4bs or MPER region even through these regions were also targeted for a more focused and in-depth analyses. Finally, as with b12 neutralization, an examination was made as to whether potent NAb responses were associated with other general features of Env that are known to affect epitope exposure, such as the extent of N-linked glycosylation and the size of the variable regions of gp120 [89-92]. The V2 loop was shorter with fewer PNLGs in Envs from subjects with potent sera, and Envs with shorter V5 loops [93] were also correlated with potent sera. The mutational patterns in all of the signature sites are highlighted in sub-region alignments in Fig. 7, ordered and colored according to the k=3 heatmap clustering scheme shown in Fig. 5A. An Env such as CH0219.e4 might be particularly promising as a vaccine antigen, because it retains the full amino acid signature associated with potent antibody responses (Fig.7), and it also has short variable loops (Figs. 13 and 14).

Structural and biological interpretation of signature sites that correlated with potent NAb sera

The combined results of the contingency table signature analyses identified five statistically significant signature sites that resided in, or proximal to, the CCR5 CoRbs of gp120 (Tables 5 and 6). These sites are shown in a crystallographic model of gp120 complexed with CD4 and the CD4i-specific mAb 17b in Figure 8. Sites 419 and 421 are located in the V4 region of gp120, immediately adjacent to the β20 strand of the bridging sheet that connects the inner and outer domains of gp120 [35,60]. Both sites make contact with the CD4i-specific mAb 17b [60] (Fig. 8) and have been shown to be critical for

CCR5 co-receptor binding [94-97]. Site 419 also makes contact with b12 [35], whereas site 421 is involved in the binding of other CD4i-specific mAbs E51 [97] and 48d [98] as well. Sites 413 and 440 in V4 and C5, respectively, are spatially close to the bridging sheet and overlap the contact surface for 17b [60]. Site 440 has been shown to be critical for CCR5 binding [95-97]. CMI analysis identified an additional site in the V2 loop, position 186, immediately adjacent to the b12 signature site at position 185. In addition to the position-based signature analysis, it was found that strong NAb responses were associated with serum Env proteins that had fewer PNLGs and shorter lengths in V2 (Table 6). It has been shown that V1/V2 stem region can impact CCR5 binding since it plays a significant role in formation of the bridging sheet [95,96]. Furthermore, site-directed mutational studies have shown that regions outside V3 loop, including site 166 (a position within V2 loop) can play a significant role in co-receptor usage/switch [93,99]. Considering the flexibility of the loop and ensuing conformational changes that take place involving V1/V2 upon CD4 binding, a position such as 186 can directly or indirectly interact with critical sites involve in the formation of bridging sheet. The fact that no other signatures were identified suggests that the CCR5 CoRbs plays a substantial and relatively consistent role in the NAb response in HIV-1-infected individuals.

In summary, assay technologies that utilize molecularly cloned Env-pseudotyped viruses with a defined sequence are powerful tools for dissecting molecular determinants of neutralization epitopes on HIV-1. In addition to enabling mutagenesis studies, data from assays with clonal Env-pseudotyped viruses have been used for computational analysis to identify Env amino acid signatures that associate with the antigenic recognition patterns of autologous [53] and heterologous [52] NAbs in sera from HIV-1-infected individuals. Although not confirmed in previous studies, such signatures could be contact sites for NAbs, or they may be determinants of epitope exposure in the quaternary structure of Env spikes. Here, partial validation was obtained of a computational strategy to accurately identifying amino acid positions that are related to NAb phenotypes. Patterns of mutations in Env proteins that correlate with b12

susceptibility were systematically studied and key positions that are known from crystallographic and mutagenesis studies to be critical sites in the b12 epitope were successfully identified. These sites were predictive of b12 susceptibility. In addition to this validation of the approach, new information was gained by defining the particular mutations in the natural virus population that most profoundly impact b12 neutralization susceptibility, and by determining the relative strength of such contributions (Table 2). Thus, 7/8 gp120 signatures were identified either directly in the contact surface for b12, or in V2, which is known to impact b12 binding (Table1). Notably, mutations in position 185 in V2 were nearly equal in strength to mutations in position 461, which are the two best predictors for assessing b12 neutralization susceptibility in natural strains. A new position, 268, was implicated in b12 binding. This signature raised a plausible hypothesis regarding the impact of electrostatic potential at the isosurface of gp120 on interactions with the positively charged b12 antibody. Two additional b12 signatures were identified in gp41 that were intriguing because they may affect exposure of the b12 epitope in the quaternary structure of Env. Interestingly, both were directly co-varying with sites at the b12-gp120 interface. Two of the ten sites identified are statistically expected to be false positives, so it is likely that two will be not be found to be relevant when experimentally tested, although each of the ten are biologically plausible.

The apparent accuracy of the b12 susceptibility signature analysis was encouraging; however, the findings highlight both limitations and virtues of these methods. Sequence-based signatures methods cannot be expected to identify all b12 contact residues in gp120 [35]; this is because some of these sites are highly conserved, whereas other sites at the contact interface may have natural variation that is well tolerated by b12. Yet other sites might reside in variable regions that cannot be aligned with confidence. In addition, since these methods start with no biological priors, they necessarily need a large number of tests that makes detecting weak signatures prohibitively data intensive. For example, two PNLGs known to affect b12 susceptibility [58] were not identified. One of these sites was at the base the V2 loop (position 197) and the other was in the V3 loop (position

301). The PNLG in position 197 is almost invariant, and so could not have been identified by this method, which relies on sequence variability. Position 301 (PNLG) reached borderline significance in the complete scan of Env when testing for an association between the preservation or loss of PNLGs and the b12 neutralization ($p = 0.019$, $q = 0.30$, $OR = 0.23$).

Signature methods focus on sites that are most impacted by common mutational patterns found in the circulating population. Such mutational patterns are directly relevant for vaccine design considerations because it is necessary to contend with natural variation for a vaccine to succeed. Indeed, signature methods provide a useful counterpoint to crystallography, which identifies the contact surface of a protein bound by antibody, but does not provide direct information about the implications of key common natural mutations [35]. Moreover, alanine scanning [63], which explores the functional impact of mutations introduced in either conserved or variable positions, is a valuable tool, but one that is limited in terms of being able to look at the consequences of natural variation at specific sites or in combinations of sites. An additional limitation is experimental, in that some sites might require concentrations of b12 that are higher than those used here for positive identification. Despite these limitations, the computational analysis appears useful for delineating the molecular determinants of complex neutralization epitopes on HIV-1 Env, including the identification distant sites that may impact b12 binding though quaternary and allosteric effects. The neutralizing impact of b12 is very specific, where slight differences in recognition sites between viruses can have major phenotypic consequences [100]. A better understanding of the impact of common natural mutations that are outside of the immediate binding surface of b12 may ultimately allow improved rational design strategies of vaccines that attempt to elicit potent anti-CD4bs antibodies.

Having confirmed that the computational analysis has utility for identifying molecular determinants of Env antigenicity in the context of the b12 epitope, an effort was made to determine whether a similar computational analysis, based on Env sequences in serum samples from HIV-1-infected

individuals, could identify amino acid signatures that associate with the magnitude and breadth of the neutralizing activity of the serum samples. Any signatures identified by this analysis might be determinants of the immunogenic as well as antigenic properties of Env, although it was beyond the scope of this study to discriminate between these two immunologic properties.

For the analyses of Env sequences in serum samples that were evaluated for neutralizing activity, a single Env sequence from each individual was obtained. There was interest in leveraging the resources to increase the number of individuals studied rather than increasing the depth of characterization of infected individuals. In part a test was being made of the feasibility of the approach for scanning a large population of HIV infected individuals with the intent of finding common features of the virus harbored in them that may have given rise to a potent NAb response. Viral evolution and quasispecies complexity in chronically infected subjects clearly were potential confounding factors; the single sequence used was randomly selected from a complex viral population within each individual and may not reflect the form of the Env that gave rise to the NAbs of interest in the serum samples. Indeed, assuming that the NAb response during chronic infection is driven by multiple viral variants, these confounding factors limit the ability to identify genetic signatures. Despite this, statistically significant signatures were revealed based on an analysis of sequences from a single Env clone from a single time point from each of 69 individuals. Notably, these signatures were focused on a single biologically interesting region, the CoRbs. An unresolved issue that is an inherent consequence of this signature-defining strategy is the uncertainty regarding whether the signature amino acids reflect common features that were useful for stimulating potent NAb responses, or if instead they reflect common patterns of escape from the NAb responses in the potent sera. Experimental comparisons to resolve this are underway; strains that retain the signature positions that are associated with potent sera, like CH0219.e4 and CH080510.e.p2 (Fig. 7), are particularly interesting candidates for immunogenicity testing.

The fact that five of the six signature sites identified, with one false-positive expected, were in the CoRbs of gp120 suggests an important role for this region in generating high titers of broadly NAb responses. This region is comprised of elements of the bridging sheet and adjacent surfaces from the outer domain of gp120, including the V3 loop, that undergo conformational changes and become exposed upon CD4 binding as an intermediate step in the membrane fusion process [60,95,96,101-103]. It is possible that in some cases CD4i-specific mAbs contribute directly to potent cross-neutralizing ability [94,86]. The CoRbs is one of the most highly conserved and protected domains on gp120 [85]. Rare variants of HIV-1 exist that exhibit spontaneous exposure of CD4i epitopes; these strains tend to infect cells independently of CD4 and to be highly sensitivity to neutralization by CoR-specific antibodies [104,105]. Owing to the presence of such antibodies in HIV-1-infected individuals [85,86,94], a mechanism of CD4-induced exposure of the CoRbs serves as an effective strategy to evade humoral immunity — a strategy that is aided by steric constraints that prevent anti-CoR antibodies from gaining accessing to their epitopes at the virus-cell interface [106]. In a systematic thermodynamic analysis by Kwong et al., in which 20 antibodies were categorized according to where they bind on the gp120 surface, it was found that 6 of 7 antibodies that bind gp120 at its receptor and coreceptor binding sites exhibited unusually high binding entropy (including 17b that binds to CoRbs) [21]. Therefore, the signature sites identified here in the CoRbs might play an indirect role in neutralization by antibodies that induce large conformational changes in gp120.

The question naturally arises as to why a region of gp120 that is so heavily guarded and difficult to target by NAbs registered in the analysis as a key determinant of potent NAb responses in HIV-1-infected individuals. One possibility is that the CoRbs of gp120 has vulnerabilities that are only beginning to be recognized. For example, using a novel combination of epitope mapping techniques, Li *et al.* [94] reported evidence that CoRbs-specific antibodies contributed to the broadly cross-reactive neutralizing activity of serum from two HIV-1 infected individuals. In addition, CoRbs residues were implicated by

alanine scanning mutagenesis as being involved to a minor extent in the epitopes for two newly described broadly neutralizing mAbs [50]. Also, vaccine-elicited CoRbs-specific antibodies correlated with viremia control in a simian-human immunodeficiency virus (SHIV) challenge model in nonhuman primates [107]. It also seems possible that amino acid residues in key positions in the CoRbs of gp120 modulate the conformation of adjacent regions, such as the CD4bs, much the same as conformational changes induced by gp120-CD4 binding modulate the CoRbs. Limited sequence variability in the CD4bs [108,109] makes this an attractive target for NAb-based vaccines. Indeed, studies have shown that the CD4bs is targeted by broadly NAbs in sera from some HIV-1-infected individuals [51].

It remains to be determined whether the genetic signatures of potent NAb responses identified here contribute to the immunogenicity as well as antigenicity of Env. By design, an attempt is being made to resolve signatures that impacted Env immunogenicity in natural infection. Clearly, strong antigenicity alone is generally not sufficient for the elicitation of NAbs [28-30, 36-39]. Other requirements may need to be met before B cells can be stimulated to produce NAbs against certain epitopes of interest. Although very little is known about what these requirements might be, proper Env configuration for B cell recognition and antibody affinity maturation should be considered. It will be interesting to test novel Env immunogens that naturally contain the genetic signatures identified in the study, or that introduce these signatures experimentally. At the very least, these findings suggest that greater attention should be paid to the CoRbs of gp120 when designing novel vaccine immunogens.

References

1. Burton DR, Desrosiers RC, Doms RW, Koff WC, Kwong PD, et al. (2004) HIV vaccine design and the neutralizing antibody problem. Nat Immunol 5: 233-6.

2. Haynes BF, Montefiori DC (2006) Aiming to induce broadly reactive neutralizing antibody responses with HIV-1 vaccine candidates. Expert Rev Vaccines 5: 579-95.

3. Mascola JR (2003) Defining the protective antibody response for HIV-1. Curr Mol Med 3: 209-216

4. Mascola JR, Montefiori DC (2009) The role of antibodies in HIV vaccines. Ann Rev Immunol, in press.

5. Belshe RB, Graham BS, Keefer MC, Gorse GJ, Wright P, et al. (1994) Neutralizing antibodies to HIV-1 in seronegative volunteers immunized with recombinant gp120 from the MN strain of HIV-1. JAMA 272: 475-80.

6. Bures R, Gaitan A, Zhu T, Graziosi C, McGrath K, et al. (2000) Immunization with recombinant canarypox vectors expressing membrane-anchored gp120 followed by gp160 protein boosting fails to generate antibodies that neutralize R5 primary isolates of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses 16: 2019-2035.

7. Mascola JR, Snyder SW, Weislow OS, Belay SM, Belshe RB, et al. (1996) Immunization with envelope subunit vaccine products elicits neutralizing antibodies against laboratory adapted but not primary isolates of human immunodeficiency virus type 1. J Infect Dis 173: 340-348.

8. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF (2005) Placebo controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J Infect Dis 191: 654-65.

9. Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, et al. (2005) Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. J Infect Dis 191: 666-677.

10. Pitisuttithum P, Gilbert P, Gurwith M, Heyward W, Martin M, et al. (2006) Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. J Infect Dis 194: 1661-1671.

11. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, et al. (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361: 2209-2220.

12. Nitayaphan S, Pitisuttithum P, Karnasuta C, Eamsila C, de Souza M, et al. (2004) Safety and immunogenicity of an HIV subtype B and E prime-boost vaccine combination in HIV-negative Thai adults. J Infect Dis 190: 702-706.

13. Burton DR (2002) Antibodies, viruses and vaccines. Nat Rev Immunol 2: 706-713.

14. Dormitzer PR, Ulmer JB, Rappuoli R (2008) Structure-based antigen design: a strategy for next generation vaccines. Trends Biotechnol 26: 659-67.

15. Phogat S, Wyatt R (2007) Rational modifications of HIV-1 envelope glycoproteins for immunogen design. Curr Pharm Des 13: 213-27.

16. Schief WR, Ban YA, Stamatatos L (2009) Challenges for structure-based HIV vaccine design. Curr. Opin. HIV AIDS 4: 431-440.

17. Center RJ, Leapman RD, Lebowitz J, Arthur LO, Earl PL, Moss B (2002) Oligomeric structure of the human immunodeficiency virus type 1 envelope protein on the virion surface. J Virol 76: 7863-7867.

18. Chan DC, Kim PS (1998) HIV entry and its inhibition. Cell 93: 681-684.

19. Wyatt R, Sodroski J (1998) The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. Science 280: 1884-1888.

20. Weissenhorn W, Dessen A, Calder LJ, Harrison SC, Skehel JJ, Wiley DC (1999) Structural basis for membrane fusion by enveloped viruses. Mol Membr Biol 16: 3-9.

21. Kwong PD, Doyle ML, Casper DJ, Cicala C, Leavitt SA, et al. (2002) HIV-1 evades antibody-mediated neutralization through conformational masking of receptor binding sites. Nature 420: 678–682.

22. Deeks SG, Schweighardt B, Wrin T, Galovich J, Hoh R, et al. (2006) Neutralizing antibody responses against autologous and heterologous viruses in acute versus chronic human immunodeficiency virus (HIV) infection: evidence for a constraint on the ability of HIV to completely evade neutralizing antibody responses. J Virol 80: 6155-6164.

23. Draenert R, Allen TM, Liu Y, Wrin T, Chappey C, et al. (2006) Constraints on HIV-1 evolution and immunodominance revealed in monozygotic adult twins infected with the same virus. J Exp Med 203: 529-539.

24. Burton DR, Stanfield RL, Wilson IA (2005) Antibody vs HIV in a clash of evolutionary titans. Proc Natl Acad Sci USA 102: 14943-14948.

25. Dey B, Pancera M, Svehla K, Shu Y, Xiang SH, et al. (2007) Characterization of human immunodeficiency virus type 1 monomeric and trimeric gp120 glycoproteins stabilized in the CD4-bound state: antigenicity, biophysics, and immunogenicity. J Virol 81: 5579-93.

26. Dey B, Svehla K, Xu L, Wycuff D, Zhou T, et al. (2009) Structure-based stabilization of HIV-1 gp120 enhances humoral immune responses to the induced co-receptor binding site. PLoS Pathog 5: e1000445.

27. Wu L, Zhou T, Yang ZY, Svehla K, O'Dell S, et al. (2009) Enhanced exposure of the CD4-binding site to neutralizing antibodies by structural design of a membrane-anchored human immunodeficiency virus type 1 gp120 domain. J Virol 83: 5077-86.

28. Ho J, Uger RA, Zwick MB, Luscher MA, Barber BH, MacDonald KS (2005) Conformational constraints imposed on a pan-neutralizing HIV-1 antibody epitope result in increased antigenicity but not neutralizing responses. Vaccine 23: 1559-1573.

29. Joyce JG, Hurni WM, Bogusky MJ, Garsky VM, Liang X, et al. (2002) Enhancement of α-helicity in the HIV-1 inhibitory peptide DP178 leads to an increased affinity for human monoclonal antibody 2F5 but does not elicit neutralizing antibody responses in vitro. J Biol Chem 277: 45811-45820.

30. Kim M, Qiao Z, Yu J, Montefiori D, Reinherz EL (2007) Immunogenicity of recombinant human immunodeficiency virus type 1-like particles expressing gp41 derivatives in a pre-fusion state. Vaccine 25: 5102-5114.

31. Cardoso RM, Zwick MB, Stanfield RL, Kunert R, Binley JM, et al. (2005) Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41. Immunity 22: 163-73.

32. Ofek G, Tang M, Sambor A, Katinger H, Mascola JR, et al. (2004) Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. J Virol 78: 10724-10737.

33. Saphire EO, Parren PW, Pantophlet R, Zwick MB, Morris GM, et al. (2001) Crystal structure of a neutralizing human IgG against HIV-1: a template for vaccine design. Science 293: 1155-1159.

34. Stanfield RL, Gorny MK, Williams C, Zolla-Pazner S, Wilson IA (2004) Structural rationale for the broad neutralization of HIV-1 by human monoclonal antibody 447-52D. Structure (Camb) 12: 193-204.

35. Zhou T, Xu L, Dey B, Hessell AJ, Van Ryk D, et al. (2007) Structural definition of a conserved neutralization epitope on HIV-1 gp120. Nature 445: 732-737.

36. Beddows S, Schülke N, Kirschner M, Barnes K, Franti M, et al. (2005) Evaluating the immunogenicity of a disulfide-stabilized, cleaved, trimeric form of the envelope glycoprotein complex of human immunodeficiency virus type 1. J Virol 79: 8812-8827.

37. Grunder C, Li Y, Louder M, Mascola J, Yang X, et al. (2005) Analysis of the neutralizing antibody response elicited in rabbits by repeated inoculation with trimeric HIV-1 envelope glycoproteins. Virology 331: 33-46.

38. Kang Y, Andjelic S, Binley JM, Crooks ET, Franti M, et al. (2009) Structural and immunogenicity studies of a cleaved, stabilized envelope trimer derived from subtype A HIV-1. Vaccine 27: 5120-5132.

39. Liao H-X, Sutherland LL, Xia S-M, Brock ME, Scearce RM, et al. (2006) A group M consensus envelope glycoprotein induces antibodies that neutralize subsets of subtype B and C primary viruses. Virology 353: 268-282.

40. Binley JM, Lybarger EA, Crooks ET, Seaman MS, Gray E, et al. (2008) Profiling the specificity of neutralizing antibodies in a large panel of plasmas from patients chronically infected with human immunodeficiency virus type 1 subtypes B and C. J Virol 82: 11651-11668.

41. Dhillon AK, Donners H, Pantophlet R, Johnson WE, Decker JM, et al. (2007) Dissecting the neutralizing antibody specificities of broadly neutralizing sera from human immunodeficiency virus type 1-infected donors. J Virol 81: 6548-6562.

42. Li Y, Migueles SA, Welcher B, Svehla K, Phogat A (2007) Broad HIV-1 neutralization mediated by CD4-binding site antibodies. Nat Med 13: 1032-1034.

43. Sather DN, Armann J, Ching LK, Mavrontoni A, Sellhorn G, et al. (2009) Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. J Virol 83: 757-769.

44. Haynes BF, Fleming J, St Clair EW, Katinger H, Stiegler G, et al. (2005) Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. Science 308: 1906-1908.

45. Haynes BF, Moody MA, Verkoczy L, Kelsoe G, Alam SM (2005) Antibody polyspecificity and neutralization of HIV-1: a hypothesis. Human Antibodies 14: 59-67.

46. Fernando K, Hu H, Ni H, Hoxie JA, Weissman D (2007) Vaccine-delivered HIV envelope inhibits CD4+ T-cell activation, a mechanism for poor HIV vaccine responses. Blood 109: 2538-2544.

47. He B, Qiao X, Klasse PJ, Chiu A, Chadburn A, et al. (2006) HIV-1 envelope triggers polyclonal IgG class switch recombination through a CD40-independent mechanism involving BAFF and C-type lectin receptors. J Immunol 176: 3931-3941.

48. Shan M, Klasse PJ, Banerjee K, Dey AK, Iyer SPN, et al. (2007) HIV-1 gp120 mannoses induce immunosuppressive responses from dendritic cells. PLoS Path 3 (11): e169.

49. Binley JM, Wrin T, Korber B, Zwick MB, Wang M, et al. (2005) Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. J Virol 78: 13232-13252.

50. Walker LM, Phogat SK, Chan-Hui P-Y, Wagner D, Phung P, et al. (2009) Broad and potent neutralizing antibodies from an African donor reveal new HIV-1 vaccine target. Science 326: 285-289.

51. Stamatatos L, Morris L, Burton DR, Mascola JR (2009) Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? Nat Med 15: 866-870.

52. Kulkarni SS, Lapedes A, Tang H, Gnanakaran S, Daniels MG, et al. (2009) Highly complex neutralization determinants on a monophyletic lineage of newly transmitted subtype C human immunodeficiency virus type 1 env clones from India. Virology 385: 505-520.

53. Rong R, Gnanakaran S, Decker JM, Bibollet-Ruche F, Taylor J, et al. (2007) Unique mutational patterns in the envelope α2 amphipathic helix and acquisition of length in gp120 hypervariable domains are associated with resistance to autologous neutralizaton of subtype C human immunodeficiency virus type 1. J Virol 81: 5658-5668.

54. Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, et al. (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science 315: 1583-1586. PMID: 17363674.

55. Carlson J, Kadie C, Mallal S, Heckerman D. (2007) Leveraging hierarchical population structure in descrete association studies. PLoS One. 7:e591. PMID: 17611623.

56. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci USA 105: 7552-7557.

57. Bunnik EM, van Gils MJ, Lobbrecht MSD, Pisas L, Nanlohy NM, et al. (2010) Emergence of b12 resistant human immunodeficiency virus type 1

variants during natural infection in the absence of humoral or cellular immune pressure. J Gen Virol, in press. PMID: 20053822

58. Wu X, Zhou T, O'Dell S, Wyatt RT, Kwong PD, Mascola JR (2009) Mechanism of human immunodeficiency virus type 1 resistance to monoclonal antibody b12 that effectively targets the site of CD4 attachment. J Virol 83: 10892-10907.

59. Huang C-C, Tang M, Zhang M-Y, Majeed S, Montabana E, et al. (2005) Structure of a V3-containing HIV-1 gp120 core. Science 310: 1025-1028.

60. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA (1998) Structure of an HIV-1 gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393: 648-659.

61. Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, et al. (1998) The antigenic structure of the HIV gp120 envelope glycoprotein. Nature 393: 705-11.

62. Mo H, Stamatatos L, Ip JE, Barbas CF, Parren PWHI, et al. (1997) Human immunodeficiency virus type 1 mutants that escape neutralization by human monoclonal antibody IgG1b12. J Virol 71: 6869-6874.

63. Pantophlet R, Saphire EO, Poignard P, Parren PWHI, Wilson I, Burton DR (2003) Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. J Virol 77: 642-658.

64. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98: 10037-10041.

65. He Y, Cheng J, Lu H, Li J, Hu J, et al. (2008) Potent HIV fusion inhibitors against enfuvirtide-resistant HIV-1 strains. Proc Natl Acad Sci USA 105: 16332-16337.

gp120-gp41 interactions:

66. Liu J, Jing W, Cheung B, Lu H, Sun J, et al. (2007) HIV gp41 C-terminal heptad repeat contains multifunctional domains. J Biol Chem 282: 9612-9620.

67. Back NKT, Smit L, Schutten M, Nara PL, Tersmette M, Goudsmit J (1993) Mutations in human immunodeficiency virus type 1 gp41 affect sensitivity to neutralization by gp120 antibodies. J Virol 67: 6897-6902.

68. Blish CA, Nguyen M-A, Overbaugh J (2008) Enhancing exposure of HIV-1 neutralization epitopes through mutations in gp41. PLoS Med. 5:e9.

69. Desmezieres E, Gupta N, Vassell R, He Y, Peden K, et al. (2005) Human immunodeficiency virus (HIV) gp41 escape mutants: cross-resistance to peptide inhibitors of HIV fusion and altrered receptor activation of gp120. J Virol 79: 4774-4781.

70. Klasse PJ, McKeating JA, Schutten M, Reitz MS Jr, Robert-Guroff M (1993) An immune-selected point mutation in the transmembrane protein of human immunodeficiency virus type 1 (HXB2-Env:Ala 582(→ Thr) decreases viral neutralization by monoclonal antibodies to the CD4-binding site. Virology 196: 332-337.

71. Park EJ, Gorny MK, Zolla-Pazner S, Quinnan GV Jr (2000) A global neutralization resistant phenotype of human immunodeficiency virus type 1 is determined by distinct mechanisms mediating enhanced infectivity and conformational changes of the envelope complex. J Virol 74: 4183-4191.

72. Park EJ, Vujcic LK, Anand R, Theodore TS, Quinnan GV (1998) Mutations in both gp120 and gp41 are responsible for the broad neutralization resistance of variant human immunodeficiency virus type 1 MN to antibodies directed at V3 and non-V3 epitopes. J Virol 72: 7099-7107.

73. Reitz MS, Wilson C, Naugle C, Gallo RC, Robert-Guroff M (1988) Generation of a neutralization resistant variant of HIV-1 is due to selection for a point mutation in the envelope gene. Cell 54: 57-63.

74. Thali M, Charles M, Furman C, Cavacini L, Posner M, et al. (1994) Resistance to neutralization by broadly reactive antibodies to the human immunodeficiency virus type 1 gp120 glycoprotein conferred by a gp41 amino acid change. J Virol 68: 674-680.

75. Watkins BA, Buge S, Aldrich K, Davis AE, Robinson J, et al. (1996) Resistance of human immunodeficiency virus type 1 to neutralization by natural antisera occurs through single amino acid substitutions that cause changes in antibody binding at multiple sites. J Virol 70: 8431-8437.

76. O'Rourke SM, Schweighaerdt B, Scott WG, Wrin T, Fonsceca DPAJ et al. (2009) Novel ring structure in the gp41 trimer of human immunodeficiency virus type 1 that modulates sensitivity and resistance to broadly neutralizing antibodies. J Virol 83: 7728-7738.

77. Cao J, Bergeron L, Helseth E, Thali M, Repke H, Sodroski J (1993) Effects of amino acid changes in the extracellular domain of the human

immunodeficiency virus type 1 gp41 envelope glycoprotein. J Virol 67: 2747-2755.

78. Helseth E, Olshevsky U, Furman C, Sodroski J (1991) Human immunodeficiency virus type 1 gp120 envelope glycoprotein regions important for association with the gp41 transmembrane glycoprotein. J Virol 65: 2119-2123.

79. Ivey-Hoyle M, Clark RK, Rosenberg M (1991) The N-terminal 31 amino acids of human immunodeficiency virus type 1 envelope protein gp120 contain a potential gp41 contact site. J Virol 65: 2682-2685.

80. Jacobs A, Sen J, Rong L, Caffrey M (2005) Alanine scanning mutants of the HIV gp41 loop. J Biol Chem 280: 27284-27288.

81. Leavitt M, Park EJ, Sidrov IA, Dimitrov DS, Quinnan GV Jr (2003) Concordant modulation of neutralization resistance and high infectivity of the primary human immunodeficiency virus type 1 MN strain and definition of a potential gp41 binding site in gp120. J Virol 77: 560-570.

82. Poumbourios P, Maerz AL, Drummer HE (2003) Functional evolution of the HIV-1 envelope glycoprotein 120 association site of glycoprotein 41. J Biol Chem 278: 42149-42160.

83. Wang J, Sen J, Rong L, Caffrey M (2008) Role of the HIV gp120 conserved domain 1 in processing and viral entry. J Biol Chem 283: 32644-32649.

84. Wilson C, Reitz MS, Aldrich K, Klasse PJ, Blomberg J, et al. (1990) The site of an immune-selected point mutation in the transmembrane protein of human immunodeficiency virus type 1 does not constitute the neutralization epitope. J Virol 64: 3240-3248.

85. Decker JM, Bibollet-Ruche F, Wei X, Wang S, Levy DN, et al. (2005) Antigenic conservation and immunogenicity of the HIV coreceptor binding site. J Exp Med 201: 1407-1419.

86. Gray ES, Taylor N, Wycuff D, Moore PL, Tomaras GD, et al. (2009) Antibody specificities associated with neutralization breadth in plasma from human immunodeficiency virus type 1 subtype C-infected blood donors. J Virol 83: 8925-8937.

87. Gray ES, Madiga MC, Moore PL, Mlisana K, Abdool Karim SS, et al. (2009) Broad neutralization of human immunodeficiency virus type 1 mediated by plasma antibodies against the gp41 membrane proximal external region. J Virol 83: 11265-11274.

88. Nandi A, Lavine CL, Wang P, Lipchina I, Goepfert PA, et al. (2010) Virology 396: 339-348.

89. Derdeyn CA, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, et al. (2004) Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science 303: 2019-22.

90. Li B, Decker JM, Johnson RW, Bibollet-Ruche F, Wei X, et al. (2006) Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1. J Virol 80: 5211-5218

91. Sagar M, Wu X, Lee S, Overbaugh J (2006) Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection and these modifications affect antibody neutralization sensitivity. J Virol 80: 9586-9598.

92. Wei X, Decker JM, Wang S, Hui H, Kappes JC, et al. (2003) Antibody neutralization and escape. Nature 422: 307-312.

93. Hoffman NG, Seillier-Moiseiwitsch F, Ahn J, Walker JM, Swanstrom R (2002) Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. J Virol 76: 3852-3864.

94. Li Y, Svehla K, Louder MK, Wycuff D, Phogat S, et al. (2009) Analysis of neutralization specificities in polyclonal sera derived from human immunodeficiency virus type 1-infected individuals. J Virol 83: 1045-1059.

95. Rizzuto C, Sodroski J (2000) Fine definition of a conserved CCR5-binding region on the human immunodeficiency virus type 1 glycoprotein 120. AIDS Res Hum Retroviruses 16: 741-749.

96. Rizzuto CD, Wyatt R, Hernández-Ramos N, Sun Y, Kwong PD, et al. (1998) A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 280: 1949-1953.

97. Xiang S-H, Wang L, Abreu M, Huang C-C, Kwong PD, et al. (2003) Epitope mapping and characterization of a novel CD4-induced human monoclonal antibody capable of neutralizing primary HIV-1 strains. Virology 315: 124-134.

98. Xiang S-H, Doka N, Choudhary RK, Sodroski J, Robinson JE (2002) Characterization of CD4-induced epitopes on the HIV type 1 gp120 envelope glycoprotein recognized by neutralizing human monoclonal antibodies. AIDS Res Hum Retroviruses 18: 1207-1217.

99. Pastore C, Nedellec R, Ramos A, Pontow S, Ratner L, Mosier DE (2006) Human immunodeficiency virus type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations. J Virol 80: 750-758.

100. Chen L, Kwon YD, Zhou T, Wu X, O'Dell S, et al. (2009) Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. Science 326: 1123-1127.

101. Chen B, Vogan EM, Gong H, Skehel JJ, Wiley DC, Harrison SC (2005) Structure of an unliganded simian immunodeficiency virus gp120 core. Nature 433: 834-41.

102. Salzwedel K, Smith ED, Dey B, Berger EA (2000) Sequential CD4-coreceptor interactions in human immunodeficiency virus type 1 Env function: soluble CD4 activates Env for coreceptor-dependent fusion and revelas blocking activities of antibodies against cryptic conserved epitopes on gp120. J Virol 74: 326-333.

103. Thali M, Moore JP, Furman C, Charles M, Ho DD, et al. (1993) Characterization of conserved human immunodeficiency virus type 1 gp120 neutralization epitopes exposed upon gp120-CD4 binding. J Virol 67: 3978-3988.

104. Kolchinsky P, Kiprilov E, Sodroski J (2001) Increased neutralization sensitivity of CD4-independent human immunodeficiency virus variants. J Virol 75: 2041-2050.

105. Zhang PF, Bouma P, Park EJ, Margolick JB, Robinson JE, et al. (2002) A variable region 3 (V3) mutation determines a global neutralization phenotype and CD4-independent infectivity of a human immunodeficiency virus

type 1 envelope associated with a broadly cross-reactive, primary virus-neutralizing antibody response. J Virol 76: 644-655.

106. Labrijn AF, Poignard P, Raja A, Zwick MB, Delgado K, et al. (2003) Access of antibody molecules to the conserved coreceptor binding site on glycoprotein gp120 is sterically restricted on primary human immunodeficiency virus type 1. J Virol 77: 10557-10656.

107. DeVico A, Fouts T, Lewis GK, Gallo RC, Godfrey K, et al. (2007) Antibodies to CD4-induced sites in HIV gp120 correlate with the control of SHIV challenge in macaques vaccinated with subunit immunogens. Proc Natl Acad Sci USA 104: 17477-17482.

108. Kowalski M, Potz J, Basiripour L, Dorfman T, Goh WC, et al. (1987) Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1. Science 237: 1351–1355.

109. Olshevsky U, Helseth E, Furman C, Li J, Haseltine W, Sodroski J (1990) Identification of individual human immunodeficiency virus type 1 gp120 amino acids important for CD4 receptor binding. J Virol 64: 5701-5707.

110. Li M, Gao F, Mascola JR, Stamatatos L, Polonis VR, et al. (2005) Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. J Virol 79: 10108-10125.

111. Li M, Salazar-Gonzalez JF, Derdeyn CA, Morris L, Williamson C, et al. (2006) Genetic and neutralization properties of subtype C human immunodeficiency virus type 1 molecular env clones from acute and early heterosexually acquired infections in southern Africa. J Virol 80: 11776-11790.

112. Blish CA, Nedellec R, Mandaliya K, Mosier DE, Overbaugh J (2007) HIV-1 subtype A envelope variants from early in infection have variable sensitivity to neutralization and to inhibitors of viral entry. AIDS 21: 693-702.

113. Abrahams J, Anderson A, Giorgi EE, Seoighe C, Mlisana K, et al. (2009) Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. J Virol 83: 3556-3567.

114. Seaman MS, Janes H, Hawkins N, Grandpre LE, Devoy C, et al. (2010) Tiered categorization of a diverse panel of HIV-1 Env pseudoviruses for assessments of neutralizing antibodies. J Virol 84: 1439-1452.

115. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, et al. (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol 82: 3952-3970.

116. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, et al. (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J Clin Microbiol 43: 406-413.

117. Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

118. Faulkner DM, Jurka J (1988) Multiple Alignment Sequence Editor (MASE). Trends Biochem Sci 13: 321-322.

119. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, et al. (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 73: 152-160.

120. Doria-Rose NA, Klein RM, Daniels MG, O'Dell S, Nason M, et al. (2010) Breadth of human immunodeficiency virus-specific neutralizing activity in sera: clustering analysis and association with clinical variables. J Virol 84: 1631-1636. PMID: 19923174.

121. Felsenstein J (1985) Phylogenies amd the comparative method. Am Nat 125: 1-15.

122. Felsenstein J (2002) In "Modern Developments in Theoretical Population Genetics: The legacy of Gustave Malecot." M. Slatkin and M. venitte, Eds. (Oxford University Press, Oxford), pp. 118-129.

123. Cover TM, Thomas JA (1991) "Elements of Information Theory", Wiley Series in Telecommunications, John Wiley & Sons, New York.

124. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440-9445.

125. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and Regression Trees. Wadsworth.

126. Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning 36: 105–139

127. Kwong PD, Wyatt R, Majeed S, Robinson J, Sweet RW, et al. (2000) Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. Structure Fold Des 8: 1329-1339.

128. Blay WM, Gnanakaran S, Foley B, Doria-Rose NA, Korber BT, Haigwood NL (2006) Consistent patterns of change during the divergence of human immunodeficiency virus type 1 envelope from that of the inoculated virus in simian/human immunodeficiency virus-infected macaques. J Virol 80: 999-1014.

129. Caffrey M, Cai M, Kaufman J, Stahl SJ, Wingfield PT, et al. (1998) Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. EMBO J 17: 4572-4584.

130. Humphrey W, Dalke A, Schulten K (1996) VMD—Visual Molecular Dynamics. J Mol Graphics 14: 33-38.

131. Nelson JD, Brunel FM, Jensen R, Crooks ET, Cardoso MF, et al. (2007) An affinity-enhanced neutralizing antibody against the membrane-proximal external region of human immunodeficiency virus type 1 gp41 recognizes an epitope between those of 2F5 and 4E10. J Virol 81: 4033-4043.

TABLES

Table 1. Sites identified as signatures of b12 sensitivity using any of the three signature-defining approaches: contingency table, CMI, and ensemble machine learning method.

HXB2 position	Signature Region	CMI[1]	Fisher's[2] Sensitive/Resistant	Recurrent top splits in decision trees[3]
gp120				
163	V2	Yes	-	-
173	V2	Yes	**Y/HS**	**Y→!Y** (32)
182	V2	Yes	-	-
185	V2	Yes	**DEN/GST**	**D→!D** (38), **!D→D** (59)
268	outer domain	Yes	**ES/KR**	**E→!E** (83)
364	b12	No	**PS/AH**	-
369	b12	No	**AP/ILQ**	-
461	b12	No	**EP**	**E→!E** (61)
gp41				
651	C-heptad repeat	No	**N/DIS**	**N→!N** (17)
655	C-heptad repeat	Yes	-	-

[1]The CMI approach does not provide specific information regarding which amino acids give rise to the signal, although particularly distinctive substitutions can be seen by examining the data (Fig. 9). A "Yes" in the CMI column means the site was associated with b12 sensitivity or resistance.

[2]The Fisher's exact contingency table is based on specific amino acids or sets of amino acids, such that the amino acids associated with signature sites are explicit; and amino acids associated with b12 resistance are underlined, whereas amino acids associated with b12 susceptibility are *not* underlined.

[3]The recurrent top splits in the decision trees (the number in parentheses indicates how many times it was found) provide information about the key signature amino acid substitutions. The exclamation point (!) means "not" in these tables and figures, thus E → !E means that "E" is found in the immediate ancestral state of the sequence, and is "not E" in the sequence.

Table 2. Summary of statistics of signature sites of b12 sensitivity.

HXB2 position[1]	Amino acid[2]	Statistic[3]	p-value[4]	q-value[4]	Odds ratio[4]	Counts				Strength[6]	Test[7]
						r1c1 Sensitive Change[5]	r1c2 Sensitive Stable[5]	r2c1 Resistant Change[5]	r2c2 Resistant Stable[5]		
163	NA	CMI	$< 10^{-3}$	$< 10^{-3}$							1aa
173	Y→!Y	Fisher	0.00024	0.042	0.14	6	74	40	105	0.2413	1aa
173	!HS→HS	Fisher	0.0017	0.0087	0.18	3	78	27	123	0.2242	>1aa
173	NA	CMI	0.001	0.013							1aa
182	NA	CMI	0.002	0.12							1aa
185[8]	**!D→D**	**Fisher**	**0.00036**	**0.04**	**6.98**	**11**	**25**	**6**	**97**	**6.4615**	**1aa**
185	**D→!D**	**Fisher**	**0.00053**	**0.059**	**0.24**	**13**	**39**	**35**	**25**	**0.2528**	**1aa**
185	DEN→!DEN	Fisher	4.4×10^{-7}	0.00013	0.109	4	82	48	106	0.1315	>1aa
185	!GST→GST	Fisher	4.0×10^{-5}	0.00088	0.05	1	85	28	128	0.1034	>1aa
185	NA	CMI	$< 10^{-3}$	$< 10^{-3}$							
268	E→!E	Fisher	0.00011	0.033	0.24	9	68	50	90	0.2586	1aa
268	!K→K	Fisher	0.00026	0.088	0.68	1	83	24	134	0.1286	1aa
268	ES→!ES	Fisher	4.7×10^{-05}	0.00006	0.21	8	69	50	90	0.2294	>1aa
268	!KR→KR	Fisher	7.8×10^{-05}	0.0011	0.06	1	83	27	131	0.1122	>1aa
268	NA	CMI	$< 10^{-3}$	$< 10^{-3}$							
364	!AH→AH	Fisher	0.0049	0.052	0.16	2	82	21	133	0.2202	b12
364	PS→!PS	Fisher	0.0018	0.0085	0.13	2	84	24	134	0.1906	b12
369	**AP→!AP**	**Fisher**	**0.0048**	**0.017**	**0.077**	**1**	**37**	**9**	**25**	**0.1368**	**b12**
369	**!ILQ→ILQ**	**Fisher**	**0.0013**	**0.047**	**0**	**0**	**37**	**8**	**24**	**0.0731**	**b12**
461	**!E→E**	**Fisher**	**0.00026**	**0.045**	**8.77**	**12**	**60**	**3**	**133**	**7.1393**	**1aa**
461	**!EP→EP**	**Fisher**	**4.5×10^{-5}**	**0.0009**	**8.59**	**15**	**57**	**4**	**132**	**7.3379**	**>1aa**
651	N→!N	Fisher	0.0007	0.064	0.24	6	68	38	101	0.2653	1aa
651	!DIS→DIS	Fisher	0.0082	0.025	0.25	2	76	20	126	0.2356	>1aa
655	NA	CMI	0.002	0.12							
149[9]	Nx[ST]→Nx[ST]	Fisher	0.0044	0.22							PNLG
V2[10]	Shorter length	Spearman	0.021	0.08							
V5[10]	Fewer PNLGs	Spearman	0.0065	0.065							

[1]**HXB2 position** refers to the amino acid position of interest in the HXB2 reference strain (www.hiv.lanl.gov: Locator tool).

[2]**Amino acid** refers to the particular amino acid or combination of amino acids that was statistically related to b12 resistance (underlined) or sensitivity (not underlined). An exclamation point means "not"; thus in the first line, when T is an ancestral state, Y mutates to "not Y" (!Y) with a statistically higher frequency in b12 resistant strains than sensitive strains.

[3]**Statistic** is the statistic that was used to identify the signature, by either the phylogenetically corrected contingency approach (Fisher exact test) employed as described in [54]; the conditional mutual information approach (CMI); or a comparison of all variable region loop lengths (length) and number of glycosylation sites (sequons with amino acid pattern Nx[ST]) with the b12 neutralization values using a Spearman rank correlation test.

[4]The **p-values**, the **q-values** (false positive rates), and the **odds ratios** are provided. The Fisher's exact test q-values were calculated for discrete tests as implemented in [54]. For the CMI analyses, p-values were acquired by shuffling phenotypes and counting the relative frequency at which random CMIs exceeded the original CMI. The q-values were calculated using the method of [124], after stripping off the highest p values (essentially a few hundred of p-value = 1). Only associations with a q-value < 0.2 are shown.

[5]Rows and columns of the 2x2 contingency table. As an example of how to read these, in position 173, **r1c1** refers to row 1 column 1 and is the number of times among b12 sensitive viruses that Y→!Y mutates to another amino acid (change). **r1c2** refers to row 1 column 2, and it is the number of times among sensitive viruses that the ancestral state was Y and it stayed Y (stable) in the Env sequence.

[6]**Strength** is a measure that expresses how predictive a given signature amino acid is of the b12 sensitive/resistant phenotype, essentially an augmented odds ratio, where each count was augmented by 1 pseudo-count to avoid issues with zeros and infinities, and strength = $(r1c1+1)(r2c2+1)/(r1c2+1)(r2c1+1)$.

[7]Several explorations of the Env alignment were used, and this is described in the "**test**". In the first screen, every amino acid found in every column was tested (1aa). Then combinations of 2 or more amino acids in every column were tested (>1aa). Then positions known to be key for the b12 binding site (Table 9) were specifically tested for all combinations of amino acids over all pairs of positions in the binding site (b12). Although pairs of positions were tested, single positions essentially accounted for the signal in that analysis. Only these single site associations are shown.

[8]Some lines are shown in bold. In these lines, the change in the amino acid is associated with a reverse in the majority of cases found among sensitive or resistant viruses; thus the change in theses sites is particularly predictive of NAb phenotype.

[9]All PNLG sites in Env were tested for phylogenetically corrected association with b12 sensitivity using the contingency table approach. None reached significance with a q-value of < 0.2; the glycosylation site at position 149 was the only one to reach even borderline significance and is included here for completeness.

[10]For the initial analysis of loop length and number of PNLGs in each loop, a phylogenetically corrected method was not used. Rather, a non-parametric Spearman's correlation test was used comparing loop length with the geometric mean 50% neutralization titer for the 25 Envs. It is reasonable to forego the phylogenetic correction in these cases because the loop lengths vary by insertion and deletion and often change dramatically within infected individuals. These parameters are less likely to be biased by phylogeny at the population level.

Table 3. Prediction strategies for b12 sensitivity applied to the 251 pseudotyped Envs included in the signature-defining training set

Sensitive Envelopes:	**Total**	**Correct**	**Incorrect**	**Sensitivity**
Signature rule	88	67	21	0.76
Logistic regression	88	53	35	0.60
Ensemble Learning Techniqu	88	64	24	0.73
Sensitive Envelopes:	**Total**	**Correct**	**Incorrect**	**Specificity**
Signature rule	163	110	53	0.67
Logistic regression	163	134	29	0.82
Ensemble Learning Techniqu	163	153	10	0.94
Summary of all Envelopes:	**Total**	**Correct**	**Accuracy**	**Fischer's p-value**
Signature rule	251	177	0.71	2.4×10^{-11}
Logistic regression	251	187	0.74	1.4×10^{-11}
Ensemble Learning Techniqu	251	217	0.86	$<2.2 \times 10^{-16}$

Table 4. Prediction strategies for b12 sensitivity applied to the 56 pseudotyped Envs included in the blinded test set

Sensitive Envelopes:	**Total**	**Correct**	**Incorrect**	**Sensitivity**
Signature rule	20	13	7	0.65
Logistic regression	20	9	11	0.45
Ensemble Learning Techniqu	20	5	15	0.25
Sensitive Envelopes:	**Total**	**Correct**	**Incorrect**	**Specificity**
Signature rule	36	26	10	0.72
Logistic regression	36	27	9	0.75
Ensemble Learning Techniqu	36	29	7	0.72
Summary of all Envelopes:	**Total**	**Correct**	**Accuracy**	**Fischer's p-value**
Signature rule	56	39	0.70	0.007
Logistic regression	56	36	0.64	0.11
Ensemble Learning Techniqu	56	34	0.61	0.44

Table 5. Sites identified as Env signatures associated with serum neutralizing breadth and potency using the tree corrected contingency table and CMI approaches. This table is organized similarly to Table 1.

HXB2 Position	Signature region[1]	CMI	Fisher's[2] Sensitive/Resistant [2]
412/413	CoRbs	-	![GS]N→[GS]N
413	CoRbs	-	Nx[ST]
419_421	CoRbs	-	R_K→!R_K
419	CoRbs	-	R→K
440	CoRbs	-	Q→!Q
186	V2	Y	

[1]All regions are in gp120. CoRbs, coreceptor binding site; V2, second variable region.

[2]Arrows are used to show the direction of the sequence change that was significant. Thus, in the three amino acids at positions 419-421, the sequence was moving from R, any amino acid, K (R_K) to a sequence that was not R, any amino acid, K (!R_K) in weakly neutralizing sera. [GS]N, means either G or S at position 412 and N at 413.

Table 6. Summary of statistics of signature sites of associated with serum neutralizing breadth and potency. This table is organized similarly to Table 2.

HXB2 Position	Amino Acid	Statistic	p-value	q-value	Odds ratio	Counts				Strength	Test[2]
						r1c1 Sensitive Change	r1c2 Sensitive Stable	R2c1 Resistant Change	r2c2 Resistant Stable		
412/413	**![GS]N→[GS]N**	**Fisher**	**2.1 x 10^{-6}**	**0.0015**	**Infinity**	**6**	**4**	**0**	**57**	**81.2**	**2 sites, Full Env 2 deep, k=3, high vs others**
413	**Nx[ST]→!Nx[ST]**	**Fisher**	**0.0083**	**0.23**	**10.17**	**5**	**2**	**10**	**43**	**8**	**PNLGs, Full Env, k=3, high vs other**
419_421	**R_K→!R_K**	**Fisher**	**0.0025**	**0.089**	**9.2**	**13**	**17**	**2**	**25**	**6.7407**	**3 sites, 2 deep CoRbs k= 3, low vs others**
419_421	**R_K→!R_K**	**Fisher**	**0.0013**		**8.1**	**13**	**17**	**3**	**33**	**6.6111**	**3 sites, 2 deep** CoRbs, **k= 2 low vs high**
419	R→!R	Fisher	0.044	0.11	5.2	9	21	2	25	3.9394	1 site, CoRbs, k= 3, low vs others
419	!K→K	Fisher	0.044	0.11	5.2	9	21	2	25	3.9394	1 site, CoRbs, k= 3, low vs others
440	Q→!Q	Fisher	0.018	0.11	0	0	5	3	0	0.0417	3 sites, CoRbs, k= 3, low vs others
440	Q→!Q	Fisher	0.012	0.13	Infinity	3	0	0	6		3 sites, CoRbs, k= 2, high vs low
186	NA	CMI	< 0.001	<0.001							
V2	Shorter length	Spearman	0.043	0.14							
V2	Fewer PNLGs	Spearman	0.017	0.043							
V2	Fewer PNLGs	[1]Contrasts	0.06	0.06							
V5	Shorter length	[1]Contrasts	0.02	0.02							

[1]Variable loop lengths and the number of glycosylations sites in each variable loop were compared as in Table 1B, using a simple Spearman's rho test.. These results were validated using a phylogenetically corrected method, phylogenetic contrasts [121,122].

[2]In this column the number of "sites" refers to the number of sites considered in combination in each test, the number "deep" refers to how amino acids at a single site were combined in each test. k = 2 or 3 refers to the k-means clusters as illustrated in Fig. 5. When k = 3, the test could either be the lowest or the highest neutralization potency cluster versus all others. When k= 2, only the high and the low clusters were compared, excluding indeterminate values. Full Env means the complete Env was scanned in the test. CoRbs means the signature was defined in the in-depth scan of the CoRbs. No significant signatures were found in comparable in-depth scans of the CD4bs and the MPER regions.

Name	Clade	Country of origin	Year	Fiebig Stage	Mode of transmission	Specimen Source	Accession number	ARRRP cat #
Q23_17_	A	Kenya		VI	M-F		AF004885	10455
Q259_d2_17_	A	Kenya		Acute/early	M-F		AF407152	10459
Q769_d22_	A	Kenya		Acute/early	M-F		AF407158	10458
Q842_d12_	A	Kenya		Acute/early	M-F		AF407160	10457
MS208_A1	A	Montserrat		VI			DQ187010	
3415_v1_c1	A	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
0260_v5_c1	A	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
0330_v4_c3	A	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
3365_v2_c20	A	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
783_v0_c51	A	Tanzania	2002	I or II	Hetro	Plasma	submitted	
3718_v3_c11	A	Tanzania	2004	I/II	Hetero	Plasma	submitted	
398_F1_F6_20	A	Tanzania		ND	Hetero	Plasma	submitted	
191955_A11	A	Uganda	2007	IV	heterosexual	plasma	submitted	
9004SS_A3_4	A	Uganda	2007	IV	Hetero	Plasma	submitted	
T280_5	A/CRF02_AG	Cameroon		Chronic		ccPBMC	EU513183	
Q461_e2_	AD	Kenya		Acute/early	M-F		AF407156	10460
0907_v4_c12	AD	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
3468_v1_c12	AD	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
191084_B7_19	A1	Uganda	2007	IV	Hetero	Plasma	submitted	
3301_v2_c6	AC	Tanzania	2004	early	Hetro	Plasma	submitted	
3301_v1_c24	AC	Tanzania	2003	I/II	Hetero	Plasma	submitted	
3589_v1_c4	AC	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
6540_v4_c1 *	AC	Tanzania	2004	early	Hetro	Plasma	submitted	
6041_v3_c23	AC	Tanzania	2004	I or II	Hetro	Plasma	submitted	
6545_v3_c13 *	AC	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
6545_v4_c1 *	AC	Tanzania	2005	early	Hetro	Plasma	submitted	
477_F3_13_55	AC	Tanzania		VI	Hetero	Plasma	submitted	
246_F3_C10_2	AC	Tanzania		VI	Hetero	Plasma	submitted	
6095_v1_c10	ACD	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
0815_v3_c3	ACD	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
3103_v3_c10	ACD	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
270015_J5_1	AD	Uganda	2006	VI	Hetero	Plasma	submitted	
192018_B1_9	AD	Uganda	2007	IV	Hetero	Plasma	submitted	
193003_B10	AD	Uganda	2007	VI	heterosexual	plasma	submitted	
3233_p12	AG	Senegal	2001	VI		Plasma	submitted	
1105_p17_1	AG	Senegal	1999	VI		Plasma	submitted	
2705_p18_1	AG	Senegal	2000	VI		Plasma	submitted	
2843_p5_1	AG	Senegal	1999	VI		Plasma	submitted	
3169_p4	AG	Senegal	1999	VI		Plasma	submitted	
3226_p15	AG	Senegal	1999	VI		Plasma	submitted	
3273_p21_1	AG	Senegal	1999	VI		Plasma	submitted	
B01	B	China/Hebei	2003	V/VI	blood	plasma	EU363825	
B03	B	China/Hebei	2003	V/VI	Sexual	plasma	EU363827	
B04	B	China/Hebei	2006	V/VI	blood	plasma	EU363828	
BZ167_12	B	Brazil		VI		ccPBMC	GQ855764	
BJOX003000_19_1	B	China/Beijing	2007	II	homosex.	plasma	submitted	

BJOX020000_03_2	B	China/Beijing	2007	II	homosex.	plasma	submitted	
B02	B	China/Gansu	2003	V/VI	blood	plasma	EU363826	
B05	B	ChinaHubei	2006	V/VI	Blood	plasma	EU363829	
HXB2	B	France		Chronic	M-M		K03455	
Bx08_16	B	France		VI		ccPBMC	GQ855765	
PVO_4_	B	Italy	1994	III	M-M	ccPBMC	AY835444	11022
TRO_11_	B	Italy	1995	III	M-M	ccPBMC	AY835445	11023
H022_7	B	Peru		VI	Sexual	ccPBMC	EF210725	
H029_12	B	Peru		VI	Sexual	ccPBMC	EF210726	
H030_7	B	Peru		VI	Sexual	ccPBMC	EF210727	
H031_7	B	Peru		VI	Sexual	ccPBMC	EF210728	
H035_18	B	Peru		VI	Sexual	ccPBMC	EF210729	
H061_14	B	Peru		VI	Sexual	ccPBMC	EF210730	
H079_2	B	Peru		VI	Sexual	ccPBMC	EF210731	
H086_8	B	Peru		VI	Sexual	ccPBMC	EF210732	
H078_14	B	Peru		VI	Sexual	ccPBMC	EF210733	
H077_31	B	Peru		VI	Sexual	ccPBMC	EF210734	
H080_23	B	Peru		VI	Sexual	ccPBMC	EF210735	
NKR_0512_8	B	Thailand				ccPBMC	submitted	
RPW_0510_2	B	Thailand				ccPBMC	submitted	
QH0692_42_	B	Trinidad	1994	V	F-M	ccPBMC	AY835439	11018
QH0515_1	B	Trinidad	1994	IV	F-M	ccPBMC	AY835440	
SC422661_8_	B	Trinidad	1995	IV	F-M	Plasma	AY835441	11058
SC05_8C11_2344	B	Trinidad	1993	II	Heterosexual	Plasma	EU289200	11576
SC45_4B5_2631	B	Trinidad	1995	II	Heterosexual	Plasma	EU289201	11577
TT29P_3A1_2769	B	Trinidad	1998		Heterosexual	Plasma	EU577190	
TT31P_2F10_2792	B	Trinidad	1998	II	Heterosexual	Plasma	EU577213	
SF162_LS	B	USA		VI			EU123924	10463
Bal_26	B	USA		VI	Mother to Child	Tissue (Lung)	DQ318211	
6101_10	B	USA	1994	V	M-M	ccPBMC	AY835434	
6535_3_	B	USA	1995	V	M-M	ccPBMC	AY835438	11017
SS1196_1	B	USA	1997	V-VI	M-M	ccPBMC	AY835442	
BG1168_1	B	USA	1996	III	M-M	ccPBMC	AY835443	
AC10_0_29_	B	USA	1998	III	M-M	ccPBMC	AY835446	11024
MN_3	B	USA		VI			submitted	
RHPA4259_7_	B	USA	2000	≤V	M-F	Plasma	AY835447	11036
THRO4156_18_	B	USA	2000	II	M-M	Plasma	AY835448	11037
REJO4541_67_	B	USA	2001	II	F-M	Plasma	AY835449	11035
TRJO4551_58_	B	USA	2001	II	M-M	Plasma	AY835450	11034
WITO4160_33_	B	USA	2000	II	F-M	Plasma	AY835451	11033
CAAN5342_A2_	B	USA	2004	≤VI	M-M	Plasma	AY835452	11038
ACH320-W61D-TCLA	B						submitted	
1006_11_C3_1601	B	USA	1997	III		Plasma	EU289183	11560
1012_11_TC21_3257	B	USA	1997	III		Plasma	EU289184	11559
1054_07_TC4_1499	B	USA	1997	II		Plasma	EU289185	11561
1056_10_TA11_1826	B	USA	1998	II		Plasma	EU289186	11562
1058_11_B11_1550	B	USA	1998	IV		Plasma	EU289187	11563
1059_09_A4_1460	B	USA	1998	III		Plasma	EU289188	11564
62357_14_D3_4589	B	USA	1996	II		Plasma	EU289189	11565
6240_08_TA5_4622	B	USA	1995	II		Plasma	EU289190	11567
6244_13_B5_4576	B	USA	1996	II		Plasma	EU289191	11566
63358_04_P3_4013	B	USA	1997	II		Plasma	EU289192	11568

700010040_C9_4520	B	USA	2006	V	MSM	Plasma	EU289193	11569
700010058_A4_4375	B	USA	2006	III	MSM	Plasma	EU289194	11570
9021_14_B2_4571	B	USA	1998	II		Plasma	EU289196	11572
PRB926_04_A9_4237	B	USA	1994	II		Plasma	EU289197	11573
PRB931_06_TC3_4930	B	USA	1995	III		Plasma	EU289198	11574
PRB958_06_TB1_4305	B	USA	2000	III		Plasma	EU289199	11575
WEAU_d15_410_5017	B	USA	1990	II	MSM	Plasma	EU289202	11578
1051_12_C22_3325	B	USA	1997	II		Plasma	EU575148	
1051_12_TD12_3291	B	USA	1997	II		Plasma	EU575170	
9014_01_TB1_4769	B	USA	1997	II		Plasma	EU575786	11571
9020_20_A13_4607	B	USA	1998	II		Plasma	EU575870	
BORI_d9_4D7_1410	B	USA	1990	II	MSM	Plasma	EU576296	
BORI_d9_4F8_1413	B	USA	1990	II	MSM	Plasma	EU576299	
SUMA_d5_8_2	B	USA	1991	II	MSM	Plasma	EU577073	
CNE9	B'	China/Henan	2005	VI	hetero	PBMC	submitted	
CNE11	B'	China/Jilin	2005	VI	blood	PBMC	submitted	
CNE14	B'	China/Jilin	2004	VI	hetero	PBMC	submitted	
CNE12	B'	China/Jilin	2004	VI	hetero	PBMC	submitted	
CNE10	B'	China/Liaoning	2004	VI	blood	PBMC	submitted	
CNE57	B'	China/Liaoning	2005	VI	MSM	PBMC	submitted	
CNE4	B'	China/Yunnan	2006	VI	IDU	PBMC	submitted	
CNE6	B'	China/Yunnan	2006	VI	IDU	PBMC	submitted	
CNE40	B'C	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE7	BC	China/Yunnan	2006	VI	IDU	PBMC	submitted	
CNE16	BC	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE15	BC	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE18	BC	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE67	BC	China/Yunnan	2007	VI	IDU	PBMC	submitted	
92BR025_9	C	Brazil		VI			HIV1U15121	
CNE30	C	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE31	C	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE23	C	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE17	C	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE52	C	China/Yunnan	2007	VI	IDU	PBMC	submitted	

CNE53	C	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE58	C	China/Yunnan	2006	VI	IDU	PBMC	submitted	
HIV_00836_2_5_	C	India	2000	VI	M-F	ccPBMC	EF117265	11499
HIV_001428_2_42_	C	India	2000	IV	M-F	ccPBMC	EF117266	11500
HIV_0013095_2_11_	C	India	2000	IV	M-F	ccPBMC	EF117267	11501
HIV_16055_2_3_	C	India	1999	II	F-M	ccPBMC	EF117268	11502
HIV_16845_2_22_	C	India	2000	V	M-F	ccPBMC	EF117269	11503
HIV_16936_2_21_	C	India	2000	III	F-M	ccPBMC	EF117270	11504
HIV_25710_2_43_	C	India	1999	V	F-M	ccPBMC	EF117271	11505
HIV_25711_2_4_	C	India	1999	III	F-M	ccPBMC	EF117272	11506
HIV_25925_2_22_	C	India	1999	III	F-M	ccPBMC	EF117273	11507
HIV_26191_2_48_	C	India	2000	III	F-M	ccPBMC	EF117274	11508
CenvFs2_pt0682_E4uncp	C	Malawi	2003	I/II	sexual	Plasma	submitted	
703010228_1C4	C	Malawi	2007	I/II	sexual	Plasma	submitted	
CenvFs4_Pt2010_F5	C	Malawi	2005	IV	sexual	Plasma	submitted	
703010217_B6	C	Malawi	2007	V/VI	sexual	Plasma	submitted	
CenvFs2_Pt1086_B2	C	Malawi	2004	I/II	sexual	Plasma	submitted	
703010054_2_A2	C	Malawi	2007	V/VI	sexual	Plasma	submitted	
BF1677F2_613a	C	Malawi	96-04		breastfeeding	Plasma	submitted	
CenvFs2_pt0985_H7uncp	C	Malawi	2004	I/II	sexual	Plasma	submitted	
CenvFs2_Pt1176_A3	C	Malawi	2004	I/II	sexual	Plasma	submitted	
CenvFs2_pt2060_G9uncp	C	Malawi	2005	I/II	sexual	Plasma	submitted	
CenvFs4_Pt0393_C3	C	Malawi	2003	IV	sexual	Plasma	submitted	
BF1266_431a	C	Malawi	96-04	I/II	breastfeeding	Plasma	submitted	
CenvFs2_pt1172_H1uncp	C	Malawi	2004	I/II	sexual	Plasma	submitted	
CenvFs2_pt2103_E8uncp	C	Malawi	2005	I/II	sexual	Plasma	submitted	
MW965_26	C	Malawi		VI	M-F	ccPBMC	U08455	3094
1394C9G1_Rev_	C	Malawi	2004	I or II	Hetero	plasma	submitted	
2010F5_Rev_	C	Malawi	2005	IV	Hetero	plasma	submitted	
7030100542A2_Rev_	C	Malawi	2007	V or VI	Hetero	plasma	submitted	
7030102001E5_Rev_	C	Malawi	2007	I or II	Hetero	plasma	submitted	
703010217B6_Rev_	C	Malawi	2007	V	Hetero	plasma	submitted	
CAP45_2_00_G3_	C	S. Africa	2005	IV	M-F	Plasma	DQ435682	11316
CAP210_2_00_E8_	C	S. Africa	2005	IV	M-F	Plasma	DQ435683	11317
CAP244_2_00_D3	C	S. Africa	2005	V	M-F	Plasma	DQ435684	
Du123_6	C	S. Africa	1998	VI	M-F	ccPBMC	DQ411850	
Du151_2	C	S. Africa	1998	V	M-F	ccPBMC	DQ411851	
Du156_12_	C	S. Africa	1999	≤IV	M-F	ccPBMC	DQ411852	11306
Du172_17_	C	S. Africa	1998	VI	M-F	ccPBMC	DQ411853	11307
Du422_1_	C	S. Africa	1998	V	M-F	ccPBMC	DQ411854	11308
TV1_21	C	S. Africa		VI			submitted	
704010042_2_E5	C	South Africa	2007	IV	sexual	Plasma	submitted	
704010083_B8	C	South Africa	2007	III	sexual	Plasma	submitted	
704809221_1B3	C	South Africa	2007	I/II	sexual	Plasma	submitted	

706010018_2E3	C	South Africa	2007	VI	sexual	Plasma	submitted	
706010164_1A7	C	South Africa	2007	IV	sexual	Plasma	submitted	
7048092211B3_Rev_	C	South Africa	2007	I or II	Hetero	plasma	submitted	
7060101641A7_Rev_	C	South Africa	2007	I or II	Hetero	plasma	submitted	
98_v3_c5	C	Tanzania	2004	early	Hetro	Plasma	submitted	
6644_v2_c33	C	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
0041_v3_c18	C	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
0921_v2_c14	C	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
3168_v4_c10	C	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
3637_v5_c3	C	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
3728_v2_c6	C	Tanzania	2004	III/IV	Hetero	Plasma	submitted	
3873_v1_c24	C	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
6022_v7_c24	C	Tanzania	2006	early	Hetro	Plasma	submitted	
6040_v4_c15	C	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
6322_v4_c1	C	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
6471_v1_c16	C	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
6631_v3_c10	C	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
6785_v5_c14	C	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
6838_v1_c35	C	Tanzania	2003	I/II	Hetero	Plasma	submitted	
6980_v1_c17_	C	Tanzania	2003	early	Hetro	Plasma	submitted	
933_v4_c4	C	Tanzania	2005	early	Hetro	Plasma	submitted	
234_F1_16_57	C	Tanzania		V	Hetero	Plasma	submitted	
410_F2_1_30	C	Tanzania		VI	Hetero	Plasma	submitted	
541_F1_A7_2	C	Tanzania		ND	Hetero	Plasma	submitted	
569_F1_37_10	C	Tanzania		V/VI	Hetero	Plasma	submitted	
98_F4_H5_13	C	Tanzania		V	Hetero	Plasma	submitted	
PWJ_0513_39	C	Thailand				ccPBMC	submitted	
96ZM651_02	C	Zambia		VI			AF286224	
ZM55F_PB28a	C	Zambia	1998	≤VI	M-F	ucPBMC	AY423971	
ZM53M_PB12_	C	Zambia	2000	≤VI	F-M	ucPBMC	AY423984	11313
ZM135M_PL10a_	C	Zambia	1998	≤VI	F-M	Plasma	AY424079	11315
ZM109F_PB4_	C	Zambia	2000	≤VI	M-F	ucPBMC	AY424138	11314
ZM106F_PB9	C	Zambia	1998	≤VI	M-F	ucPBMC	AY424163	
ZM249M_PL1_	C	Zambia	2003	II	F-M	Plasma	DQ388514	11319
ZM197M_PB7_	C	Zambia	2002	≤VI	F-M	ucPBMC	DQ388515	11309
ZM214M_PL15_	C	Zambia	2003	≤VI	F-M	Plasma	DQ388516	11310
ZM233M_PB6_	C	Zambia	2002	≤VI	F-M	ucPBMC	DQ388517	11311
ZM215F_PB8	C	Zambia	2002	≤VI	M-F	ucPBMC	DQ422948	
246F_C1G	C	Zambia	2003	II	M to F	Plasma	submitted	
249M_B10	C	Zambia	2003	IV	F to M	Plasma	submitted	
ZM247v1_Rev_	C	Zambia	2003	II	Hetero	plasma	submitted	
3326_v4_c3	CD	Tanzania	2005	V/VI	Hetero	Plasma	submitted	
6952_v1_c20	CD	Tanzania	2003	early	Hetro	Plasma	submitted	
3337_v2_c6	CD	Tanzania	2004	V/VI	Hetero	Plasma	submitted	
3817_v2_c59	CD	Tanzania	2004	early	Hetro	Plasma	submitted	
6480_v4_c25	CD	Tanzania	2004	early	Hetro	Plasma	submitted	
6650_v1_c8	CD	Tanzania	2003	V/VI	Hetero	Plasma	submitted	
6811_v7_c18	CD	Tanzania	2006	early	Hetro	Plasma	submitted	
401_F1_8_10	CD	Tanzania		VI	Hetero	Plasma	submitted	
252_7	Clade G	W. Afr.		VI			EU513190	

CNE59	CRF_01 AE	China/Yunnan	2006	VI	IDU	PBMC	submitted	
AE03	CRF01_AE	China/Shanghai	2005	V/VI	Sexual	plasma	EU363851	
BJOX005000_09_2	CRF01_AE	China/Beijing	2007	II	homosex.	plasma	submitted	
BJOX009000_02_4	CRF01_AE	China/Beijing	2007	IV	homosex.	plasma	submitted	
BJOX010000_06_2	CRF01_AE	China/Beijing	2007	II	homosex.	plasma	submitted	
BJOX025000_01_1	CRF01_AE	China/Beijing	2007	II	homosex.	plasma	submitted	
BJOX028000_10_3	CRF01_AE	China/Beijing	2007	II	homosex.	plasma	submitted	
AE01	CRF01_AE	China/Guangdong,	1999	V/VI	Blood	plasma	EU363849	
CNE5	CRF01_AE	China/Henan	2006	VI	hetero	PBMC	submitted	
AE02	CRF01_AE	China/Yunnan	2006	V/VI	IDU	plasma	EU363850	
CNE28	CRF01_AE	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE3	CRF01_AE	China/Yunnan	2006	VI	IDU	PBMC	submitted	
CNE55	CRF01_AE	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE56	CRF01_AE	China/Yunnan	2007	VI	IDU	PBMC	submitted	
CNE8	CRF01_AE	China/Yunnan	2006	VI	IDU	PBMC	submitted	
C1080rsga_c3	CRF01_AE	Thailand	1999	>VI	Hetero	Plasma	submitted	
R1166rsga_c1	CRF01_AE	Thailand	1998	>VI	Hetero	Plasma	submitted	
R2184rsga_c4	CRF01_AE	Thailand	2001	>VI	Hetero	Plasma	submitted	
R3265rsga_c6	CRF01_AE	Thailand	1999	>VI	Hetro	Plasma	submitted	
KSS_0514_13	CRF01_AE	Thailand				ccPBMC	submitted	
PSR_0508_2	CRF01_AE	Thailand				ccPBMC	submitted	
SPK_0525_13	CRF01_AE	Thailand				ccPBMC	submitted	
T266_60	CRF02_AG	Cameroon		Chronic		ccPBMC	EU513193	
T278_50_	CRF02_AG	Cameroon		Chronic		ccPBMC	EU513198	
DJ263_8	CRF02_AG	Djibouti		VI	Sexual	ccPBMC	AF063223	
263_8_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513182	
T255_34_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513184	
T257_31_	CRF02_AG[1]	Cameroon		Acute/early		ccPBMC	EU513185	
T33_7_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513186	
211_9_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513187	
242_14_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513188	
T250_4_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513189	
T253_11	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513191	
269_12	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513194	
235_47_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513195	
T251_18_	CRF02_AG[1]	Cameroon		Chronic		ccPBMC	EU513196	
271_11_	CRF02_AG[1]	Cameroon		Acute/early		ccPBMC	EU513197	

928_28_	CRF02_AG[1]	Cote d'Ivoire		Acute/early		ccPBMC	EU513199	
1656_p21	CRF06_cpx	Senegal	1999	VI		Plasma	submitted	
CH038_12	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF042692	
CH064_20	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117254	
CH070_1	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117255	
CH091_9	CRF07_BC	China	2003	VI	IVDU	ccPBMC	EF117256	
CH110_2	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117257	
CH111_8	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117258	
CH181_12	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117259	
CH120_6	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117260	
CH119_10	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117261	
CH117_4	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117262	
CH115_12	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117263	
CH114_8	CRF07_BC	China	2004	VI	IVDU	ccPBMC	EF117264	
BC14	CRF07_BC	China/Beijing	2003	V/VI	Sexual	plasma	EU363844	
BC15	CRF07_BC	China/Beijing	2003	V/VI	Sexual	plasma	EU363845	
BJOX002000_03_2	CRF07_BC	China/Beijing	2007	I /II	homosex.	plasma	submitted	
BC04	CRF07_BC	China/Sichuan	2007	V/VI	IDU	plasma	EU363834	
BC05	CRF07_BC	China/Sichuan	2007	V/VI	IDU	plasma	EU363835	
BC03	CRF07_BC	China/Xinjiang	2003	V/VI	IDU	plasma	EU363833	
BC18	CRF07_BC	China/Xinjiang	2005	V/VI	IDU	plasma	EU363848	
CNE20	CRF07_BC	China/Xinjiang	2007	VI	hetero	PBMC	submitted	
CNE19	CRF07_BC	China/Xinjiang	2007	VI	hetero	PBMC	submitted	
CNE21	CRF07_BC	China/Xinjiang	2007	VI	hetero	PBMC	submitted	
BC07	CRF07_BC	China/Yunnan	2005	V/VI	IDU	plasma	EU363837	
BC09	CRF07_BC	China/Yunnan	2003	V/VI	Sexual	plasma	EU363839	
BC10	CRF07_BC	China/Yunnan	2005	V/VI	IDU	plasma	EU363840	
BC01	CRF08_BC	China/Yunnan	2006	V/VI	IDU	plasma	EU363831	
BC06	CRF08_BC	China/Yunnan	2005	V/VI	IDU	plasma	EU363836	
BC08	CRF08_BC	China/Yunnan	2006	V/VI	IDU	plasma	EU363838	
BC11	CRF08_BC	China/Yunnan	2005	V/VI	IDU	plasma	EU363841	
BC17	CRF08_BC	China/Yunnan	2006	V/VI	IDU	plasma	EU363847	
X2252_c7	CRF14_BG	Portugal	2007	Chronic	Hetero	Plasma	EU885766	
X1100_c7	CRF14_BG	Switzerland	2002	NA	IDU	Plasma	EU885760	
3016_v5_c45	D	Tanzania	2005	I or II	Hetro	Plasma	submitted	
6405_v4_c34	D	Tanzania	2004	early	Hetro	Plasma	submitted	

A07412M1_vrc12	D	Uganda	1999	n/a		ccPBMC	submitted	
X2088_c9	G_{IB}	Ghana	2006	Chronic	Hetero	Plasma	EU885764	
P0402_c2_11	G_{IB}	Portugal	2002	Chronic	Hetero	Plasma	EU885759	
X1193_c1	G_{IB}	Spain	2002	Chronic	IDU	Plasma	EU885761	
X1254_c3	G_{IB}	Spain	2003	Chronic	IDU	Plasma	EU885762	
X1854_c2_10	G_{IB}	Spain	2005	V/VI	IDU	Plasma	EU885763	
X2160_c25	G_{IB}	Spain	2007	Chronic	IDU	Plasma	EU885765	
UNC6316_11	H	USA		Chronic			submitted	

* 6540 and 6545 are a likely epidemiologically linked pair, with very similar virus

TABLE 8

Name	b12neut	b12	Decision Forest	Logistic Regression	173 Y/HS	185 DEn/GST	268 Es/KR	364 PS/AH	369 AP/Ilq	461 EP/	651 N/DIS	susceptit resistan diff			Linear Regression: Predicted neutralization value
398_F1_F6_20	1	0.1	0.75	0.37979461	Y		E	S	L		N	4	1	3	28.17795027
98_v3_c5	1	0.1	0.28125	0.598138266	Y	D		S			N	5	0	5	21.56188791
RHPA4259_7_	1	0.1	0.84375	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
SF162_LS	1	0.1	0.5625	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
3326_v4_c3	1	0.11	0.5625	0.37979461	Y	D	E	A	P			4	1	3	28.17795027
Bal_26	1	0.2	0.84375	0.826209723	Y	D	E	S	P	E	N	7	0	7	12.42629333
Du123_6	1	0.2	0.5	0.201241328	Y		E	H	L		N	3	2	1	34.79401264
Du422_1_	1	0.2	0.65625	0.522540379	Y		E	S	L	E	N	5	1	4	23.61015298
H061_14	1	0.2	0.78125	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
MW965_26	1	0.2	0.84375	0.522540379	Y		E	S	L	E	N	5	1	4	23.61015298
SC422661_8_	1	0.2	0.625	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
QH0692_42_	1	0.3	0.5625	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
1051_12_TD12_3291	1	0.5	0.6875	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
CNE9	1	0.5	0.46875	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
THRO4156_18_	1	0.5	0.375	0.598138266	Y	E		S	P		N	5	0	5	21.56188791
TT31P_2F10_2792	1	0.6	0.65625	0.726782011	Y	D	E	S	P	P		6	0	6	16.99409062
CAP45_2_00_G3_	1	0.7	0.59375	0.37979461	Y		E	P	L		N	4	1	3	28.17795027
REJO4541_67_	1	0.7	0.125	0.454394877	Y			S	P		N	4	0	4	26.1296852
6095_v1_c10	1	0.78	0.5625	0.598138266	Y	D	E	P	P			5	0	5	21.56188791
6644_v2_c33	1	0.8	0.59375	0.073112045		H		S	L			1	2	-1	43.92960722
CNE11	1	0.8	0.71875	0.826209723	Y	D	E	S	P	E	N	7	0	7	12.42629333
Du156_12_	1	0.8	0.875	0.522540379	Y		E	P	L	E	N	5	1	4	23.61015298
H031_7	1	0.9	0.03125	0.37979461		H	D	S	P		N	4	1	3	28.17795027
Du172_17_	1	1	0.9375	0.661696916	Y	D	E	S	L	E	N	6	1	5	19.04235569
MS208_A1	1	1	0.90625	0.661696916	Y	D	E	P	L	P	N	6	1	5	19.04235569
WEAU_d15_410_5017	1	1.1	0.78125	0.826209723	Y	D	E	S	P	E	N	7	0	7	12.42629333
700010040_C9_4520	1	1.2	0.25	0.598138266	Y		E	S	P		N	5	0	5	21.56188791
PRB926_04_A9_4237	1	1.3	0.5	0.598138266	Y	D	E	S	P			5	0	5	21.56188791
Du151_2	1	1.4	0.46875	0.37979461	Y		E	P	L		N	4	1	3	28.17795027
CenvFs2_pt0682_E4uncp	1	1.62	0.375	0.201241328	Y	D	K	S	L			3	2	1	34.79401264
3233_p12	1	1.9	0.75	0.37979461	Y		E	S	L		N	4	1	3	28.17795027
AC10_0_29_	1	1.9	0.59375	0.726782011	Y	E	E	S	P		N	6	0	6	16.99409062
TT29P_3A1_2769	1	1.9	0.625	0.598138266	Y	D	E	S	P			5	0	5	21.56188791
1056_10_TA11_1826	1	2	0.96875	0.826209723	Y	D	E	S	P	E	N	7	0	7	12.42629333
CNE7	1	2.3	0.9375	0.37979461	Y	D		S	L		N	4	1	3	28.17795027
3016_v5_c45	1	2.4	0	0.37979461		D	E	S	A		I	4	1	3	28.17795027
SS1196_1	1	2.4	0.71875	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
BORI_d9_4F8_1413	1	2.5	0.375	0.522540379	Y	G	E	S	P		N	5	1	4	23.61015298
CH181_12	1	2.5	0.96875	0.522540379	Y	D	E	S	L		N	5	1	4	23.61015298
QH0515_1	1	2.6	0.5625	0.598138266	Y	D	E	S			N	5	0	5	21.56188791
CH038_12	1	2.8	1	0.661696916	Y	D	E	S	L	E	N	6	1	5	19.04235569
ZM214M_PL15_	1	3	0.1875	0.093919682		H	K	S	L		N	2	3	-1	41.41007501
SC05_8C11_2344	1	3.1	0.6875	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
WITO4160_33_	1	3.1	0.5625	0.726782011	Y	E	E	S	P		N	6	0	6	16.99409062
ZM249M_PL1_	1	3.2	0.59375	0.093919682	Y		K	S	L		S	2	3	-1	41.41007501
1051_12_C22_3325	1	3.6	0.71875	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062
CNE67	1	3.6	0.75	0.37979461	Y		E	S	L		N	4	1	3	28.17795027

T278_50_	1	3.6	0.875	0.522540379	Y		D		E		P			L		N		5	1	4	23.61015298
1006_11_C3_1601	1	3.9	0.84375	0.661696916	Y		D		E		S		P		E		S	6	1	5	19.04235569
3301_v2_c6	1	3.9	0.6875	0.522540379		S	D		E		S		P			N		5	1	4	23.61015298
6952_v1_c20	1	4.1	0.71875	0.310472494	Y		D		E		S			L			I	4	2	2	30.22621535
Bx08_16	1	4.2	0.46875	0.726782011	Y		D		E		S		P			N		6	0	6	16.99409062
SUMA_d5_8_2	1	4.2	0.65625	0.726782011	Y		D		E		S		P			N		6	0	6	16.99409062
CNE30	1	4.6	0.3125	0.201241328		H	D				S			L		N		3	2	1	34.79401264
CNE16	1	4.8	0.46875	0.522540379	Y				E		S			L	E	N		5	1	4	23.61015298
HIV_26191_2_48_	1	4.9	0.6875	0.522540379	Y		D		E		P			L		N		5	1	4	23.61015298
CNE23	1	5.3	0.625	0.522540379	Y				E		S			L	E	N		5	1	4	23.61015298
BORI_d9_4D7_1410	1	6.2	0.6875	0.826209723	Y		D		E		S		P		E	N		7	0	7	12.42629333
H079_2	1	6.5	0.59375	0.726782011	Y		E		E		S		P			N		6	0	6	16.99409062
CNE20	1	7.4	0.28125	0.15629664	Y			S			S			L	E		I	3	3	0	36.84227772
T266_60	1	7.6	0.09375	0.317871323	Y						S					N		3	0	3	30.69748249
CNE15	1	8.8	0.0625	0.160878565	Y						S			L				2	1	1	37.31354485
703010228_1C4	1	8.9	0.90625	0.522540379	Y		D		E		S			L		N		5	1	4	23.61015298
H022_7	1	9.3	0.625	0.726782011	Y		D		E		P		P			N		6	0	6	16.99409062
CenvFs4_Pt2010_F5	1	10.4	0.65625	0.201241328	Y		D				S			L			I	3	2	1	34.79401264
CH110_2	1	10.8	0.53125	0.522540379	Y		D				S			L	E	N		5	1	4	23.61015298
1054_07_TC4_1499	1	11.1	0.78125	0.726782011	Y		D		E		S		P			N		6	0	6	16.99409062
703010217_B6	1	11.3	0.59375	0.201241328	Y					K	S			L		N		3	2	1	34.79401264
477_F3_13_55	1	11.7	0.71875	0.255201012	Y						S			L		N		3	1	2	32.74574756
270015_J5_1	1	12.9	0.59375	0.522540379	Y		D		E		S			L		N		5	1	4	23.61015298
CenvFs2_Pt1086_B2	1	14.8	0.6875	0.310472494		H			E		S			L	E	N		4	2	2	30.22621535
6535_3_	1	17.5	0.84375	0.826209723	Y		E		E		S		P		E	N		7	0	7	12.42629333
ZM197M_PB7_	1	19.9	0.09375	0.37979461			E		E		S			L		N		4	1	3	28.17795027
CAP210_2_00_E8_	1	20.4	0.15625	0.522540379	Y				E		P			L	E	N		5	1	4	23.61015298
246F_C1G	1	22.7	0.46875	0.093919682		H	D		E			H		L				2	3	-1	41.41007501
CNE18	1	23.2	0.5625	0.37979461	Y				E		S			L		N		4	1	3	28.17795027
249M_B10	1	23.6	0.59375	0.093919682	Y					K	S			L			S	2	3	-1	41.41007501
CH115_12	1	23.8	0.375	0.093919682		H	D				S			L			I	2	3	-1	41.41007501
3415_v1_c1	1	25	0.875	0.522540379	Y				E		P			I	E	N		5	1	4	23.61015298
HIV_25711_2_4_	1	25.9	0.84375	0.522540379	Y		D		E		S			L		N		5	1	4	23.61015298
ZM53M_PB12_	1	25.9	0.90625	0.522540379	Y		D		E		S			L		N		5	1	4	23.61015298
CNE5	1	26.3	0.03125	0.37979461			E		E		P			L		N		4	1	3	28.17795027
401_F1_8_10	1	29.2	0	0.255201012			E				S			L	P			3	1	2	32.74574756
CNE14	1	34.6	0.5625	0.826209723	Y		D		E		S		P		E	N		7	0	7	12.42629333
0815_v3_c3	1	34.8	0.125	0.37979461	Y		D				S			L		N		4	1	3	28.17795027
271_11_	1	38.7	0.59375	0.201241328	Y				E			H		I		N		3	2	1	34.79401264
CNE31	1	41	0.625	0.255201012	Y						S			L		N		3	1	2	32.74574756
6540_v4_c1	1	46.1	0.65625	0.073112045		S					S			I				1	2	-1	43.92960722
1012_11_TC21_3257	0	25	0	0.093919682	Y			T		K	S			L				2	3	-1	41.41007501
1105_p17_1	0	25	0.3125	0.37979461	Y		D		E		S			L				4	1	3	28.17795027
1656_p21	0	25	0	0.37979461	Y				E		S			L		N		4	1	3	28.17795027
191084_B7_19	0	25	0	0.123553999				G			S			L		N		2	2	0	39.36180993
192018_B1_9	0	25	0	0.123553999	Y					K	S			L				2	2	0	39.36180993
2705_p18_1	0	25	0.03125	0.255201012	Y				E			H				N		3	1	2	32.74574756
2843_p5_1	0	25	0.0625	0.454394877	Y				E		S					N		4	0	4	26.1296852

3169_p4	0	25	0	0.123553999		S	E					H				N		2	2	0	39.36180993		
3226_p15	0	25	0.09375	0.255201012			D				S			I		N		3	1	2	32.74574756		
3273_p21_1	0	25	0	0.123553999	Y					R	P			L				2	2	0	39.36180993		
62357_14_D3_4589	0	25	0.53125	0.598138266	Y				E		S		P			N		5	0	5	21.56188791	Transmitted B	USA
6240_08_TA5_4622	0	25	0.03125	0.37979461	Y		D		E		S						I	4	1	3	28.17795027		
6244_13_B5_4576	0	25	0.03125	0.598138266	Y				E		S		P			N		5	0	5	21.56188791	Transmitted B	USA
703010054_2_A2	0	25	0.59375	0.093919682				S	E		S			L			I	2	3	-1	41.41007501	C	Malawi
704010042_2_E5	0	25	0	0.201241328		H			E		S			L		N		3	2	1	34.79401264		
704010083_B8	0	25	0	0.123553999	Y					K	S			L				2	2	0	39.36180993		
704809221_1B3	0	25	0.59375	0.522540379	Y		E		E		P			L		N		5	1	4	23.61015298	C	South Africa
706010018_2E3	0	25	0.03125	0.123553999	Y				E					L			I	2	2	0	39.36180993		
706010164_1A7	0	25	0.59375	0.160878565					E		S			L				2	1	1	37.31354485		
9004SS_A3_4	0	25	0.03125	0.37979461	Y			S			S		P			N		4	1	3	28.17795027		
9014_01_TB1_4769	0	25	0.03125	0.598138266	Y		D				S		P			N		5	0	5	21.56188791	Transmitted B	USA
9020_20_A13_4607	0	25	0.0625	0.598138266	Y		D		E		S		P					5	0	5	21.56188791	Transmitted B	USA
9021_14_B2_4571	0	25	0.53125	0.598138266	Y				E		S		P			N		5	0	5	21.56188791	Transmitted B	USA
92BR025_9	0	25	0	0.093919682		H			E			H		L		N		2	3	-1	41.41007501		
BF1677F2_613a	0	25	0	0.15629664		H	E		E		S			L			I	3	3	0	36.84227772		
BZ167_12	0	25	0.0625	0.726782011	Y		D		E		S		P			N		6	0	6	16.99409062	B	Brazil
C1080rsga_c3	0	25	0	0.37979461	Y		E				P			L		N		4	1	3	28.17795027		
CenvFs2_pt0985_H7uncp	0	25	0	0.15629664	Y		E			K		A		L		N		3	3	0	36.84227772		
CenvFs2_Pt1176_A3	0	25	0	0.093919682	Y					K	S			L			I	2	3	-1	41.41007501		
CenvFs2_pt2060_G9uncp	0	25	0.0625	0.37979461	Y		D				S			L		N		4	1	3	28.17795027		
CenvFs4_Pt0393_C3	0	25	0	0.201241328	Y				E		S			L			I	3	2	1	34.79401264		
CH064_20	0	25	0.0625	0.15629664	Y			T	E		S			L			D	3	3	0	36.84227772		
CH070_1	0	25	0	0.040901701		H			E			H		L			I	1	4	-3	48.02613737		
CH091_9	0	25	0	0.248729004	Y			T		K	S			L	E	N		4	3	1	32.27448043		
CH114_8	0	25	0	0.310472494	Y			T	E		S			L		N		4	2	2	30.22621535		
CH117_4	0	25	0	0.160878565							S			L		N		2	1	1	37.31354485		
CH119_10	0	25	0	0.310472494	Y			T			S			L	E	N		4	2	2	30.22621535		
CH120_6	0	25	0	0.040901701		S		T			S			L			I	1	4	-3	48.02613737		
DJ263_8	0	25	0	0.201241328		S			E		S			I		N		3	2	1	34.79401264		
KSS_0514_13	0	25	0	0.15629664		H		G	E		P			L		N		3	3	0	36.84227772		
NKR_0512_8	0	25	0.03125	0.454394877	Y				E		S		P					4	0	4	26.1296852		
PSR_0508_2	0	25	0	0.093919682	Y					K	S			L			D	2	3	-1	41.41007501		
PWJ_0513_39	0	25	0.03125	0.522540379	Y			G	E		S		P			N		5	1	4	23.61015298	C	Thailand
R1166rsga_c1	0	25	0.09375	0.310472494	Y			G	E		P			L		N		4	2	2	30.22621535		
R2184rsga_c4	0	25	0	0.054819459		H					P			L			D	1	3	-2	45.9778723		
R3265rsga_c6	0	25	0	0.255201012	Y						P			L		N		3	1	2	32.74574756		
RPW_0510_2	0	25	0	0.37979461			D			K	S		P			N		4	1	3	28.17795027		
SPK_0525_13	0	25	0	0.15629664		H		G	E		P			L		N		3	3	0	36.84227772		
TV1_21	0	25	0	0.201241328	Y				E			H		I		N		3	2	1	34.79401264		
0041_v3_c18	0	50	0	0.37979461	Y				E		S			L		N		4	1	3	28.17795027		
0260_v5_c1	0	50	0	0.201241328		H			E		S			L		N		3	2	1	34.79401264		
0330_v4_c3	0	50	0	0.123553999			E				S			L			I	2	2	0	39.36180993		
0907_v4_c12	0	50	0	0.15629664	Y			G		K	S			L		N		3	3	0	36.84227772		
0921_v2_c14	0	50	0	0.054819459		S					P			L			I	1	3	-2	45.9778723		
211_9_	0	50	0	0.201241328	Y			G	E			H				N		3	2	1	34.79401264		

234_F1_16_57	0	50	0	0.310472494		H			E		P			L	E	N		4	2	2	30.22621535			
235_47_	0	50	0	0.255201012	Y						P			L		N		3	1	2	32.74574756			
242_14_	0	50	0	0.093919682		S			E			H		L		N		2	3	-1	41.41007501			
246_F3_C10_2	0	50	0	0.123553999	Y					K	S			L				2	2	0	39.36180993			
252_7	0	50	0	0.201241328	Y		D					H		L		N		3	2	1	34.79401264			
263_8_	0	50	0.03125	0.37979461	Y				E		S			L		N		4	1	3	28.17795027			
269_12	0	50	0	0.123553999	Y							H		L		N		2	2	0	39.36180993			
3103_v3_c10	0	50	0	0.255201012					E		S			L		N		3	1	2	32.74574756			
3168_v4_c10	0	50	0	0.123553999	Y					K	S			L				2	2	0	39.36180993			
3301_v1_c24	0	50	0.625	0.522540379		S	D		E		S		P			N		5	1	4	23.61015298	AC		Tanzania
3337_v2_c6	0	50	0	0.255201012	Y					K	S		P					3	1	2	32.74574756			
3365_v2_c20	0	50	0.0625	0.15629664	Y			S		K	S			L		N		3	3	0	36.84227772			
3468_v1_c12	0	50	0	0.070818806		H	D			K		A		L		N		2	4	-2	43.45834008			
3589_v1_c4	0	50	0	0.123553999	Y							H		L		N		2	2	0	39.36180993			
3637_v5_c3	0	50	0.15625	0.37979461	Y		D				S			L		N		4	1	3	28.17795027			
3718_v3_c11	0	50	0	0.123553999	Y					K	S			L				2	2	0	39.36180993			
3728_v2_c6	0	50	0	0.017242996				G		K		H		L			I	0	5	-5	54.64219974			
3817_v2_c59	0	50	0	0.255201012		H			E		S		P					3	1	2	32.74574756			
3873_v1_c24	0	50	0	0.255201012	Y					K	S		P					3	1	2	32.74574756			
410_F2_1_30	0	50	0.21875	0.445895539	Y		D		E		S			L	E		S	5	2	3	25.65841806			
541_F1_A7_2	0	50	0	0.255201012	Y						S			L		N		3	1	2	32.74574756			
569_F1_37_10	0	50	0	0.054819459			D			R		H		L				1	3	-2	45.9778723			
6022_v7_c24	0	50	0	0.445895539		H	D		E		S			L	E	N		5	2	3	25.65841806			
6040_v4_c15	0	50	0	0.201241328	Y					K	S			L		N		3	2	1	34.79401264			
6041_v3_c23	0	50	0	0.201241328	Y					K	S			L		N		3	2	1	34.79401264			
6101_10	0	50	0	0.093919682	Y			G			S			L			I	2	3	-1	41.41007501			
6322_v4_c1	0	50	0.0625	0.070818806		S	D			K		H		L		N		2	4	-2	43.45834008			
6405_v4_c34	0	50	0	0.070818806		H		G	E			A	P				I	2	4	-2	43.45834008			
6471_v1_c16	0	50	0	0.123553999	Y				E			H		L				2	2	0	39.36180993			
6480_v4_c25	0	50	0	0.123553999		S	D				P			L				2	2	0	39.36180993			
6545_v3_c13	0	50	0.59375	0.073112045		S					S			I				1	2	-1	43.92960722			
6545_v4_c1	0	50	0.5625	0.123553999		S			E		S			I				2	2	0	39.36180993			
6631_v3_c10	0	50	0	0.15629664	Y					K		H		L	P	N		3	3	0	36.84227772			
6650_v1_c8	0	50	0.5	0.310472494	Y		D		E		S			L			I	4	2	2	30.22621535			
6785_v5_c14	0	50	0.09375	0.073112045		H					S			L				1	2	-1	43.92960722			
6811_v7_c18	0	50	0.09375	0.073112045							P			L			S	1	2	-1	43.92960722			
6838_v1_c35	0	50	0	0.054819459		H				K	S			L				1	3	-2	45.9778723			
6980_v1_c17_	0	50	0	0.093919682		H	D				P			L			S	2	3	-1	41.41007501			
783_v0_c51	0	50	0.34375	0.37979461	Y		D				S			L		N		4	1	3	28.17795027			
928_28_	0	50	0	0.123553999	Y							H		L		N		2	2	0	39.36180993			
933_v4_c4	0	50	0	0.093919682	Y				E			H		L			I	2	3	-1	41.41007501			
98_F4_H5_13	0	50	0	0.160878565					E		P						I	2	1	1	37.31354485			
BG1168_1	0	50	0	0.598138266	Y				E		S		P			N		5	0	5	21.56188791		B	USA
CAAN5342_A2_	0	50	0.09375	0.598138266	Y				E		S				E	N		5	0	5	21.56188791		B	USA
CAP244_2_00_D3	0	50	0	0.093919682	Y					K	S			L			I	2	3	-1	41.41007501			
CH111_8	0	50	0.0625	0.37979461	Y				E		S			L		N		4	1	3	28.17795027			
CNE10	0	50	0.09375	0.310472494	Y			G	E		S			L		N		4	2	2	30.22621535			
CNE12	0	50	0.25	0.826209723	Y		D		E		S		P		E	N		7	0	7	12.42629333		B'	China/Jilin

CNE17	0	50	0	0.201241328				S	E		S			L		N		3	2	1	34.79401264		
CNE19	0	50	0.03125	0.522540379	Y				E		S			L	E	N		5	1	4	23.61015298	B'C	China/Xinjiang
CNE21	0	50	0.125	0.37979461	Y						S			L	E	N		4	1	3	28.17795027		
CNE28	0	50	0	0.37979461	Y		E				P			L		N		4	1	3	28.17795027		
CNE3	0	50	0	0.093919682		H			E			H		L		N		2	3	-1	41.41007501		
CNE4	0	50	0.03125	0.522540379	Y			G	E		S		P			N		5	1	4	23.61015298	B'	China/Yunnan
CNE40	0	50	0.03125	0.522540379	Y				E		S			L	E	N		5	1	4	23.61015298	B'C	China/Yunnan
CNE52	0	50	0	0.093919682		H				K	S			L		N		2	3	-1	41.41007501		
CNE53	0	50	0.03125	0.201241328	Y				E		S			L			S	3	2	1	34.79401264		
CNE55	0	50	0	0.15629664		S	D		E		P			L			S	3	3	0	36.84227772		
CNE56	0	50	0.125	0.522540379	Y		E		E		P			L		N		5	1	4	23.61015298	CRF01_AE	China/Yunnan
CNE57	0	50	0.0625	0.522540379	Y		D		E		P			L		N		5	1	4	23.61015298	B'	China/Liaoning
CNE58	0	50	0	0.37979461	Y				E		S			L		N		4	1	3	28.17795027		
CNE59	0	50	0.25	0.37979461	Y				E		P			L		N		4	1	3	28.17795027		
CNE6	0	50	0	0.522540379	Y		D		E			H	P			N		5	1	4	23.61015298	B'	China/Yunnan
CNE8	0	50	0	0.255201012	Y						P			I		N		3	1	2	32.74574756		
H029_12	0	50	0.03125	0.726782011	Y		E		E		S		P			N		6	0	6	16.99409062	B	Peru
H030_7	0	50	0.09375	0.726782011	Y		D		E		S		P			N		6	0	6	16.99409062	B	Peru
H035_18	0	50	0	0.598138266	Y		D				S		P			N		5	0	5	21.56188791	B	Peru
H077_31	0	50	0	0.522540379	Y		E			K	P		P			N		5	1	4	23.61015298		
H078_14	0	50	0	0.201241328		S			E		S			L		N		3	2	1	34.79401264		
H080_23	0	50	0.0625	0.522540379	Y		D		E		S			L		N		5	1	4	23.61015298	B	Peru
H086_8	0	50	0	0.726782011	Y		D		E		S		P			N		6	0	6	16.99409062	B	Peru
HIV_0013095_2_11_	0	50	0.125	0.054819459		H					S			L			I	1	3	-2	45.9778723		
HIV_001428_2_42_	0	50	0	0.522540379	Y		E		E		S			L		N		5	1	4	23.61015298	C	India
HIV_00836_2_5_	0	50	0	0.37979461	Y						S			L	E	N		4	1	3	28.17795027		
HIV_16055_2_3_	0	50	0	0.522540379	Y		E				P			L	E	N		5	1	4	23.61015298	C	India
HIV_16845_2_22_	0	50	0	0.310472494		H	E		E		S			L		N		4	2	2	30.22621535		
HIV_16936_2_21_	0	50	0.03125	0.522540379	Y				E		S			L	P	N		5	1	4	23.61015298	C	India
HIV_25710_2_43_	0	50	0	0.093919682		H					S			L	E		I	2	3	-1	41.41007501		
HIV_25925_2_22_	0	50	0	0.040901701		S				K	S			L			I	1	4	-3	48.02613737		
P0402_c2_11	0	50	0	0.201241328		H			E		S			L		N		3	2	1	34.79401264		
PVO_4_	0	50	0.03125	0.598138266	Y				E		S		P			N		5	0	5	21.56188791		
Q23_17_	0	50	0.09375	0.201241328	Y					K	S			L		N		3	2	1	34.79401264		
Q259_d2_17_	0	50	0	0.093919682		H		G	E		S			L				2	3	-1	41.41007501		
Q461_e2_	0	50	0	0.15629664	Y			S		R	S			I		N		3	3	0	36.84227772		
Q769_d22_	0	50	0	0.255201012	Y		E				S			L				3	1	2	32.74574756		
Q842_d12_	0	50	0	0.201241328	Y					K	S			L		N		3	2	1	34.79401264		
T250_4_	0	50	0.03125	0.201241328	Y				E			H		L		N		3	2	1	34.79401264		
T251_18_	0	50	0.03125	0.255201012	Y			S	E		S							3	1	2	32.74574756		
T253_11	0	50	0	0.040901701				S	E			A		L			I	1	4	-3	48.02613737		
T255_34_	0	50	0	0.093919682		H			E			H		I		N		2	3	-1	41.41007501		
T257_31_	0	50	0	0.093919682	Y			S				A		L		N		2	3	-1	41.41007501		
T280_5	0	50	0	0.37979461	Y				E		S			L		N		4	1	3	28.17795027		
T33_7_	0	50	0.59375	0.201241328	Y				E			H		I		N		3	2	1	34.79401264		
TRJO4551_58_	0	50	0	0.37979461			D		E		P		P				I	4	1	3	28.17795027		
TRO_11_	0	50	0	0.37979461	Y		E		E		S						I	4	1	3	28.17795027		
X1100_c7	0	50	0	0.310472494	Y			T	E		S			L		N		4	2	2	30.22621535		

Name	b12neut	b12	Decision Forest	Logistic Regression	173 Y/HS	185 DEn/GST	268 Es/KR	364 PS/AH	369 AP/Ilq	461 EP/	651 N/DIS	susceptit	resistan	diff	Linear Regression: Predicted neutralization value		
X1193_c1	0	50	0.125	0.37979461	Y			S	L	E	N	4	1	3	28.17795027		
X1254_c3	0	50	0	0.310472494	Y	T	E	P	L		N	4	2	2	30.22621535		
X1854_c2_10	0	50	0	0.160878565	Y			S	L			2	1	1	37.31354485		
X2088_c9	0	50	0	0.093919682	Y	S		P	L		I	2	3	-1	41.41007501		
X2160_c25	0	50	0	0.37979461	Y	D	E	S			I	4	1	3	28.17795027		
X2252_c7	0	50	0.03125	0.522540379	Y	G	E	P	P		N	5	1	4	23.61015298	Clade G -	W. Afr.
ZM106F_PB9	0	50	0	0.040901701	S	T		S	L		S	1	4	-3	48.02613737		
ZM109F_PB4_	0	50	0	0.054819459		S	E	H	L			1	3	-2	45.9778723		
ZM135M_PL10a_	0	50	0.3125	0.201241328		D	E	S	L		S	3	2	1	34.79401264		
ZM215F_PB8	0	50	0	0.123553999			E	S	L		I	2	2	0	39.36180993		
ZM233M_PB6_	0	50	0	0.201241328		T	E	S	L		N	3	2	1	34.79401264		
ZM55F_PB28a	0	50	0	0.255201012			E	S	L		N	3	1	2	32.74574756		
Name	**b12neut**	**b12**	**Decision Forest**	**Logistic Regression**	**173 Y/HS**	**185 DEn/GST**	**268 Es/KR**	**364 PS/AH**	**369 AP/Ilq**	**461 EP/**	**651 N/DIS**	**susceptit**	**resistan**	**diff**	**Linear Regression: Predicted neutralization value**		
B02	1	1.45	0.59375	0.826209723	Y	D	E	S	P	E	N	7	0	7	12.42629333		
SC45_4B5_2631	1	2.79	0.15625	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062		
1059_09_A4_1460	1	4.71	0.21875	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062		
HXB2	1	0.06**	0.28125	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062		
W61D_TCLA_71	1	1.2	0.28125	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062		
PRB958_06_TB1_4305	1	2.31	0.28125	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062		
MN_3	1	0.01*	0.1875	0.598138266	Y	D	E	S	P			5	0	5	21.56188791		
700010058_A4_4375	1	19.62	0	0.522540379	Y	D	K	P	P		N	5	1	4	23.61015298		
BC01	1	7.01	0.71875	0.522540379	Y	D	E	P	L		N	5	1	4	23.61015298		
BC10	1	2.43	0.15625	0.37979461	Y		E	S	L		N	4	1	3	28.17795027		
A07412M1_vrc12	1	9.89	0.21875	0.454394877		E	E	S	P			4	0	4	26.1296852		
BC03	1	20.08	0.5625	0.37979461	Y	D		S	L	E		4	1	3	28.17795027		
1058_11_B11_1550	1	16.15	0.15625	0.37979461	Y		E	S	P		I	4	1	3	28.17795027		
193003_B10	1	2.31	0.25	0.310472494	Y	D	E	S	L		I	4	2	2	30.22621535		
191955_A11	1	22.7	0.3125	0.310472494	Y	S	E	S	L		N	4	2	2	30.22621535		
1394C9G1_Rev_	1	19	0	0.160878565				S	L		N	2	1	1	37.31354485		
BC11	1	1.15	0.03125	0.201241328	Y		R	S	L		N	3	2	1	34.79401264		
703010217B6_Rev_	1	21.2	0.59375	0.201241328	Y		K	S	L		N	3	2	1	34.79401264		
2010F5_Rev_	1	11.53	0.65625	0.201241328	Y	D		S	L		I	3	2	1	34.79401264		
AE02	1	10.12	0.21875	0.093919682	H		E	P	L		D	2	3	-1	41.41007501		
B03	0	25	0.03125	0.726782011	Y	D		P	P	E	N	6	0	6	16.99409062	B	China/Hebei
PRB931_06_TC3_4930	0	25	0.34375	0.726782011	Y	D	E	S	P		N	6	0	6	16.99409062	Transmitted B	USA
B04	0	25	0.25	0.661696916	Y	G	E	P	P	E	N	6	1	5	19.04235569	B	China/Hebei
63358_04_P3_4013	0	25	0.1875	0.598138266	Y		E	S	P		N	5	0	5	21.56188791	Transmitted B	USA
B01	0	25	0.03125	0.598138266	Y		E	S	P		N	5	0	5	21.56188791	B	China/Hebei
BJOX003000_19_1	0	25	0.25	0.598138266	Y	D		P	P		N	5	0	5	21.56188791		
704809221 1B3_Rev_	0	25	0.59375	0.522540379	Y	E	E	P	L		N	5	1	4	23.61015298	C	South Africa
BC05	0	25	0.6875	0.522540379	Y	D	E	S	L		N	5	1	4	23.61015298	CRF07_BC	China/Sichuan
BC07	0	25	0.15625	0.522540379	Y	E	E	S	L		N	5	1	4	23.61015298	CRF07_BC	China/Sichuan
BC04	0	25	0.125	0.37979461	Y		E	S	L		N	4	1	3	28.17795027		
B05	0	25	0	0.310472494	H	D	K	S	P		N	4	2	2	30.22621535		
BC15	0	25	0.3125	0.310472494	Y	T	E	S	L		N	4	2	2	30.22621535		
BJOX002000_03_2	0	25	0.5625	0.310472494	Y		K	S	L	P	N	4	2	2	30.22621535		
BJOX020000_03_2	0	25	0.5625	0.310472494	Y		K	S	L	P	N	4	2	2	30.22621535		

CenvFs2_pt2103_E8uncp	0	25	0.25	0.310472494	Y		E		E		S			L			S	4	2	2	30.22621535		
ZM247v1_Rev_	0	25	0.125	0.310472494	Y					K	S			L	E	N		4	2	2	30.22621535		
BC08	0	25	0	0.255201012					E		S			L		N		3	1	2	32.74574756		
BC18	0	25	0	0.255201012	Y						S			L		N		3	1	2	32.74574756		
96ZM651_02	0	25	0.21875	0.201241328	Y				E		S			L			S	3	2	1	34.79401264		
AE01	0	25	0.40625	0.201241328		H			E		P			L		N		3	2	1	34.79401264		
AE03	0	25	0.46875	0.201241328		H			E		P			L		N		3	2	1	34.79401264		
BC06	0	25	0.3125	0.201241328	Y					K	S			L		N		3	2	1	34.79401264		
BC14	0	25	0.1875	0.201241328	Y				E		S			L			S	3	2	1	34.79401264		
BC17	0	25	0.1875	0.201241328	Y				E		S			L			I	3	2	1	34.79401264		
7030102001E5_Rev_	0	25	0	0.160878565							S			L		N		2	1	1	37.31354485		
7060101641A7_Rev_	0	25	0.59375	0.160878565					E		S			L				2	1	1	37.31354485	C	South Africa
BC09	0	25	0.03125	0.15629664	Y			T		K	S			L		N		3	3	0	36.84227772		
BJOX005000_09_2	0	25	0.125	0.15629664		H	E		E		S			L			S	3	3	0	36.84227772		
BJOX010000_06_2	0	25	0.53125	0.15629664		H		G	E		S			L		N		3	3	0	36.84227772		
UNC6316_11	0	25	0	0.123553999			E					H		L		N		2	2	0	39.36180993		
7030100542A2_Rev_	0	25	0.59375	0.093919682				S	E		S			L			I	2	3	-1	41.41007501	C	Malawi
BJOX025000_01_1	0	25	0.09375	0.093919682		H			E		P			L			D	2	3	-1	41.41007501		
CenvFs2_pt1172_H1uncp	0	25	0	0.093919682					E			H		L	P		I	2	3	-1	41.41007501		
BJOX009000_02_4	0	25	0.21875	0.070818806		H		G	E		P			L			D	2	4	-2	43.45834008		
BF1266_431a	0	25	0	0.040901701				G	E			H		L			S	1	4	-3	48.02613737		
BJOX028000_10_3	0	25	0	0.040901701		H		S				H		L		N		1	4	-3	48.02613737		
57128_vrc15		nt	0.28125	0.522540379	Y		D		E		S		P				I	5	1	4	23.61015298		
704010017_B7		nt	0.3125	0.522540379	Y				E		S			L	E	N		5	1	4	23.61015298		
A08483M1_vrc9		nt	0.09375	0.454394877	Y				E		S		P					4	0	4	26.1296852		
E08364M4_ec3		nt	0.3125	0.37979461		H	D		E		S		P					4	1	3	28.17795027		
BJOX019000_10_2		nt	0.125	0.310472494	Y			S	E		S			L		N		4	2	2	30.22621535		
9009SA_A4_2		nt	0	0.201241328		H	D		E		S	H	P					3	2	1	34.79401264		
BC02		nt	0.375	0.201241328	Y			T		R	S					N		3	2	1	34.79401264		
BC16		nt	0	0.160878565							S			L		N		2	1	1	37.31354485		
CNE62		nt	0.40625	0.15629664		H		G	E		P			L		N		3	3	0	36.84227772		
NKU3006_ec1		nt	0.15625	0.15629664		H	D		E		S			L			I	3	3	0	36.84227772		
CenvFs3_pt703010131_1E2uncja		nt	0.28125	0.073112045								A		L		N		1	2	-1	43.92960722		
93UG065_ec3		nt	0	0.054819459		H		G		K	S							1	3	-2	45.9778723		

* Data from HSC (Hagen von Briesen)

** Data from Binley2004

Yellow highlights indicate env clones with sequencing issues; virus not prepared; assays not performed

Beige highlights indicate viruses which are non-infectious;; virus not prepared; assays not performed

Table 9. Sets of sites used for deeper combinatorial analyses of signatures.

Region[1]	Env glycoprotein	Amino acid positions (HXB2 numbering)	Reference(s)
b12 epitope	gp120	A281, S364, S365, G367, P369, T373, Y384, N386, P417, C418, R419, D474	[35]
CD4bs	gp120	E102, H105, T123, P124, L125, C126, V127, S128, L129, T194, S195, C196, N197, T198, V255, S256, T257, V275, T278, D279, N280, A281, K282, T283, S364, S365, G366, G367, D368, P369, E370, I371, V372, S375, Y384, I424, N425, M426, W427, Q428, K429, V430, G431, K432, L453	[60]
CoRbs	gp120	K117, K121, K207, N377, E381, R419, I420, K421, Q422, P438, R440, G441, R444	[95, 96]
MPER	gp41	Contiguous positions 664-680: D664, K665, W666, A667, S668, L669, W670, N671, W672, F673, D674, I675, T676, N677, W678, L679, W680	[87, 131]

[1]b12 epitope, the region of gp120 that is bound by mAb b12; CD4bs, the CD4 binding site; CoRbs, the CCR5 coreceptor binding site; MPER, membrane proximal external region. The entire protein was scanned for simple signatures, but for more complex signatures (multiple amino acids per position and multiple positions in combination) to make the analyses computationally feasible, only regions of known biological relevance were scanned. An examination was made of the preservation or loss of glycosylation sequons (potential N-linked glycosylation sites PNLGs) in conjunction with neutralization susceptibility, testing for the acquisition or loss of the amino acid pattern Nx[ST], where N is an Asp, x is any amino acid, and [ST] is either a Ser or Thr.

Table 10. Summary of charged residues in the gp120 core structure. Qualitative evaluation of all acidic residues in the recent X-ray structure of b12-bound to the JRFL gp120 [35] that was used in the electrostatic potential calculations.

Residue	Salt bridge partner	Conservation	Correlated Mutations
E83	None	Conserved	-
E91	K284	Conserved	-
D99	R480	-	-
E102	R480	D/E/Q/N	-
E106	-	A/T/Q/E/K	-
D107	-	Conserved	-
D113	-	Conserved	-
E211	-	D/Q/T	-
E267	K231	-	230/231
E268	-	-	-
E269	K348	-	348
D279	K282	D/N/S/K	-
E293	K337	-	337
E351	K348	-	348
D368	-	Conserved	-
E370	-	Conserved	-
E381	-	Conserved	-
D412	R335	-	335
D457	b12 interface	Conserved	-
E464	K357	-	-
E466	R456	-	456
D474	R480	No charge flip	-
D477	R480	Conserved	-
E482	K485	Variable	-
E492	K490	-	K490

Table 11: A list of all sites that co-vary with b12 signature sites. All sites are found to co-vary in a contingency table analysis with a q-value < 0.2. Co-variation sets among signature sites are highlighted in bold or underlined.

b12 signature site[1]	Env glycoprotein	Co-varying site(s) [1]
163	gp120	None
173	gp120	8, 10, 47, 150, 151, 155, 156, 172, 181, 292, 305, 340, 364, 707, 732, 813
182	gp120	None
185	gp120	20, 29, 97, 130, 134, 152, 200, **268,** 271, 275, 300, 304, 311, 320, 340, 397, 429, 476, 490, 496, 588, 595, 607, 633, **655,** 724, 792, 821, 837
268	gp120	4, 5, 108, 121, 149, 159, 169, 170, 178, 183, 184, **185,** 189, 202, 231, 271, 273, 297, 341, 348, 412, 429, 430, 462, 464, 502, 629, 641, 727, 777, 840, 855
364	gp120	173
369	gp120	None
461	gp120	133, 336, 464, 515, 779, 831
651	gp41	80, 84, 169, 429[2], 432, 602, 798, 817, 822
655	gp41	**185**

[1]All site positions are based on HXB2 numbering.

Table 12. HIV-1-positive serum samples used for signature analysis. Single SGA Env clones were sequenced from each sample. All samples were taken during chronic infection, at the same time the sample was tested for cross-reactive neutralizing antibodies. All sequences have been submitted to GenBank (in progress).

Patient code	Subtype	Sample Country	Year	sequence name	GenBank Accession Number
700010025	B	USA	2006	CH010025.w48.p1	submitted
704010028	C	South Africa	2007	CH010028.w24.p1	submitted
700010032	B	USA	2006	CH010032.48.p1	submitted
703010073	C	Malawi	2007	CH010073.w16.p1	submitted
703010085	C	Malawi	2007	CH010085.w4.p1	submitted
706010090	C	South Africa	2007	CH010090.w8.p1	submitted
700010094	B	USA	2006	CH010094.w48.p1	submitted
703010098	C	Malawi	2007	CH010098.w16.p1	submitted
703010102	C	Malawi	2007	CH010102.e.p2	submitted
700010111	B	USA	2006	CH010111.w48.p1	submitted
704010124	C	South Africa	2007	CH010124.w24.p1	submitted
702010141	C	Malawi	2007	CH010141.w12.p1	submitted
703010167	C	Malawi	2007	CH010167.w8.p2	submitted
703010180	C/F	Malawi	2007	CH010180.w12.p1	submitted
704010207	C	South Africa	2007	CH010207.w4.p1	submitted
704010210	C	South Africa	2007	CH010210.w2.p2	submitted
701010211	B	USA	2008	CH010211.w2.p1	submitted
702010259	C	Malawi	2008	CH010259.w16.p1	submitted
704010273	C	South Africa	2007	CH010273.w4.p1	submitted
702010293	C	Malawi	2008	CH010293.w8.p1	submitted
704010298	C	South Africa	2007	CH010298.w12.p1	submitted
704010301	C	South Africa	2007	CH010301.w12.p1	submitted
704010316	C	South Africa	2007	CH010316.w16.p1	submitted
704010327	C/G	South Africa	2007	CH010327.w12.p1	submitted
704010330	C	South Africa	2007	CH010330.w16.p1	submitted
704010343	C	South Africa	2007	CH010343.w12.p1	submitted
704010355	C	South Africa	2007	CH010355.w2.p1	submitted
704010368	C	South Africa	2007	CH010368.w8.p2	submitted
706010383	C	South Africa	2007	CH010383.w12.p1	submitted
704010384	C	South Africa	2007	CH010384.w16.p2	submitted
704010392	C	South Africa	2007	CH010392.w4.p2	submitted
704010408	C	South Africa	2007	CH010408.w12.p1	submitted
704010420	C	South Africa	2007	CH010420.w16.p1	submitted
702010432	C	Malawi	2008	CH010432.w4.p1	submitted
702010440	C	Malawi	2008	CH010440.w4.p1	submitted
704010453	C	South Africa	2007	CH010453.w12.p3	submitted

704010461	C	South Africa	2007	CH010461.w12.p1	submitted
704010540	C	South Africa	2008	CH010540.e.p1	submitted
704010581	C	South Africa	2008	CH010581.e.p1	submitted
704010605	C	South Africa	2008	CH010605.w12.p1	submitted
707010175	A	Tanzania	2008	CH0175.e2	submitted
707010219	A1	Tanzania	2008	CH0219.e4	submitted
703010269	C	Malawi	2007	CH0269.e3	submitted
707010457	C	Tanzania	2008	CH0457.e1	submitted
705010534	C	South Africa	2008	CH0534.e1	submitted
707010536	A1/C	Tanzania	2008	CH0536.e2	submitted
713080024	B	England	2008	CH080024.e.p1	submitted
713080038	B	England	2008	CH080038.e.p2	submitted
713080046	B	England	2008	CH080046.e.p1	submitted
713080052	B	England	2008	CH080052.e.p1	submitted
713080060	B	England	2008	CH080060.e.p1	submitted
713080071	B	England	2008	CH080071.e.p2	submitted
713080087	B	England	2008	CH080087.e.p1	submitted
713080095	B	England	2008	CH080095.e.p1	submitted
713080100	CRF01_AE	England	2008	CH080100.e.p1	submitted
713080117	A1	England	2008	CH080117.e.p1	submitted
713080128	B	England	2008	CH080128.e.p1	submitted
713080134	B	England	2008	CH080134.e.p1	submitted
713080142	B	England	2008	CH080142.e.p1	submitted
713080156	B	England	2008	CH080156.e.p1	submitted
713080169	B	England	2008	CH080169.e.p1	submitted
713080175	B	England	2008	CH080175.e.p2	submitted
713080183	B	England	2008	CH080183.e.p1	submitted
713080191	B	England	2008	CH080191.e.p1	submitted
713080203	B	England	2008	CH080203.e.p1	submitted
713080219	B	England	2008	CH080219.e.p2	submitted
713080225	B	England	2008	CH080225.e.p2	submitted
713080258	B	England	2008	CH080258.e.p2	submitted
713080510	A1	England	2008	CH080510.e.p2	submitted

Table 13. HIV-1 strains used for NAb assays to identify signatures in serum-derived Env sequences.

Virus name	Subtype	Country	Year	Fiebig stage	Mode of transmission	Accession
6535.3	B	USA	1995	V	M-M	AY835438
QH0692.42	B	Trinidad	1994	V	F-M	AY835439
SC422661.8	B	Trinidad	1994	IV	F-M	AY835441
PVO.4	B	Italy	1996	III	M-M	AY835444
AC10.0.29	B	USA	1998	III	M-M	AY835446
RHPA4259.7	B	USA	2000	≤V	M-F	AY835447
BB1006-11.C3.1601	B	USA	1997	III	Sexual	EU289183
BB1054-07.TC4.1499	B	USA	1997	II	Sexual	EU289185
700010040.C9.4520	B	USA	2006	V	MSM	EU289193
WEAU-d15.410.787	B	USA	1990	II	MSM	EU289202
Du156.12	C	S. Africa	1998	≤IV	M-F	DQ411852
Du172.17	C	S. Africa	1998	VI	M-F	DQ411853
Du422.1	C	S. Africa	1998	V	M-F	DQ411854
ZM197M.PB7	C	Zambia	2002	≤VI	F-M	DQ388515
ZM214M.PL15	C	Zambia	2003	≤VI	F-M	DQ388516
Ce1086_B2	C	Malawi	2004	I/II	Sexual	FJ444395
Ce0393_C3	C	Malawi	2003	IV	Sexual	FJ444215
Ce1176_A3	C	Malawi	2004	I/II	Sexual	FJ444437
Ce2010_F5	C	Malawi	2005	IV	Sexual	FJ444561
Q23.17	A	Kenya	1994	VI	M-F	AF004885
Q168.a2	AD	Kenya	1995	Acute/early	M-F	AF407148
Q259.d2.17	A	Kenya	1994	Acute/early	M-F	AF407152
Q461.e2	AD	Kenya	1995	Acute/early	M-F	AF407156
Q769.d22	A	Kenya	1996	Acute/early	M-F	AF407158
Q842.d12	A	Kenya	1994	Acute/early	M-F	AF407160

* * *

All documents and other information sources cited herein are hereby incorporated in their entirety by reference.

WHAT IS CLAIMED IS:

1. An HIV envelope protein comprising the signature regions of CHO219.e4 or CHO80510.ep2.

2. The protein according to claim 1 wherein said protein comprises the signature regions of CHO219.e4.

3. The protein according to claim 1 wherein said protein is a gp160 or gp140 protein.

4. The protein according to claim 1 wherein said protein is a gp120 or gp41 protein.

5. The protein according to claim 1 wherein said protein comprises the amino acid sequence of CHO219.e4 gp160 or CHO219.e4 gp140 shown in Fig. 13.

6. An isolated nucleic acid encoding the protein according to claim 1.

7. The nucleic acid according to claim 6 wherein said nucleic acid is present in a vector.

8. The nucleic acid according to claim 7 wherein said vector is a viral vector.

9. A composition comprising the protein according to claim 1 or the nucleic acid according to claim 6 and a carrier.

10. The composition according to claim 9 wherein said composition further comprises an adjuvant.

11. A method of inducing an immune response in a mammal comprising administering said protein according to claim 1 or said nucleic acid according to claim 6 to said mammal in an amount sufficient to induce said response.

12. The method according to claim 11 wherein said mammal is a human.

13. An isolated antibody specific for said protein according to claim 1, or antigen binding fragment thereof.

14. A method of inhibiting infection of a mammalian cell by HIV-1 comprising contacting said cell with said antibody according to claim 13, or said fragment thereof, under conditions so that said inhibition is effected.

15. The method according to claim 14 wherein said cell is a human cell.

Fig. 1

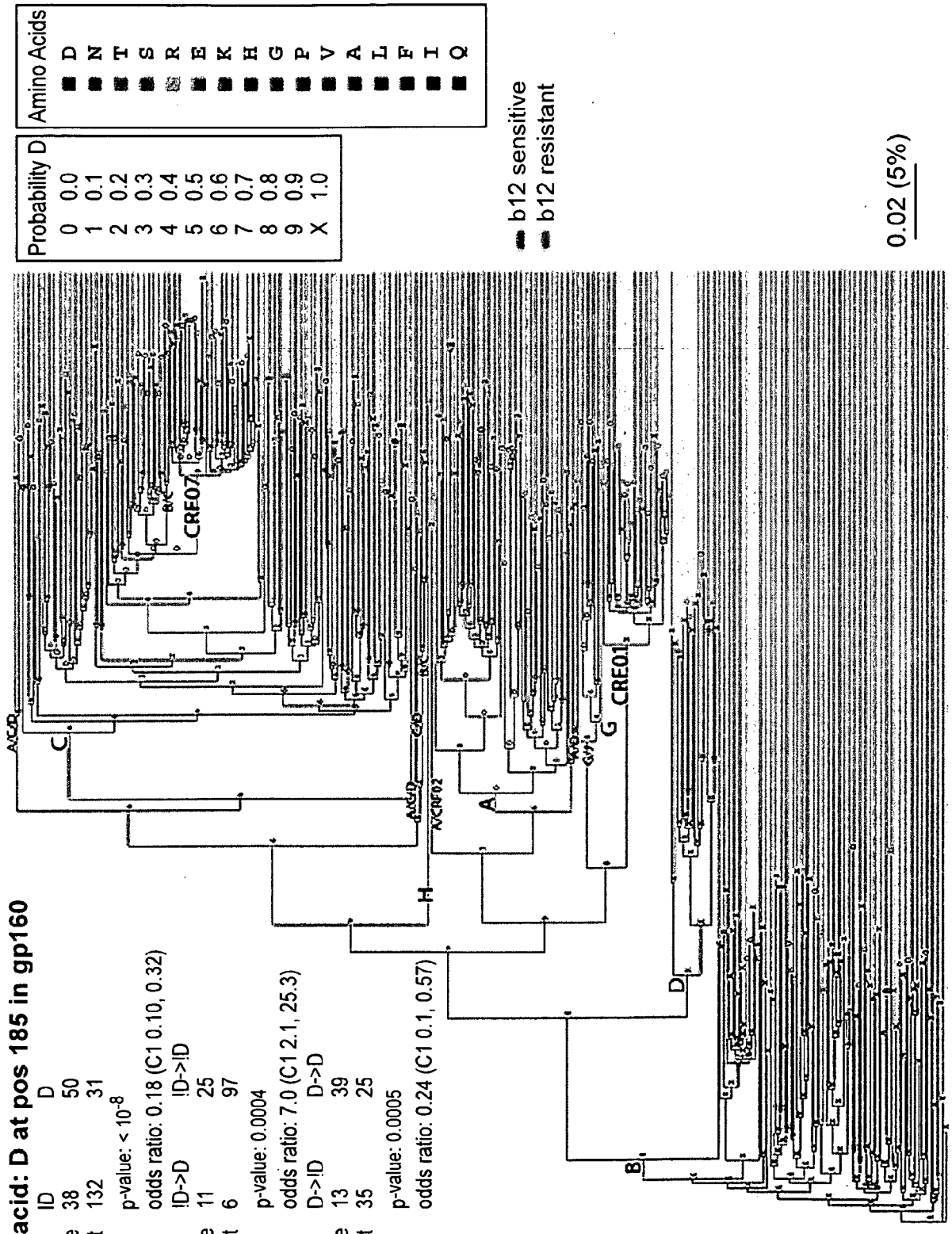
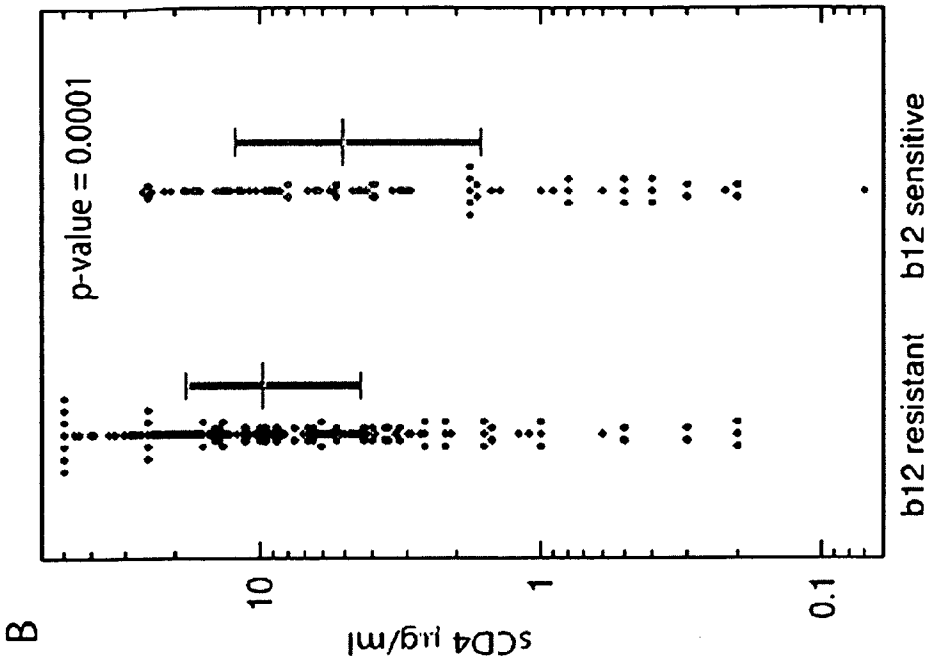

Amino acid: D at pos 185 in gp160
ID D
b12 sensitive 38 50
b12 resistant 132 31
p-value: < 10-8
odds ratio: 0.18 (C1 0.10, 0.32)
!D->D !D->!D
b12 sensitive 11 25
b12 resistant 6 97
p-value: 0.0004
odds ratio: 7.0 (C1 2.1, 25.3)
D->!D D->D
b12 sensitive 13 39
b12 resistant 35 25
p-value: 0.0005
odds ratio: 0.24 (C1 0.1, 0.57)
Probability D
0 0.0
1 0.1
2 0.2
3 0.3
4 0.4
5 0.5
6 0.6
7 0.7
8 0.8
9 0.9
X 1.0
Amino Acids
D
N
T
S
R
E
K
H
G
P
V
A
L
F
I
Q
b12 sensitive
b12 resistant
0.02 (5%)
CRF07
CRF01
AXCRF02
A
B
C
D
G
H

Fig. 2


B
sCD4 µg/ml
10
1
0.1
p-value = 0.0001
b12 resistant
b12 sensitive

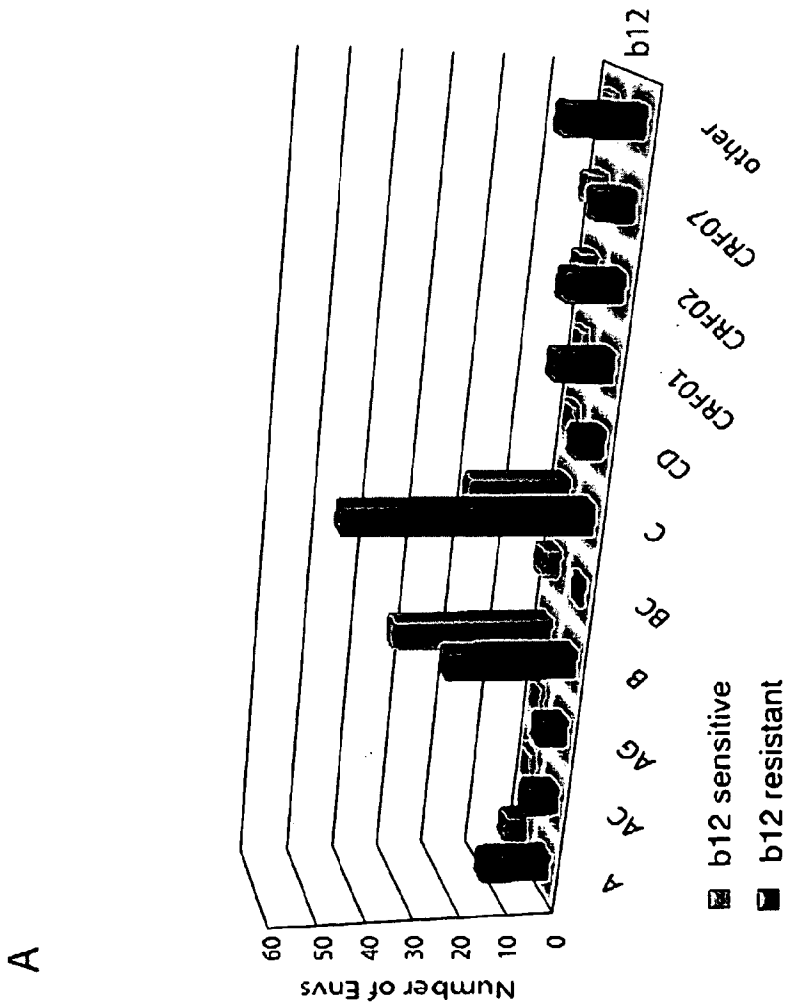

A
Number of Envs
60
50
40
30
20
10
0
A
AC
AG
B
BC
C
CD
CRF01
CRF02
CRF07
other
b12
b12 sensitive
b12 resistant

Fig. 3

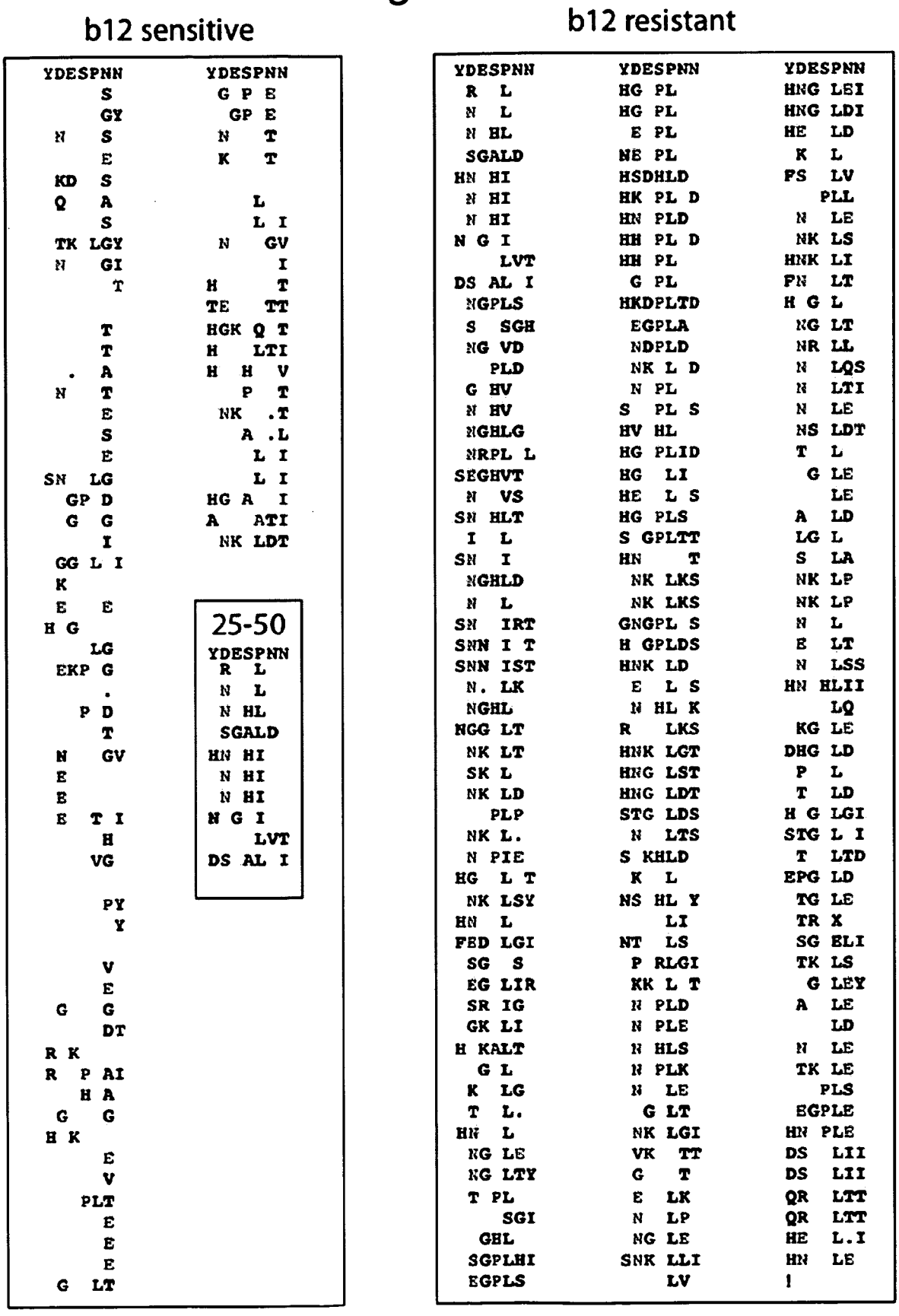
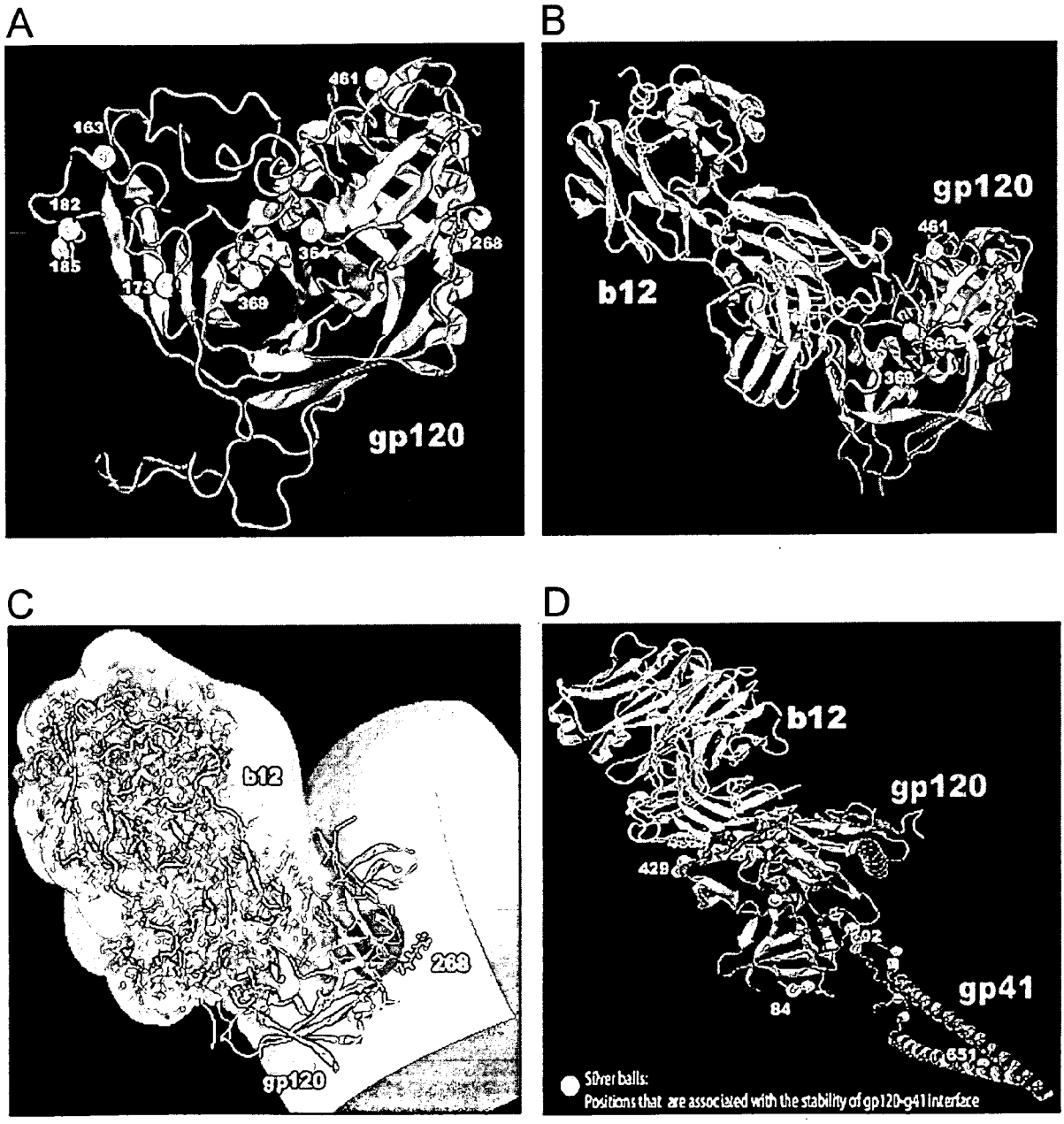

b12 sensitive
YDESPNN
S
GY
N S
E
KD S
Q A
S
TK LGY
N GI
T
T
T
. A
N T
E
S
E
SN LG
GP D
G G
I
GG L I
K
E E
H G
LG
EKP G
.
P D
T
N GV
E
E
E T I
H
VG
PY
Y
V
E
G G
DT
R K
R P AI
H A
G G
H K
E
V
PLT
E
E
E
G LT
YDESPNN
G P E
GP E
N T
K T
L
L I
N GV
I
H T
TE TT
HGK Q T
H LTI
H H V
P T
NK .T
A .L
L I
L I
HG A I
A ATI
NK LDT
25-50
YDESPNN
R L
N L
N HL
SGALD
HN HI
N HI
N HI
N G I
LVT
DS AL I
b12 resistant
YDESPNN
R L
N L
N HL
SGALD
HN HI
N HI
N HI
N G I
LVT
DS AL I
NGPLS
S SGH
NG VD
PLD
G HV
N HV
NGHLG
NRPL L
SEGHVT
N VS
SN HLT
I L
SN I
NGHLD
N L
SN IRT
SNN I T
SNN IST
N. LK
NGHL
NGG LT
NK LT
SK L
NK LD
PLP
NK L.
N PIE
HG L T
NK LSY
HN L
FED LGI
SG S
EG LIR
SR IG
GK LI
H KALT
G L
K LG
T L.
HN L
NG LE
NG LTY
T PL
SGI
GHL
SGPLHI
EGPLS
YDESPNN
HG PL
HG PL
E PL
NE PL
HSDHLD
HK PL D
HN PLD
HH PL D
HH PL
G PL
HKDPLTD
EGPLA
NDPLD
NK L D
N PL
S PL S
HV HL
HG PLID
HG LI
HE L S
HG PLS
S GPLTT
HN T
NK LKS
NK LKS
GNGPL S
H GPLDS
HNK LD
E L S
N HL K
R LKS
HNK LGT
HNG LST
HNG LDT
STG LDS
N LTS
S KHLD
K L
NS HL Y
LI
NT LS
P RLGI
KK L T
N PLD
N PLE
N HLS
N PLK
N LE
G LT
NK LGI
VK TT
G T
E LK
N LP
NG LE
SNK LLI
LV
YDESPNN
HNG LEI
HNG LDI
HE LD
K L
FS LV
PLL
N LE
NK LS
HNK LI
FN LT
H G L
NG LT
NR LL
N LQS
N LTI
N LE
NS LDT
T L
G LE
LE
A LD
LG L
S LA
NK LP
NK LP
N L
E LT
N LSS
HN HLII
LQ
KG LE
DHG LD
P L
T LD
H G LGI
STG L I
T LTD
EPG LD
TG LE
TR X
SG ELI
TK LS
G LEY
A LE
LD
N LE
TK LE
PLS
EGPLE
HN PLE
DS LII
DS LII
QR LTT
QR LTT
HE L.I
HN LE
!

Fig. 4


A
461
163
182
185
173
369
268
gp120
B
gp120
461
b12
364
369
C
b12
268
gp120
D
b12
gp120
429
202
84
gp41
651
Silver balls:
Positions that are associated with the stability of gp120-g41 interface

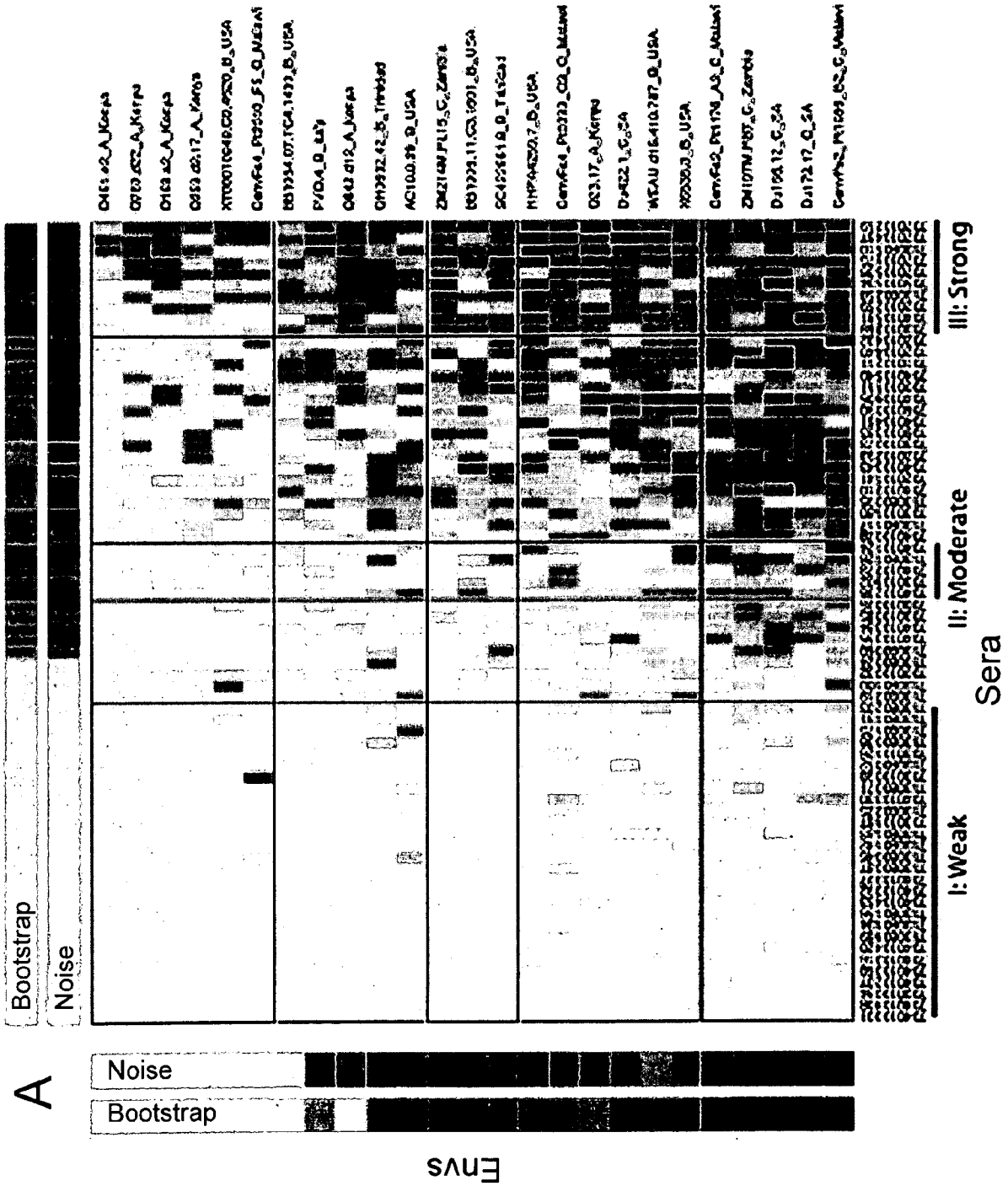

A
Noise
Bootstrap
Envs
Bootstrap
Noise
I: Weak
II: Moderate
III: Strong
Sera

Fig. 5

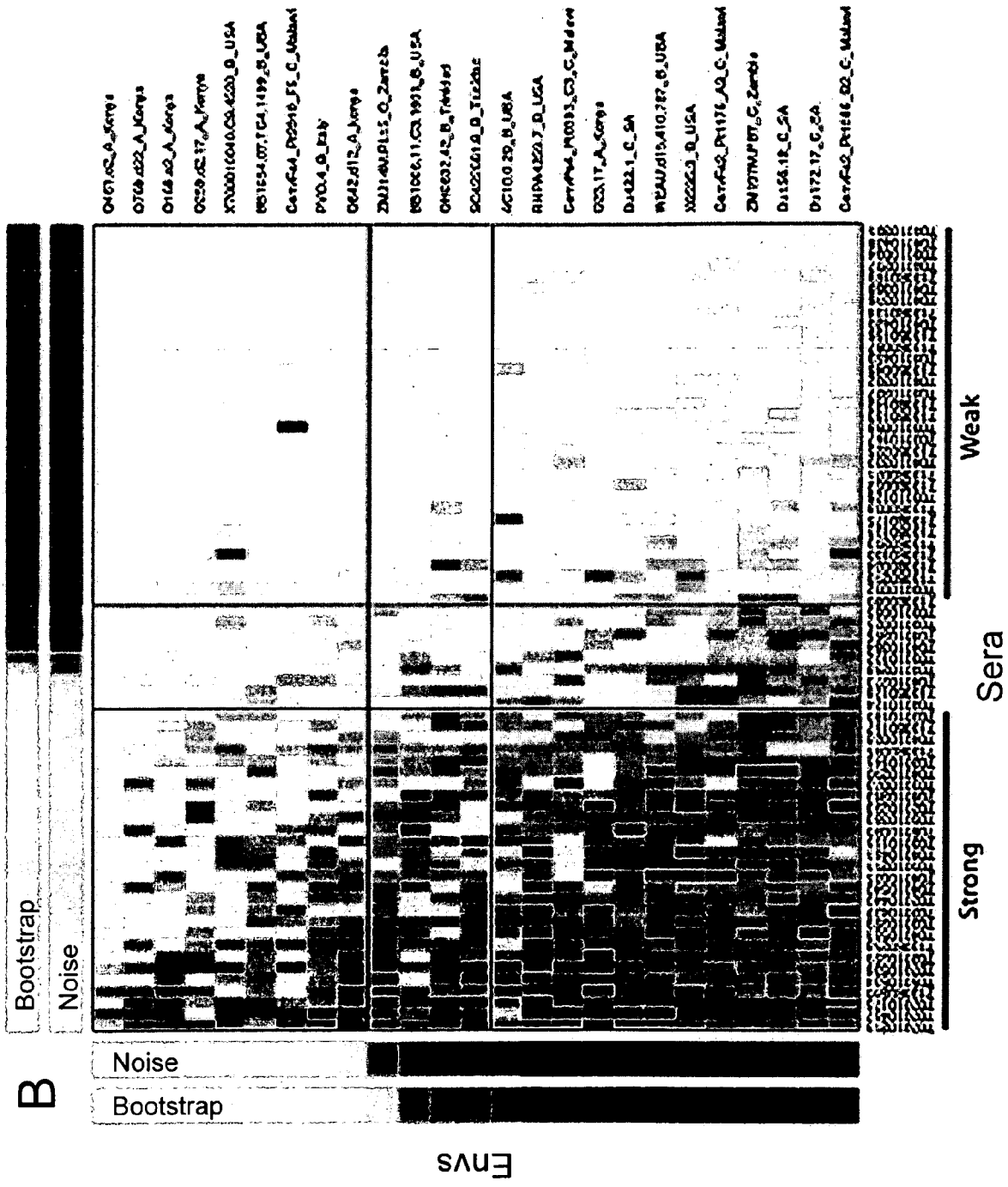


Fig. 5

Fig. 6

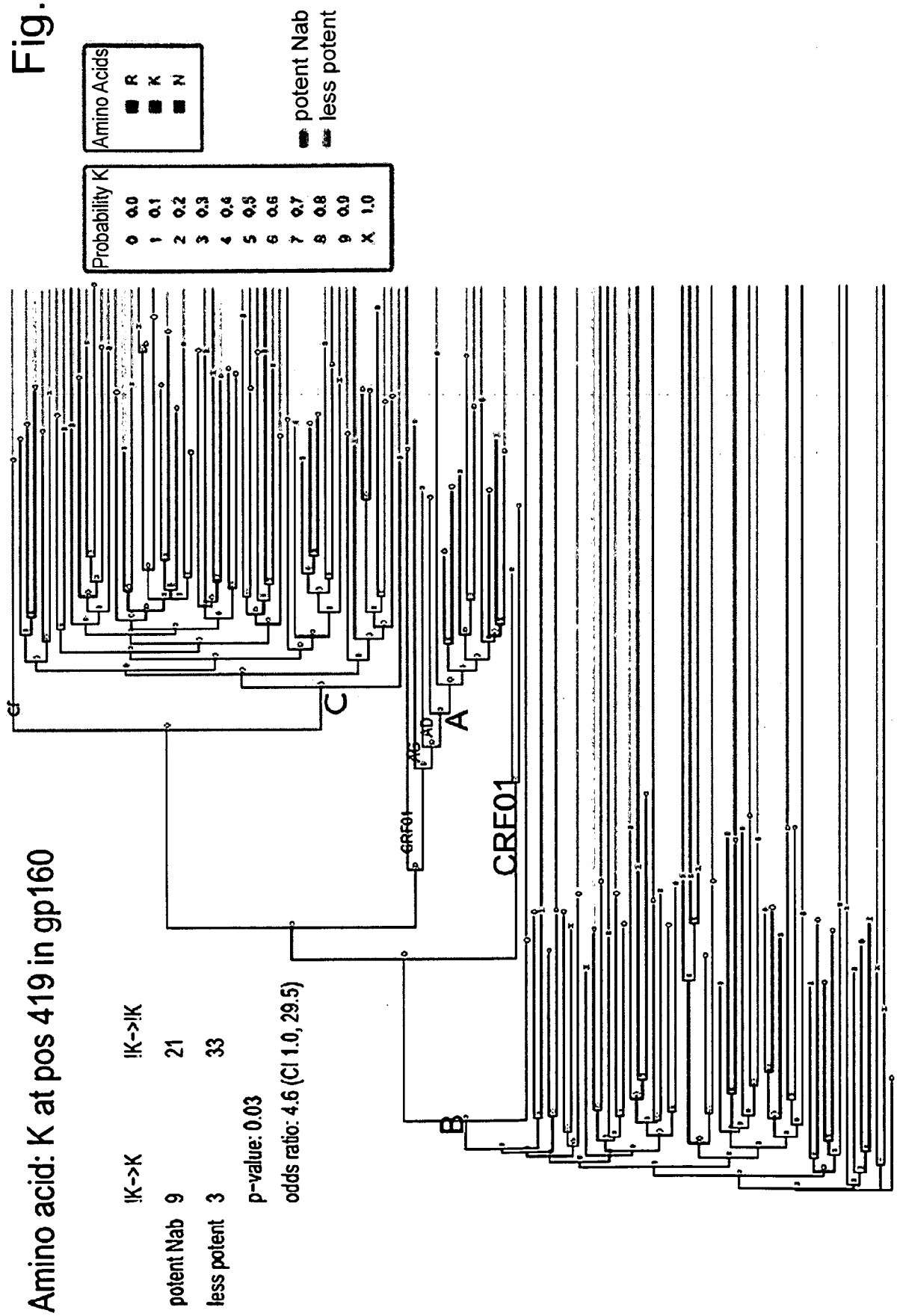

Amino acid: K at pos 419 in gp160
!K->K
!K->!K
potent Nab 9
21
less potent 3
33
p-value: 0.03
odds ratio: 4.6 (CI 1.0, 29.5)
Probability K
0 0.0
1 0.1
2 0.2
3 0.3
4 0.4
5 0.5
6 0.6
7 0.7
8 0.8
9 0.9
X 1.0
Amino Acids
R
K
N
potent Nab
less potent
CRF01
A
B
C
D
AG
GRF01

Fig. 7

Names	Signature Regions				Heatmap Cluster
	412	421	440	447	186
CH010301.w12.p1	NTITLPCRIK		@AMYAPPI		E
CH010408.w12.p1	EVITLPCRIK		RAMYAPPV		E
CH010094.w48.p1	EVITLPCKIK		KAMYAPPI		G
CH010207.w4.p1	TTLTLPCRIK		RAMYAPPI		E
CH010368.w8.p2	NTITIPCKIK		RAMYAPPI		E
CH010316.w16.p1	TNITLSCRIK		RAMYAPPI		E
CH010420.w16.p1	NITTIPCKIK		@AMYAPPI		E
CH080169.e.p1	TILTLPCKIK		KAMYAPPI		N
CH080134.e.p1	DTITLPCRIK		KAMYAPPI		N
CH080156.e.p1	INITLPCRIK		KAMYXPPI		D
CH010453.w12.p3	KTITLPCRIK		RAMYAPPI		N
CH010392.w4.p2	KNITIYCRIK		RAMYAPPI		E
CH010327.w12.p1	STIILPCRIR		RAMYAPPI		N
CH080087.e.p1	TNITLPCRIK		KAMYAPPI		N
CH080046.e.p1	RNITLQCRIR		KAMYAPPI		S
CH080142.e.p1	SNITLPCRIK		KAMYAPPI		N
CH080183.e.p1	PNITLQCKIK		KAMYAPPI		N
CH010167.w8.p2	AIITLPCRIK		RAMYAPPI		E
CH080225.e.p2	DTITLPCKIK		RAMYNPPI		D
CH010330.w16.p1	NSTTIPCKIK		RAMYAPPI		K
CH080071.e.p2	GTITLPCKIK		RAMYAPPI		D
CH010098.w16.p1	DTITLQCKIK		RAMYAPPI		K
CH080052.e.p1	TTISLPCRIK		KAMYAPPI		D
CH080038.e.p2	KTITLPCRIK		KAMYAPPI		N
CH080175.e.p2	DTIILQCRIR		KAMYAPPI		G
CH080191.e.p1	EIVNIPCRIR		KAMYAPPI		N
CH080203.e.p1	DTIILPCKIK		KAMYAPPI		N
CH080219.e.p2	DTITLPCRIK		RAMYAPPI		D
CH080100.e.p1	GTIILPCRIK		@AMYAPPI		E
CH080024.e.p1	DTITLPCRIK		KAMYAPPI		N
CH080060.e.p1	NTITLPCRIK		KAMYAPPI		D
CH010355.w2.p1	TTITLQCRIK		RAMYAPPI		G
CH010384.w16.p2	ANITLQCRIK		@AMYAPPI		N
CH010085.w4.p1	SIITLQCRIK		KAMYAPPI		G
CH010025.w48.p1	TIITLPCRIK		KAMYAPPI		S
CH010298.w12.p1	KNITLQCKIK		RAMYAPPI		N
CH010102.e.p2	ATIYIQCRIK		RAIYAPPI		D
CH010273.w4.p1	DTIILPCRIK		RAMYAPPI		N
CH080128.e.p1	DTITLPCRIK		KAMYAPPI		G
CH010124.w24.p1	ENITLQCRIK		RAIYAPPI		S
CH010028.w24.p1	ATITLQCRIK		RAIYAPPI		N
CH080117.e.p1	DTITLQCRIK		@AMYAPPI		D
CH010180.w12.p1	TNITIPCRIK		@AMYAPPI		S
CH080095.e.p1	TNITLPCRIK		RAMYAPPI		S
CH010210.w2.p2	EDITLPCRIK		KAMYAPPI		E
CH010141.w12.p1	TNITLPCRIK		RAMYAPPI		G
CH010343.w12.p1	ANITLQCRIK		RAIYAAPI		G
CH010293.w8.p1	STITLQCKIK		RAMYASPI		G
CH010073.w16.p1	STIILRCRIK		RAMYAPPI		N
CH010383.w12.p1	EVITLPCRIK		RAMYAPPI		P
CH010461.w12.p1	NTITLPCRIK		RAMYASPI		T
CH010090.w8.p1	SIITLPCRIK		RAMYAPPI		E
CH010605.w12.p1	DTITLPCRIK		RAMYAPPI		G
CH010111.w48.p1	GTITLPCRIK		RAMYAPPI		E
CH010540.e.p1	TEIILQCRLK		KAMYAPPI		G
CH010211.w2.p1	DTIILPCRIK		@AMYAPPI		K
CH010440.w4.p1	STIILPCRIK		KAMYAPPI		N
CH010432.w4.p1	KVISLPCRIK		RAIYAPPI		E
CH0269.e3	TNITLPCRLK		RAMYAPPI		N
CH010581.e.p1	SNITLPCRIK		RAMYAPPI		N
CH010259.w16.p1	SNITLPCRIK		RAMYAPPI		K
CH080258.e.p2	GNITLPCKIR		KAMYAPPI		E
CH0534.e1	ANITLPCRIK		RAMYAPPI		E
CH0536.e2	DTITLQCRIK		@AMYAPPI		N
CH0175.e2	GTITLPCRIK		@AIYAPPI		N
CH080510.e.p2	GNITLQCRIK		@AMYAPPI		N
CH0457.e1	SNITIPCKIK		KAMYAPPI		N
CH0219.e4	SNITIQCRIK		@AMYAPPI		E

Cluster I: low

Cluster I/II indeterminate

Cluster II: moderate

Cluster II/III indeterminate

Cluster III: high

Fig. 8

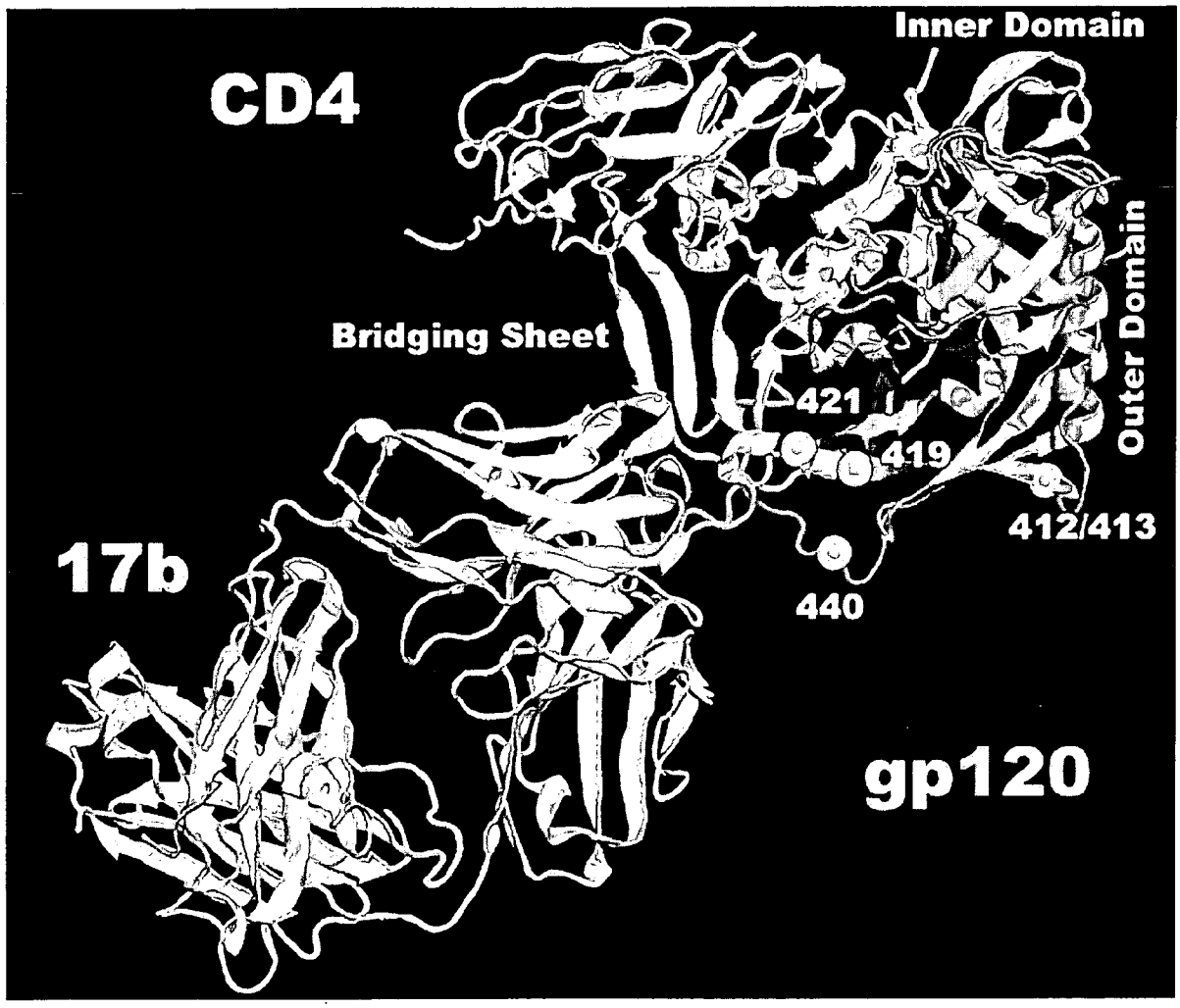

Inner Domain
CD4
Outer Domain
Bridging Sheet
421
419
412/413
17b
440
gp120

Fig. 9

CMI sites: HXB2 positions 163, 182, 665

b12 sensitive: 163 TS, 182 IV, 655 KE?

TVK	TVK
---	---
---	---
S--	--R
---	--T
---	--S
---	--R
---	---
---	K--
S--	---
--R	---
---	--Q
-I-	--R
S--	G--
---	---
---	---
---	S--
--Q	-T-
---	---
---	---
---	---
---	--L
--E	---
---	-L-
---	---

--Q	
-IQ	
---	---
---	---
---	--E
-M-	---
---	---
S--	---
---	---
---	--Q
---	---
---	---

--E	
-I-	
--E	
S--	
S--	

-L-	

---	

b12 resistant: 163 A, 182 AE, 655 QRNI?

TVK	TVK	TVK
---	---	--Q
---	--Q	--R
--Q	-I-	---
---	-IR	--N
---	A--	--I
---	---	--R
--I	---	---
--Q	---	---
---	-I-	-I-
---	---	--R
--R	--M	---
---	-EN	---
---	---	--R
---	--Q	A--
---	---	-A-
---	---	-I-
---	---	A--
---	--R	-E-
---	---	---
---	---	--Q
--Q	---	--N
---	--I	--Q
---	---	--I
--Q	---	--R
---	--Q	---
-EN	-AI	--S
---	---	---
---	---	---
---	-M-	---
---	---	--I
-IQ	-E-	---
-E-	---	---
--R	---	---
---	---	-E-
---	---	---
--R	---	---
---	---	---
---	---	---
--S	--I	---
---	---	--Q
---	--R	-A-
---	--	A--
A-N	-II	---
--R	--R	---
--Q	--N	---
-AR	---	---
--R	---	-E-
--R	---	---
-I-	A--	---
--R	---	---
---	---	--Q
---	---	-E-
---	--R	---
---	--R	---

Fig. 10

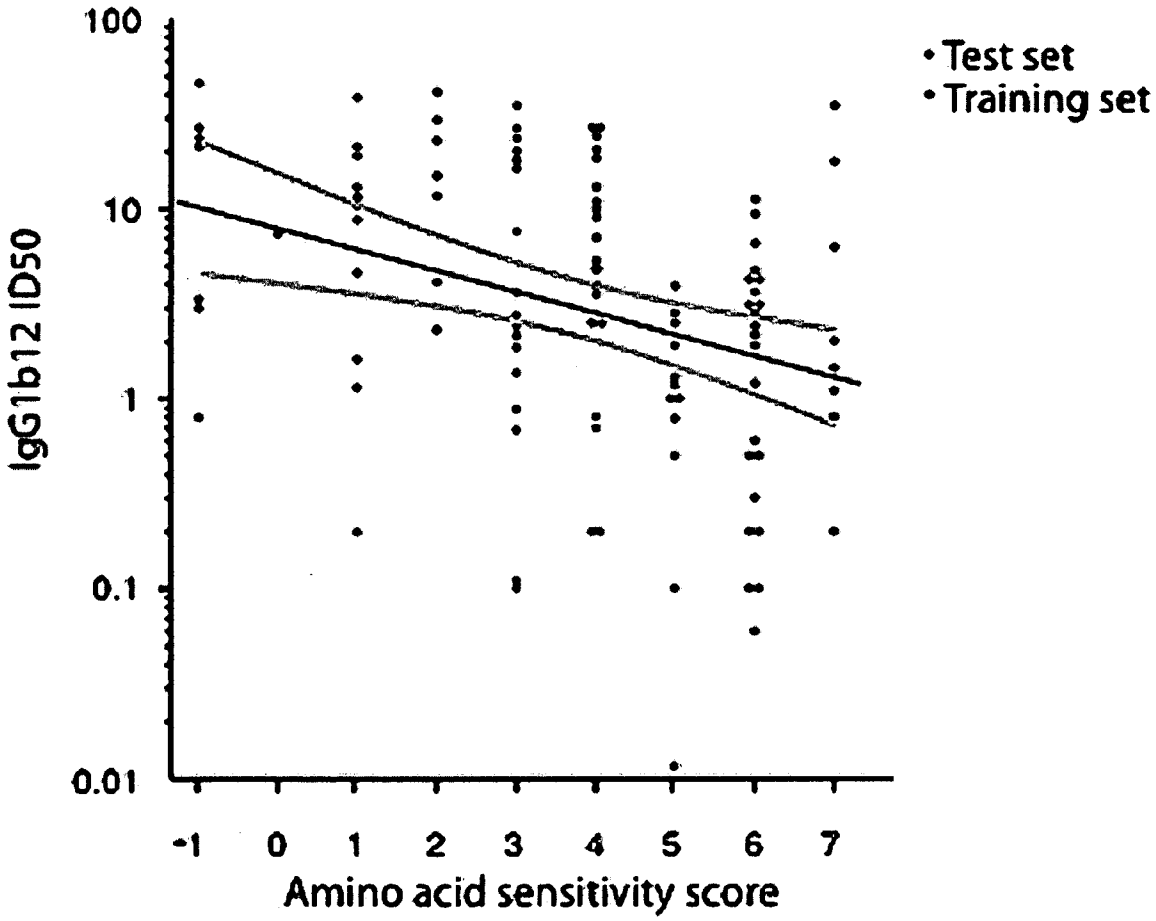

100
10
1
0.1
0.01
IgG1b12 ID50
-1
0
1
2
3
4
5
6
7
Amino acid sensitivity score
• Test set
• Training set

Fig. 11

Genetic Signatures of Broadly NAb Responses

Computational Analysis Based on Env Sequences in Serum Samples

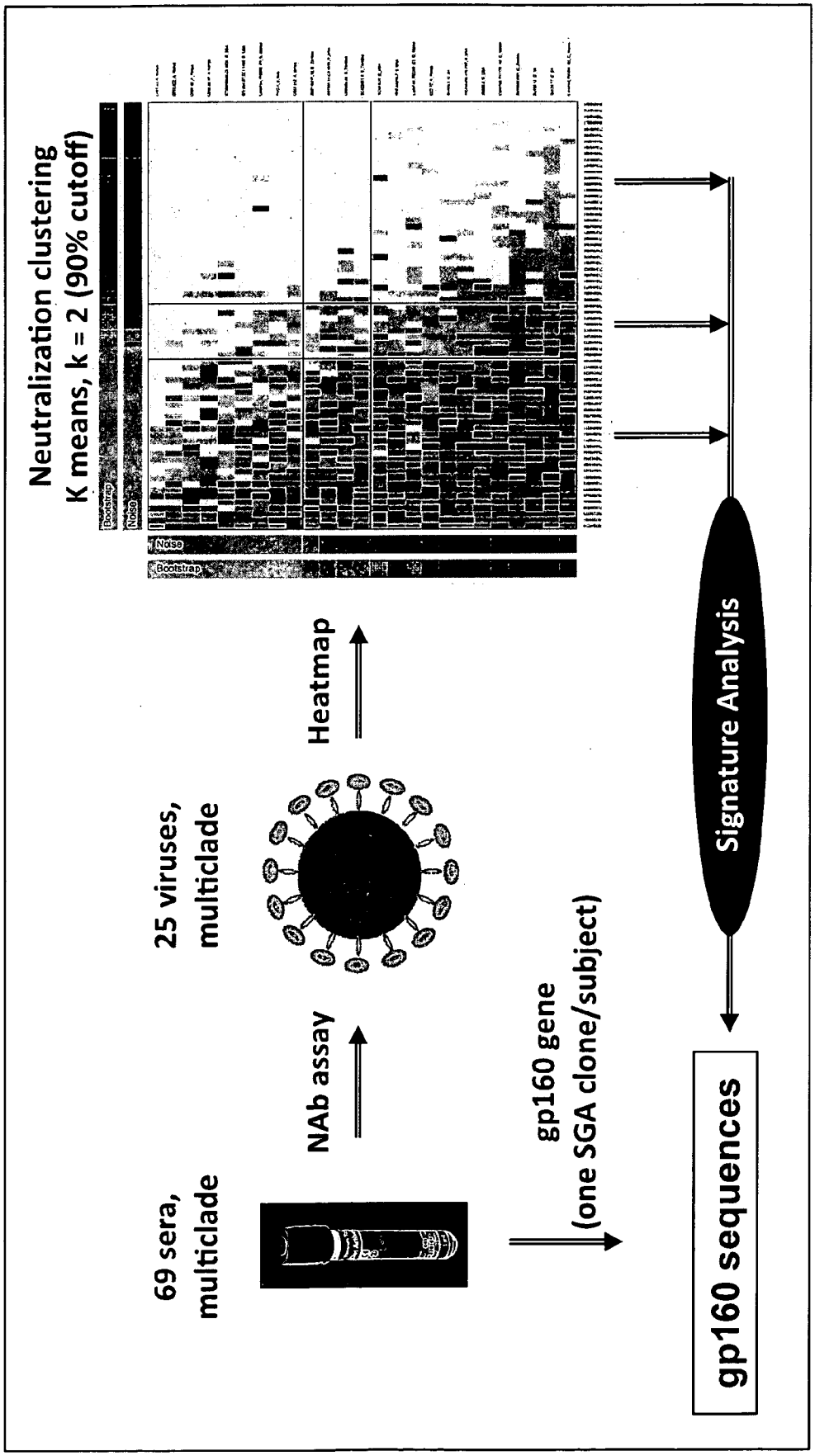
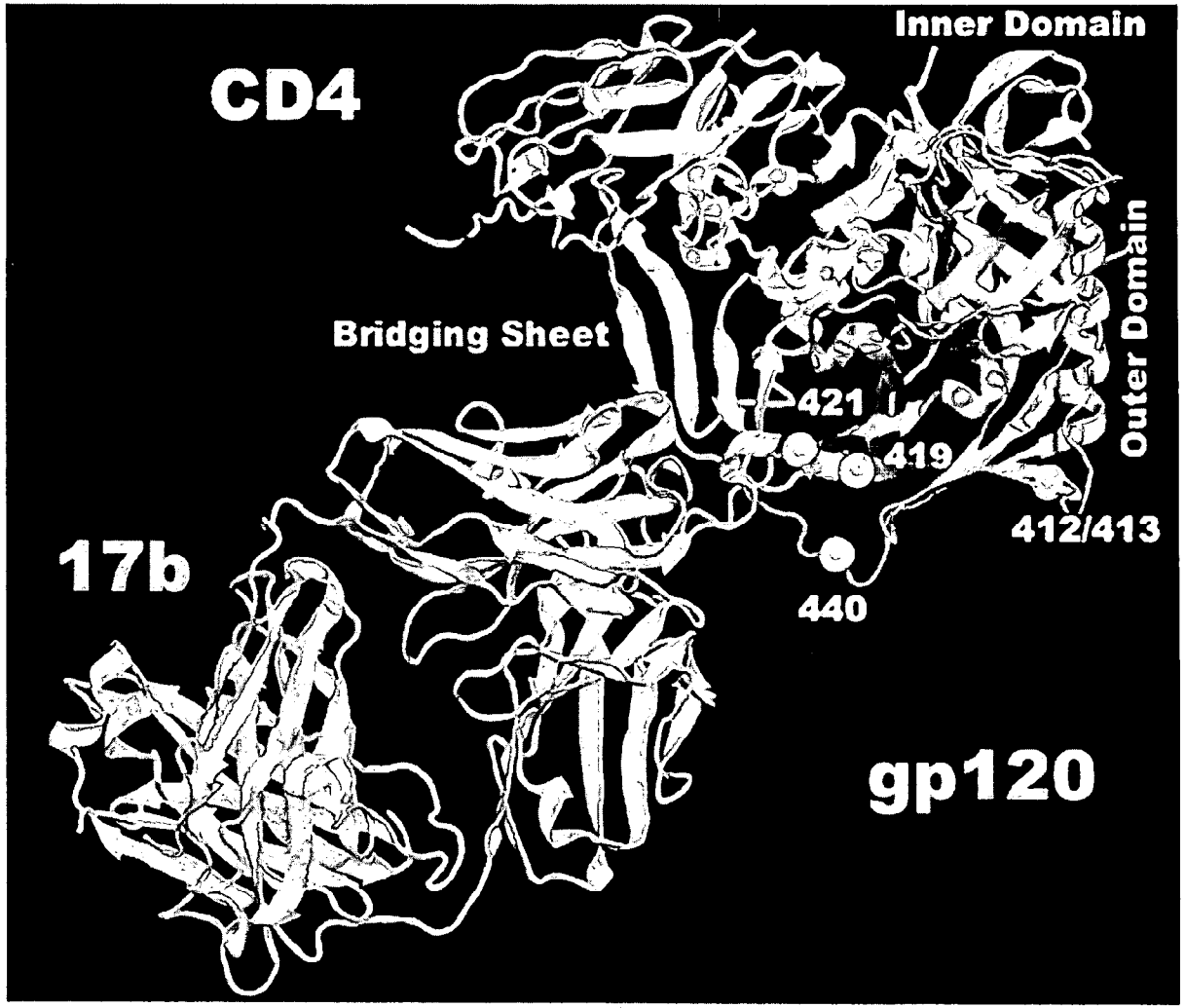


Fig. 12

- Five signatures sites in Env were associated with potent NAbs
- All five sites are in the CD4i region of gp120
- Mutagenesis studies have shown that 419/421 (V4) and 440 (C5) are critical for CCR5 binding
- 413 is often a PNLG in V4



Also, site 186 in V2 was found to be correlated by conditional mutual information

Fig. 13

CH0219_e4 gp160
TAGAACTGCTAGAAGAGCTCAAACATGAAGCTGTTAGACATTTTCCTAGACCATGGCTTCATGGATTAGGACAACATATCTATAACACCTATGGAGAC
ACTTGGGAAGGAGTTGAAGCTATAATCAGAATTCTGCAACACAGCTACTGTTTGTTCATTTCAGAATTGGGTGCCTTCATAGCAGAATAGGCATTAGTCC
AGGGAGAAGAGGCAGGAATGGAACCAGTAGATCCTAACCTAGAAGTCAGCCTACAACTGCTTGTAACAAGTGTTACTGT
AAAAAGTGTTGCTATCATTGTCAAGTTTGCTTTCTGAACAAGGGCTTAGGCATCTCCTATGGCAGGAAGAAGCGGAGACAGCGACGAAGCTCCTCC
AAGCAGTAAGGATCATCAAAATCCTATATATCAAAGCAGTAAGTAGTAATTAGTAAATGCTCTCTTTACAAATCTGTGCAATAGTAGGACTG
ATAGTGGCGATAAATCTTAGCAATAGTTGTGTGGACTATAGTAGGTATAGGAAATTGCTAAGGCAAAAGAAAATAGATAGGCTAATTGAAAG
AATAAGTGAAAGAGCAGAAGATAGTGGCAATGAGAGTGATGGGGACACAGAGGAATTATCCAAACTTGTGTGGAAAGAAGCAGAGACCACTTTATTTTGTGCATCTG
GATAATAATTTGTAGTGCTGCAGAAAACTTGTGGGTTAATGTCTACTATGGGGTACCTGTGTGGAAAGAAGCAGAGACCACTTTATTTTGTGCATCTG
ATGCTAAAGCATATTCTACAGAAGCACATAATGTTTGGGCTACACATGCCTGTGTACCCACAGACCCCAACCCACAAGAAGTATATTTAGAAAATGTG
ACAGAAGAGTTCAACATGTGGAGAAATAAAATGGTAGATCAGATGCAGGAAGATATAGCCAGTCTATGGGACCAATTCCTACAACCATGTGTGAAGTT
AACCCCTCTCTGTGTTACCTTAAATTGCAGTAACCCCAAGAACCCCGATAACAGTACAGATAACAGTACAGGCATAGGCAGAGAAGACATGAAAGACA
TGAAAAACTGCTCTTTCAATATGACCACAGAATTAAGGGATAAACACCAAAAAATGTATTCTCTTTTTATAGACTTGATATAGAAGAACTCAATGAA
AATAGTAATAGCAGTAATAGCAGTAGCAGTGAGTATAGATTAATAAATTGTATAATACCTCAACCATTGCACAAGCTTGTCCAAAGGTATC
CTTTGAGCCAATCCCCATACATTATTGTGCCCCAGCTGGTTTGCGATTCTAAAGTGTAGGGATAAAAAGTTCAATGGAACAGGACCATGCAGGAATG
TCAGCACAGTACAATGCACACATGGAATCAAGCCAGTAGTATCAACTCAACTGCTGTTAAATGGCAGTCTAGCAGAAAAAGGGATAAAGATTAGATCT
AGAATATCTCAGAAAGTGCTAAAACTATAATAGTACAACTTGACCAGCCTGTAGTAATTAATTGTACCAGACCTAACAACAATACAAGAACAAGTAT
ACCTATGGGACCAGGACGAGCATTATATATGCAACAGGGATCCAAGACAAGCACATTGTAATATTAGTAGAGAAAAATGGAATGAAA
CTTTAAGCAAGGTAGCTAAAAAATTAAAAGAATATTTTAATAATACATCAAACTGCCTCAGGGGGGATGTAGAGGTTACAACACAT
AGTTTTAATTGTGGAGGAGATTTTTCTATTGTAATACATCAAACCTGTTTAATAGCACTTGGGGTAATGGCAGTTACAGCACACAGGGGA
CGCAAATTCAAATATAACTATCCAATGCAGAATAAAGCAAATTGTAAGGATGTGGCAGAGGACAGGACAAGCAATGTATGCCCCTCCCATCAAAGGAA
TAATAAGATGTATGTCAAATATTTACAGGCCTACTATTAACAAGAGATGGTGGGATTAATAGGACGAATGAACCTTCAGACCTATAGGAGGAGATATG
ATGGATAATTGGAGAAGTGAGCTATATAAATATAAAGTAGTAAGAATTGAACCAATAGGAGTAGCACCCAACAGGGCAAAGAGAAGTGGTGGAGAG
AGAAAAAAGAGCAGTCTTTGGAATGGGAGCTGTCTTCCTTGGGTTCTTAGGAGGCAGCCGGAAGCACTATGGGCGCAGCGTCGATAACGCTGACGTTAC
AGGCCAGACAATTGTTGTCTGGTATAGTGCAACAGCAAAGCAATTTGCTGAGGGCTATAGAGGCTCAACACAGCATCTCGTTGAAACTCACGGTCTGGGGC
ATTAAACAGCTCCAGGCAAGAGTCCTGGCTGTGGAAAGATACCTACAGGATCAACAGGTCCTAGGACTTTGGGGCTGCTCTGGGAAAATCATCTGCGC
CACTAATGTGCCCTGGAATTCTAGTTGGAGTAATAAATCCTATGGGGAGATATGGGATAACATGACCTGGCTGGAATGGGATAAAGAAGTTAGCAATT
ACACAGACATAATCTATGATCTAATTGCAAAATCGCAAAATCAACAGGAAAAGAATGAACAAGACTTATTGGCATTGGACAAGTTTGTGG
GGTTGGTTTGAAATATCAAGATGGCTGTGGTATATAAAAATATTCATAATGATAGTAGGAGGCTTAATAGGATTAAGAATAGTTTTTGTTATACTGTC
TATAATAAATAGAGTTAGGCAGGGATACTCACCTTTGTCGTTCCAGACCCTTGCCCCGCACCCAAGGGGACTCGACAGGCCCGGAGGAATCGAAGAAG
AAGGTGGAGAGCAAGGCAGAGACAGATCGATTCGATTAGTGAGCGGATTCTTAGCACTTGCCTGGGACGACCTGCGGAGCCTGTGCCTCTTCAGCTAC
CACCGCTTGAGAGACTTCATCTTGATTGTAGCGAGGACTGTGGAACTTCTGGGACACACAACAGTCTCAAGGGGCTGAGACTGGGGTGGAGGGCCTCAA
GTATCTGGGGAACCTCCTGCTATATTGGGGTCAGGAAATAAGAATTAGTGCTTGATACCATAGCAATAGCAGTAGCTGGCTGGACAG
ATAGGGTTATAGAAGTAGGACAAAGAATTGGTAGAGCTATCCTCAACATACCTAGAAGAATCAGACAAGGTTTAGAAAGGCTTCTCTGCTATAACATGG
GTGGCAAATGGTCAAAAAGTAGCATAGTGGGATGGCCTGAGATTAGGGAAAGAATAAGAAGAACTCCAGCAGCGGCAGAAGGAGTAGGAGCAGTATCT
CAAGATTTAGCTAAGCATGGAGCAATCACAAGCAGTAATATAAATCACCCTAGTTGTGCCTGGTTGGAAGAGCAAAAGGAAGAGGAGGGTAGGCTTTCC
AGTCAGGCCACAGGTACCTTTAAGACCAATGACTTACAAGGCAGCTATAGACCTCAGCCATTTTTTAAAAGAAAAGGGGGGACTGGAATAGGTTAATTT
ACTCCAGGAAGAGACAAGAAATCCTTGATCTGTGGGTCTATAACACACAAGGCTTCTTCCCCTGAT

>CH0219_e4 gp160
MRVMGTQRNYPNLWRWGTMLFLGIIICSAAENLWVNVYYGVPVWKEAETTLFCASDAKAYSTEAHNVWATHACVPTDPNPQEVYLENVTEEFNMWRNK
MVDQMQEDIASLWDQFLQPCVKLTPLCVTLNCSNPKNPDNSTDNSTGIGREDMKDMKNCSFNMTTELRDKHQKMYSLFYRLDIEELNENSNSSNSSS

Fig. 13 Cont.-1

NSSEYRLINCNTSTIAQACPKVSFEPIPIHYCAPAGFAILKCRDKKFNGTGPCRNVSTVQCTHGIKPVVSTQLLLNGSLAEKGIKIRSENISESAKTI
IVQLDQPVVINCTRPNNNTRTSIPMGPGRALYATGAITGDPRQAHCNISREKWNETLSKVAKKLKEYFNNRTIIFTNASGGDVEVTTHSFNCGGEFFY
CNTSNLFNSTWGNGSYSTNDTGDANSNITIQCRIKQIVRMWQRTGQAMYAPPIKGIIRCMSNITGLLLTRDGGINRTNETFRPIGGDMMDNWRSELYK
YKVVRIEPIGVAPNRAKRRVVEREKRAVFGMGAVFLGFLGAAGSTMGAASITLTLQARQLLSGIVQQQSNLLRAIEAQQHLLKLTVWGIKQLQARVLA
VERYLQDQQVLGLWGCSGKIICATNVPWNSSWSNKSYGEIWDNMTWLEWDKEVSNYTDIIYDLLAKSQNQQEKNEQDLLALDKWTSLWGWFEISRWLW
YIKIFIMIVGGLIGLRIVFVILSIINRVRQGYSPLSFQTLAPHPRGLDRPGGIEEEGGEQGRDRSIRLVSGFLALAWDDLRSLCLFSYHRLRDFILIV
ARTVELLGHNSLKGLRLGWEGLKYLGNLLLYWGQEIRISAIKLLDTIAIAVAGWTDRVIEVGQRIGRAILNIPRRIRQGLERLLL*

>HV13340 (CH0219_e4 gp160.opt)
SalI--BamHI
GTCGACACCATGCGCGTGATGGGCACCCAGCGCAACTACCCCAACCTGTGGCGCTGGGGCACCATGCTGTTCCTGGGCATCATCATCTGCT
CCGCCGCCGAGAACCTGTGGGTGAACGTGTACTACGGCGTGCCCGTGTGGAAGGAGGCCGAGACCACCCTGTTCTGCGCCTCCGACGCCA
AGGCCTACTCCACCGAGGCCCACACGTGTGGGCCACCCACGCCTGCGTCCCACCGACCCCAACCCCAGGAGGTGTACCTGGAGAACG
TGACCGAGGAGTTCAACATGTGGCGCAACAAGATGGTGGACCAGATGCAGGAGGACATCGCCTCCCTGTGGGACCAGTTCCTGCAGCCCT
GCGTGAAGCTGACCCCCCTGTGCGTGAACTGCTCCAACCCCAAGAACCCCGACAACTCCACCGACGGCATCGGCC
GCGAGGACATGAAGGACATGAAGAACTGCTCCTTCAACATGACCACCGAGCTGCGCGACAAGGCACCAGAAGATGTACTCCCTGTTCTACC
GCCTGGACATCGAGGAGCTGAACGAGAACTCCAACTCCTCCTCCAACTCCTCCGAGTACCGCCTGATCAACTGCAACA
CCTCCACCATCGCCCAGGCCTGCCCCAAGGTGTCCTTCGAGCCCATCCCCATCCACTACTGCGCCCCCGCCGGCTTCGCCATCCTGAAGT
GCCGCGACAAGAAGTTCAACGGCACCGGCCCCTGCCGCAACGTGTCCACCGTGCAGTGCACACCCAGTGGTGTCCACCC
AGCTGCTGCTGAACGGCTCCCTGGCCGAGAAGGGCATCAAGATCCGCTCCGAGTCCGCCAAGACCATCATCGTGCAGC
TGGACCAGCCCGTGGTGATCAACTGCACCCGCCCCAACAACACCCGCACCTCCATCCCCATGGGCCCCGGCCGCGCCCTGTACGCCA
CCGGCGCCATCACCGGCGACCCCCAGGCCCACTGCAACATCTCCCGCGAGAAGTGGAACGAGACCCTGTCCAAGGTGGCCAAGAAGC
TGAAGGAGTACTTCAACGACCCGCCCATCATCTTCACCAACGCCTCCGGCGGCGACGTGGAGGTGACCACCCACTCCTTCAACTGCGGCG
GCGAGTTCTTCTACTGCAACACCTCCAACCTGTTCAACTCCACCTGGGGCAACGGCTCCTACTCCACCAACGACACCGGCGACGCCAACT
CCAACATCACCATCCAGTGCCGCATCAAGCAGATCGTGCGCATGTGGCAGGCCATGTACGCCCCCCCCATCAAGGGCA
TCATCCGCTGCATGTCCAACATCACCGGCCTGCTGACCCGCGACGGCGGCATCAACCGCACCAACGAGACCTTCCGCCCCATCGGCG
GCGACATGATGGACAACTGGCGCTCCGAGCTGTACAAGTACAAGGTGGTGCGCATCGAGCCCATCGGCGTGGCCCCCAACCGCGCCAAGC
GCCGCGTGGTGGAGCGCGAGAAGCGCGCCGTGTTCGGCATGGGCGTTCCTGGGCTTCCTGGGCGCCGCTCCACCATGGGCG
CCGCCTCCATCACCCTGACCCTGCAGGCCCGCCAGCTGCTGTCCGGCATCGTGCAGCAGCAGTCCAACCTGCTGCGCGCCATCGAGGCCC
AGCAGCACCTGCTGAAGCTGACCGTGTGGGGCATCAAGCAGCTGCAGGCCCGCGTGCTGGCGCTACCTGCAGGACCAGCAGG
TGCTGGGCCTGTGGGGCTGCTCCGGCAAGATCATCTGCGCCACCAACGTGCCCTGGAACTCCTCCTGGTCCAACAAGTCCTACGGCGAGA
TCTGGGACAACATGACCTGGCTGGAGTGGGACAAGGAGGTGTCCAACTACACCGACATCATCTACGACCTGATCGCCAAGTCCCAGAACC
AGCAGGAGAAGAACGAGCAGGACCTGCTGGCCCTGGACAAGTGGACCTCCCTGTGGGGCTGGTTCGAGATCTCCCGCTGGCTGTGGTACA
TCAAGATCTTCATCATGATCGTGGGCGGCCTGATCGGCCTGCGCATCGTGTTCGTGATCCTGTCCATCATCAACCGCGTGCGCCAGGGCT
ACTCCCCCCTGTCCTTCCAGACCCTGGCCCCCCCACCCCCGCGGCCTGGACCGCCCCGGCGGCATCGAGGAGGAGGGCGGCGAGCAGGGCC
GCGACCGCTCCATCCGCCTGGTGTCCGGCTTCCTGGCCCTGGGACGACCTGCGCTCCCTGTGCCTGTTCTCCTACCACCGCCTGC
GCGACTTCATCCTGATCGTGGCCCGCACCGTGGAGCTGCTGGGCCACAACTCCCTGAAGGGCCTGCGCCTGGGCTGGGAGGGCCTGAAGT
ACCTGGGCAACCTGCTGCTGTACTGGGGCCAGGAGATCCGCATCTCCGCCATCAAGCTGCTGGACACACCATCGCCGTGGCCGGCT

Fig. 13 Cont.-2

GGACCGACCGCGTGATCGAGGTGGGCCAGCGCATCGGCCATCCTGAACATCCCCCCGCCGCATCCGCCAGGGCCTGGAGGCGCCTGC
TGCTGTAGGAGGATCCA*

>CH0219_e4 gp140C
MRVMGTQRNYPNLWRWGTMLFLGIIICSAAENLWVNVYYGVPVWKEAETTLFCASDAKAYSTEAHNVWATHACVPTDPNPQEVYLENVTEEFNMWRNK
MVDQMQEDIASLWDQFLQPCVKLTPLCVTLNCSNPKNPDNSTDNSTGIGREDMKDMKNCSFNMTTELRDKHQKMYSLFYRLDIEELNENSNSNSSSS
NSSEYRLINCNTSTIAQACPKVSFEPIPIHYCAPAGFAILKCRDKKFNGTGPCRNVSTVQCTHGIKPVVSTQLLLNGSLAEKGIKIRSENISESAKTI
IVQLDQPVVINCTRPNNNTRTSIPMGPGRALYATGAITGDPRQAHCNISREKWNETLSKVAKKLKEYFNNRTIIFTNASGGDVEVTTHSFNCGGEFFY
CNTSNLFNSTWGNGSYSTNDTGDANSNITIQCRIKQIVRMWQRTGQAMYAPPIKGIIRCMSNITGLLLTRDGGINRTNETFRPIGGDMMDNWRSELYK
YKVVRIEPIGVAPNRAKREVVEREKEAVFGMGAVFLGFLGAAGSTMGAASITLTLQARQLLSGIVQQQSNLLRAIEAQQHLLKLTVWGIKQLQARVLA
VERYLQDQQVLGLWGCSGKIICATNVPWNSSWSNKSYGEIWDNMTWLEWDKEVSNYTDIIYDLIAKSQNQQEKNEQDLLALDKWTSLWGWFEISRWLW
YIK_

>HV13341 (CH0219_e4 gp140C.opt)
HindIII-SalI----BamHI-XbaI
AAGCTTGTCGACACCATGCGCGTGATGGGCACCCAGCGCAACTACCCCACCTGTGGCGCTGGGGCACCATGCTGTTCCTGGGCATCATCAT
CTGCTCCGCCGCCGAGAACCTGTGGGTGAACGTGTACTACGGCGTGCCCGTGTGGAAGGAGGCCGAGACCACCCTGTTCTGCGCCTCCGA
CGCCAAGGCCTACTCCACCGAGGCCCACAACGTGTGGGCCACCCACGCCTGCGTGCCCACCGACCCCAACCCCCAGGAGGTGTACCTGGA
GAACGTGACCGAGGAGTTCAACATGTGGCGCAACAAGATGGTGGACCAGATGCACGAGGACATCGCCTCCCTGTGGGACCAGTTCCTGCA
GCCCTGCGTGAAGCTGACCCCCCTGTGCGTGACCCTGAACTGCTCCAACCCCAAGAACCCCGACAACTCCACCGACAACTCCACCGGCAT
CGGCCGCGAGGACATGAAGGACATGAAGAACTGCTCCTTCAACATGACCACCGAGCTGCGCGACAAGCACCAGAAGATGTACTCCCTGTTT
CTACCGCCTGGACATCGAGGAGCTGAACGAGAACTCCAACTCCTCCTCCTCCAACTCCTCCGAGTACCGCCTGATCAACTG
CAACACCTCCACCATCGCCCAGGCCTGCCCCAAGGTGTCCTTCGAGCCCATCCCCATCCACTACTGCGCCCCCGCCGGCTTCGCCATCCT
GAAGTGCCCGCGACAAGAAGTTCAACGGCACCGGCCCCTGCCGCAACGTGTCCACCGTGCAGTGCACCCACGGCATCAAGCCCGTGGTGTC
CACCCAGCTGCTGCTGAACGGCTCCCTGGCCGAGAAGGGCATCAAGATCCGCTCCGAGTCCGCCAAGACCATCATCGT
GCAGCTGGACCAGCCCGTGGTGATCAACTGCACCCGCCCCAACAACACCCGCACCTCCATCCCCATGGGCCCCGCCGCCCTGTA
CGCCACCGGCGCCATCACCGGCGACCCCCGCCAGGCCCACTGCAACATCTCCCGCGAGAAGTGGAACGAGACCCTGTCCAAGGTGGCCAA
GAAGCTGAAGGAGTACTTCAACAACCGCACCATCATCTTCACCAACGCCTCCGGCGCGACGTGGAGGTGACCACCCACTCCTTCAACTG
CGGCGGCGAGTTCTTCTACTGCAACACCTCCAACCTGTTCAACTCCACCTGGGGCAACGGCTCCTACTCCACCAACGACACCGGCGACGC
CAACTCCAACATCACCATCCAGTGCCGCATCAAGCAGATCGTGCGCATGTGGCAGCGCACCGGCCAGGCCATGTACGCCCCCCCCATCAA
GGGCATCATCCGCTGCATGTCCAACATCACCGGCCTGCTGACCCGCGACGGCGGCATCAACCGCACCAACGAGACCTTCCGCCCCCAT
CGGCGGCGACATGAGGACAACTGGCGCTCCGAGCTGTACAAGTACAAGGTGGTGCGCATCGAGCCCATCGGCGTGGCCCCCAACCGCGC
CAAGCGCGAGGTGGTGGAGCGCGAGAAGGAGGCCGTGTGTTCCTGGGCGCCGCCGGCTCCACCAT
GGGCGCCGCCTCCATCACCCTGACCCTGCAGGCCCGCCAGCTGCTGTCCGGCATCGTGCAGCAGCAGTCCAACCTGCTGCGCGCCATCGA
GGCCCAGCAGCACCTGCTGAAGCTGACCGTGTGGGGCATCAAGCAGCTGCAGGCCCGCGTGCTGGCCGTGGAGCGCTACCTGCAGGACCA
GCAGGTGCTGGGCCTGTGGGGCTGCTCCGGCAAGATCATCTGCGCCACCAACGTGCCCTGGAACTCCTCCTGGTCCAACAAGTCCTACGG
CGAGATCTGGGACAACATGACCTGGCTGGAGTGGGACAAGGAGGTGTCCAACTACACCGACATCATCTACGACCTGATCGCCAAGTCCCA
GAACCAGCAGGAGAAGAACGAGCAGGACCTGCTGGCCCTGGACAAGTGGACCTCCCTGTGGGGCTGGTTCGAGATCTCCCGCTGGCTGTG
GTACATCAAGTAGGGATCCTCTAGA

SDS-PAGE and Western Blot of of Purified HV13341 (CH0219_e4Env gp140C)

Fig. 14