Geospatial HIV-1 subtype C gp120 sequence diversity and its predicted impact on broadly neutralizing antibody sensitivity

Jyoti Sutar[1, 2], Suprit Deshpande[1], Ranajoy Mullick[1, 2], Nitin Hingankar[1], Vainav Patel[3], Jayanta Bhattacharya[1, 2]*

1. Translational Health Science & Technology Institute, NCR Biotech Science Cluster, Faridabad, Haryana-121001, India, 2. International AIDS Vaccine Initiative, New Delhi & New York, USA, 3. ICMR-National Institute for Research in Reproductive Health, Mumbai – 400012, India

Key words:
HIV-1, subtype C, envelope, intra-clade diversity, neutralizing antibodies

Running title:
Genetic diversity associated with neutralization of HIV-1 subtype C

***Corresponding author**:

E-mail: JBhattacharya@iavi.org, JBhattacharya@thsti.res.in

Abstract

Evolving diversity in globally circulating HIV-1 subtypes presents formidable challenge in defining and developing neutralizing antibodies for prevention and treatment. HIV-1 subtype C is responsible for majority of global HIV-1 infections. Broadly neutralizing antibodies (bnAbs) capable of neutralizing distinct HIV-1 subtypes by targeting conserved vulnerable epitopes on viral envelope protein (Env) are being considered as promising antiviral agents for prevention and treatment. In the present study, we examined the diversity in genetic signatures and attributes that differentiate region-specific global HIV-1 subtype C *gp120* sequences associated with virus neutralization outcomes to key bnAbs having distinct epitope specificities. A total of 1814 full length HIV-1 subtype C gp120 sequence from 37 countries were retrieved from Los Alamos National Laboratory HIV database (www.hiv.lanl.gov). The amino acid sequences were assessed for their phylogenetic association, variable loop lengths and prevalence of potential N-linked glycosylation sites (pNLGS). Responses of these sequences to bnAbs were predicted with a machine learning algorithm 'bNAb-ReP' and compared with those reported in the CATNAP database. Phylogenetically, sequences from Asian countries including India clustered together however differed significantly when compared with pan African subtype C sequences. Variable loop lengths within Indian and African clusters were distinct from each other, specifically V1, V2 and V4 loops. Furthermore, V1V2 and V2 alone sequences were also found to vary significantly in their charges. Pairwise analyses at each of the 25 pNLG sites indicated distinct country specific profiles. Highly significant differences ($p<0.001$***) were observed in prevalence of four pNLGS (N130, N295, N392 and N448) between South Africa and India, having most disease burden associated with subtype C. Our findings highlight that the distinctly evolving clusters within global intra-subtype C *gp120* sequences are likely to influence the disparate region-specific sensitivity of circulating HIV-1 subtype C to bnAbs.

Introduction

The extraordinary diversity of *env* targeting neutralizing antibodies is a barrier to achieving the desired vaccine-induced and antibody-mediated protection. The evolving antigenic diversity in global and region-specific circulating HIV-1 subtypes is complex which not only poses significant roadblock to developing preventive vaccine but also poses challenge in neutralizing antibody mediated prophylaxis and treatment. Broadly neutralizing antibodies (bnAbs) act solely on the HIV-1 envelope glycoprotein (Env) in neutralizing genetically distinct HIV-1 subtypes. Till date a number of bnAbs have been discovered from elite neutralizers and few of them have been found to be prevent acquisition as well as significantly reducing the plasma viral loads, when tested both in animal models and humans (Escolano et al. 2017; R. Kumar et al. 2018a; Mendoza et al. 2018), thus justifying their importance as products for prevention and treatment. Some of the bnAbs that are currently being evaluated through human clinical trials should provide additional possibilities for prevention (Sok and Burton 2016, 2018), such as their extent, in addition to virus neutralization, in persistent viral clearance (Caskey et al. 2019; Nishimura and Martin 2017) and eliminating HIV-1 infected cells (Caskey et al. 2015; Caskey et al. 2017; R. Kumar et al. 2018a). While some single bnAbs have been found to show significant breadth across subtypes, combination of bnAbs with distinct epitope-specificity is believed to provide the most effective response against globally diverse HIV-1 subtypes and also would likely prevent the development of antibody-escape variants (R. Kumar et al. 2018a). The trimeric Env glycoproteins on the virus surface are the most diverse of all proteins encoded by HIV-1; which differs by greater than 20% of amino acids between matched subtype (Gnanakaran et al. 2007; Han et al. 2020; Hraber et al. 2014; Korber et al. 2001; Lynch et al. 2009) and which continues to diversify at a population level (Bouvin-Pley et al. 2013; Bunnik et al. 2010; DeLeon et al. 2017; Hraber et al. 2014). This has been substantiated by observation that several epitopes that are targeted by different bnAbs those have been found to vary over time (DeLeon et al. 2017). While bnAbs target both surface gp120 and membrane proximal external region (MPER) of gp41, gp120 exhibits extraordinary sequence divergence compared to that of gp41 MPER and variation in this region is believed to represent distinct genetic subtypes, or clades, which are prevalent in distinct geographic regions (Buonaguro et al. 2007; Taylor et al. 2008). Although bnAbs isolated from individuals infected with one particular subtype are generally effective at neutralizing

viruses belonging to other subtypes, antibody potency is often found to be correlated with matched subtypes as described elsewhere (Binley and Burton. 2004; Bures et al. 2002; Hraber et al. 2014; Kulkarni et al. 2009; Li et al. 2006; Seaman et al. 2010). Moreover, diversity has been found to have an impact even within matched subtype, as demonstrated by the fact that the subtype-matched neutralization advantage was more apparent in regions with distinct viral diversities (Hraber et al. 2014).

HIV-1 subtype C accounts for approximately half of the global infections (Novitsky et al. 2002), which predominates in India and South Africa. Yet, robust, comparative *env* sequence diversity and evolution analysis, critical for bnAb based intervention strategies in these regions is severely lacking (Sutar et al. 2019). While recent development of bnAbs has considerably improved our knowledge on conserved epitopes that they target, and Env structure associated with broad and potent virus neutralizing antibodies, a greater understanding of the antigenic diversity of global HIV-1 subtype C Env would facilitate understanding potential bnAb combination that could potentially overcome the intra-clade C diversity. In the present study, we examined the variation in signature sequences, loop length, N-linked glycosylation and key epitopes within the existing globally circulating HIV-1 subtype C *gp120* targeted by potent bnAbs and predicted their potential impact on virus neutralization.

Results

Evidence of phylogenetic divergences of globally circulating HIV-1 subtype C *gp120* sequences.

Previous studies have demonstrated that HIV-1 subtype C is the most abundant globally circulating subtype responsible for approximately 46.6% of the global HIV infections (Hemelaar et al. 2019), In the present study, we retrieved from HIV database (www.hiv.lanl.gov) a total of 23750 sequences covering the complete *gp120* gene (HXB2 coordinates: 6225-7758). To mitigate the intra-individual quasispecies bias, we applied 'one sequence/individual filter' retaining 1927 full length sequences. As shown in Table 1, there is uneven distribution of *gp120* sequences from different countries. Of the 37 countries analyzed, South Africa (ZA) was found to contribute in the database (www.hiv.lanl.gov) more than 1000 sequences, while Malawi (MW), Botswana (BW) and Zambia (ZM) contributed between 100 to 1000 sequences. Interestingly, complete *gp120* sequences were found from only 84 unique individuals from India (IN), while all other countries contributed approximately 50 or less sequences. Upon removal of identical sequences as well as those harbouring internal stop codons, a total of 1814 full length HIV-1C gp120 protein sequences were retained for further analysis. We next analyzed the phylogenetic properties of the subtype C *gp120* sequences that uniquely represent globally circulating subtype C across different geography. Average number of amino acid differences per site between sequences grouped by source country were estimated by MEGA (S. Kumar et al. 2018b). Sequences from Asian Countries (IN, China (CN) and Nepal (NP)) were observed to be closer to each other (0.183-0.208) compared to African countries (ZA and MW: 0.207-0.234). This trend continued in the maximum likelihood tree constructed as indicated in Figure 1A, wherein sequences from Asian countries (IN, CN and NP) were observed to form a unique sub-cluster. To further validate these observations statistically, a subset of sequences (N=251) forming the aforementioned diverging node were reanalysed with 1000 ultrafast bootstrap replicates and SH-aLRT. As indicated in figure 1B, sequences from Asian countries clustered together with 97% bootstrap and 92.1% SH-aLRT support. They also clustered distinctly from African sequences supported by 84.5% SH-aLRT and 79% ultrafast bootstrap replicates.

Geography based subtype C *gp120* loop length and charge variation.

Changes in the lengths of variable loops with gp120 has been previously documented to be an evasion mechanism of HIV-1 to escape neutralizing antibody driven humoral responses (Deshpande et al. 2016; Ringe et al. 2012; van Gils et al. 2011). Therefore, we next assessed lengths of variable regions between sequences from different countries. We included sequences from six countries in this analysis that could contribute more than 50 sequences each (Figure 2A). Comparison of variable region loop (V1-V5) lengths with Kruskal Wallis non-parametric test indicated several statistically significant differences between sequences from diverse countries. For the V1 loop, sequences from UK (GB) had significantly lower length (median:16) compared to all other countries (median range: 23-27). This trend continued in V2 loop with lower length in UK sequences (median: 34) compared to rest of the countries (median range: 42-48). Additionally, sequences from BW and NP were observed to be significantly distinct. However, this was an effect of outlier values and minimal standard deviation in the loop lengths respectively. These differences remained consistent when V1V2 loop were taken together indicating no loop length compensation. The V3 loop length was observed to be highly conserved across all the countries with some outlier values in South Africa. Interestingly, V4 loop length was observed to be most variable among the variable domains. Sequences from UK were observed to have significantly shorter V4 loop (Median: 13) compared to rest of the countries (Median range: 27-34). Of note, sequences from India (IN) had longer V4 loop (Median: 31) than those from South Africa (Median: 27). Sequences from African countries (BW, ET, KE, MW, TZ, ZA and ZM) had shorter V4 loop length (Median range: 27-30) compared to the rest of the countries (Median range: 29-34) except UK. Similar to the V3 loop, V5 loop length was conserved across all the countries under consideration (Median: 11) except UK (median: 6). Overall, length of the gp120 protein was observed to be shortest in sequences from GB (Median: 462) and longest in those reported from NP (Median: 518). Consistent with the V4 loop observations, gp120 length was observed to be lower (Median range: 503-511) in countries from Africa (BW, ET, KE, MW, TZ, ZA and ZM) compared to rest of the countries (Median range: 507-518) except UK. A subset of six countries with more than 50 sequences (IN, ZA, MW, TZ and BW) was further selected for assessment of sequence charge distribution (Figure 2B). V1, V1-hypervariable region, V2, V2 hypervariable region as well as V1V2 cumulatively were found to have significantly different distributions. Of note, V1V2 hypervariable region charge was distinct in India as compared to South Africa.

Comparison of abundance of potential N linked glycosylation sites (pNLGs).

HIV-1 gp120 is a heavily glycosylated protein with host derived N-linked glycans making up ~50% of its total mass. These glycans play an important role in ensuring viral infectivity as well as evading neutralizing antibodies (Doores et al. 2015) emphasizing the importance of their assessment. In the present study, proportion of pNLGs were compared between sequences from the 14 aforementioned countries. pNLGs ranging from 11 to 33 were observed in the sequences. There were no overall significant differences between the number of average pNLGs between countries (Kruskal Wallis test, $p>0.05$). Out of a total of 3003 pairwise analyses performed with Fisher's exact test at each NLG position between each of the countries, 342 combinations were found to be statistically significant (Fisher's exact test, $p<0.01$*) involving 31 of the 33 pNLG sites. The two invariable sites were N140 (abundance: 17-33%) and N301 (abundance: 89-100%). N140, a position in the V1 hypervariable region is perhaps a displaced glycosylation position from N139/141 due to insertion/deletion, while N301 is an important glycosylation site that plays critical role in various envelope functions including membrane fusion and thus is highly conserved (Ogert et al. 2001). Figure 3 represents abundance at of the 25 well characterized pNLG sites denoted along with their functional domains. As apparent in figure 4 as well as observed in statistical comparisons, there was no significant difference between glycosylation profiles at any of the sites between India, China and Nepal. Similar observations were also made for a cluster of African countries South Africa, Tanzania, Kenya and Malawi. Furthermore, sequences from Ethiopia and Sweden were also observed to have similar NLG profiles. Upon comparison of sequences from India and south Africa, highly significant differences (Fisher's exact test, $p <0.001$***) were observed in 4 pNLG sites (N130, N295, N392 and N448), which are present in C1, C2, V4 and C4 domains respectively. These sites are important for integrity of the 'mannose patch' and interaction with several bnAbs such as 2G12, VRC-PG05, PGT135 and PGT151 (Doores et al. 2015).

Variation in Shannon entropy indicate significant intra-subtype C diversity.

Shannon entropy is a measure of variability wherein higher entropy indicates higher variability. Indian sequence 'query' data set (N=83) was compared against South African sequence 'background' data set (N=910) through Entropy-TWO tool on LANL HIV database which generated Shannon entropy values with

statistical confidence measured through Monte-Carlo randomization with 100 replacements. Overall, 133 amino acid positions across gp120 were detected to have differential entropy between India and South Africa, of which 83 sites had higher entropy in South Africa while 50 sites had higher entropy in India (Figure 4B). As indicated in Figure 4B, many of the differential entropy sites are located in the surface accessible V1-V2 and V3 regions, also targeted by several bnAbs. To assess the entropy differences associated with bnAb contact sites, we plotted entropy data for South Africa, India and HIV-1 subtype C overall, as reported in the HIV-LANL database. To prevent bias because of hypervariable nature of certain gp120 domains, the residue positions present in hypervariable regions were removed from the subsequent comparisons. While no difference was observed in entropy profiles at key bnAb sites between overall subtype C entropy and that observed in South Africa (Mann-Whitney test, $p = 0.0759$), corresponding entropy profile in India was significantly different both from overall subtype C ($p<0.0001$) and South Africa ($p<0.0001$). These observations suggest a differential pattern of conservation across the two populations and thus may result in variable breadth of neutralization by these bnAbs (Bai et al. 2019).

Variation in abundances of epitopes associated with resistance and susceptibility to bnAbs.

Next, we examined the abundance of HIV- 1 subtype C resistance phenotype as defined in the CATNAP database of key bnAbs having varied epitope specificities in *gp120* across different countries. We took sequence datasets of different counties (www.hiv.lanl.gov) and calculated the frequency of key amino acid residues associated with the sensitivity and resistance to individual bnAbs. We examined the select key bnAbs targeting CD4bs (VRC01, VRC07, 3BNC117 and N6), V1V2 region (PG9, PG16, PGT145, PGDM1400 and CAP256.VRC26.25) and V3 supersite (PGT121, PGT128 and 10-1074). As shown in Figure 5, we found considerable variation in the abundance of amino acid residues that form epitopes associated with neutralization resistance to bnAbs with unique specificity. Variation in abundances of the following residues that form epitopes to bnAbs with unique specificities were observed in geographically divergent globally circulating HIV-1 subtype C (Figure 5). For CD4bs directed bnAbs: variation in the sensitivity to global HIV-1 subtype C to VRC01 was found to be majorly associated with N234 and S364. Similarly, for V1V2 directed bnAbs: K130, I161, Q170, Y173, S291, T297, N332 and N340 associated with significant variation in PGT145 sensitivity; K130, I161,I165, V169 and N332 (most significant) associated

with significant variation in PGDM1400 sensitivity; K130, I161, R166, V169, Q170 and N332 (most significant) associated with significant variation in PG9 sensitivity; K130, I161 and K171 associated with significant variation in PG16 sensitivity; I165. R166, V169 and N332 associated with significant variation in CAP256.VRC26.25 sensitivity was observed. Finally, for V3 directed bnAbs: K155, I165, R252, N289, E293, I307, H330, N332, S334, A336 and N448 associated with significant variation in PGT121 sensitivity; R252, N295, N300, H330, S334, A336, Q344, and T415 associated with significant variation in PGT128 sensitivity; K155, N156, I165, N230, T240, R252, K282, I307, A316, H330, S334, A336, Q344 and N448 associated with significant variation in 10-1074 sensitivity were observed. Taken together, our observation indicates the existence of variation in the abundance of key bnAb contact sites across global circulating HIV-1 subtype C which further highlights that majority of the bnAbs likely would be effective when administered in proper combination against diverse region-specific circulating HIV-1 subtype C.

Evidence of accumulation of bnAb resistance phenotype in globally circulating HIV-1 C over time.

With abundance of bnAb neutralization data against specific envelope sequences obtained *in vitro* as well as through clinical trials, several machine learning-based algorithms are increasingly becoming available that can predict probable sensitivity to bnAbs on the basis of gp120 sequences. In the present study we employed one such recently published algorithm bNAb-ReP (Rawi et al. 2019) to predict sensitivity of 1466 gp120 sequences selected in the present study from countries India, China, South Africa, Malawi, Zambia and Tanzania to following bnAbs: 3BNC117, VRC01, VRC07, PGT145, CAP256:VRC26.25, PGDM1400, PG9, PG16, PGT121, PGT128 and 10-1074. The prediction data were plotted along with country-wise *in vitro* data available through CATNAP database for a total of 283 sequences as indicated in Table 2. As indicated in Figure 6, probability values greater than 0.5 point towards sensitivity to bnAbs while those lower than 0.5 indicate probable resistance. For VRC01, while CATNAP database indicated unequivocal sensitivity of majority of sequences from all 6 countries, bNAb-ReP predicted a significant fraction of sequences from African countries to be resistant. Predictions for VRC26.25, 10-1074, PGT121 and PGDM1400 matched those reported in the CATNAP database. Similar to VRC01, predictions for 3BNC117, PGT128, PG9, PG16 and VRC07 did not match the data from CATNAP and indicated probable resistance in many of the sequences from all 6 countries. To assess if bnAb sensitivity differed over time,

prediction data for each of the 11 bnAbs was plotted against three periods based on the reporting date of the sequences (Figure 7). The three periods considered were 1986-2000 (N=244), 2001-2010 (N=1187) an 2011-2019 (N=333). Except for PGT121, PGT128 and 10-1074, all bnAbs showed significant decrease in sensitivity over time. Despite this decrease, most sequences were predicted to be susceptible to PGDM1400 and VRC26.25. However, susceptibility to 3BNC117, VRC01, VRC07, PGT145, PG9 as well as PG16 was predicted to have reduced significantly over the period of time assessed.

Discussion

Given that subtype C accounts for approximately half of the global HIV burden, limited information on the intra-clade C *env* diversity and its association with variation in their neutralization phenotypes exists. In the present study, we compared the genetic attributes of the globally circulating HIV-1 subtype C *gp120* sequences that differentiate the region-specific intra-clade HIV-1 subtype C neutralization diversity. For this, we used the available information in the HIV database (www.hiv.lanl.gov) and established algorithm (Rawi et al. 2019) to suitably predict the association between genetic features that potentially dissects the HIV-1 intra-clade C neutralization diversity. Although there is a disparity between the number of existing region-specific unique HIV-1 subtype C sequences in the database, nonetheless, region-specific distinct genetic clustering was observed by phylogenetic analysis of *gp120* amino acid sequences. It is to be noted that our analysis was based on one sequence per individual to avoid any sampling bias on an individual level in our analysis. The region-specific subtype C *gp120* divergence could possibly be due to diversity in the population level across geography (Hraber et al. 2014). Indeed, studies have shown that HIV-1 can selectively incorporate broad range of biologically active host proteins in the process of viral egress, which can potentially exhibit altered pathogenicity and neutralization phenotypes (Burnie and Guzzo 2019). This also indicates possible association between intra clade C genetic diversity with ethnically distinct population.

The evolutionary genetic drift within subtype C that we observed from phylogenetic analysis is likely due to differential host characteristics which include immune response and differential genetic bottlenecks. For example, *gp120* sequences from Asian countries were found to demonstrate monophylectic clustering compared to that observed with those obtained from African countries. A number of studies (Deshpande et al. 2016; Ringe et al. 2012; van Gils et al. 2011) have demonstrated the role of loop length in the hypervariable regions, (with particular reference to V1V2 region), charge and N-linked glycosylation on altered neutralization phenotype. In the present study, we observed that while there is an existence of region-specific variation in V1V2 loop length, V4 loop length variation between subtype C sequences was found to be most profound across geographic boundaries. In addition, we also observed that the average *gp120* length of African countries was found to be smaller compared to other regions. Our data demonstrates subtle but significant variations in these attributes thereby indicating that a common set of bnAbs are not

likely to be equally effective against the HIV-1 subtype C circulating globally. The above conclusion was further substantiated by our observation where we found significant variation in the entropy profiles (variation in sites/positions associated with different bnAbs) between the geographically distinct *gp120* sequences. With limited data available in CATNAP, indeed we found evidence of variation in susceptibility region-specific clade C to different bnAbs, which substantiate our observation. Interestingly, as reported elsewhere (Bouvin-Pley et al. 2013; Bouvin-Pley et al. 2014; Hake and Pfeifer 2017), our data also predicted potential likelihood of accumulation of resistance phenotype overtime to existing bnAbs. This observation indicate that it is necessary for continuous surveillance of evolving viruses in the context of subtype C along towards prioritizing bnAb combinations that will optimally dissect and overcome the evolving genetic diversity. Interestingly, a similar accumulation in ART resistance has been documented and studies that concurrently evaluate the interplay of these two evolutionary patterns, heretofore considered to be mutually exclusive, may highlight novel and synergistic therapeutic strategies. In summary, the differences in HIV-1 clade C *gp120* sequences observed herein indicate disparate and distinctly evolving clusters within clade C with differential predicted responses to bnAbs. Elucidation of neutralization diversity of subtype C particularly in context of evolution of *gp120* overtime will be essential for selecting appropriate bnAb combination for effective prophylaxis and treatment and also in informing rational vaccine design. In summary, our study highlights that towards developing HIV-1 bnAbs as products for prevention and treatment, continued surveillance of the evolution of genetic features with particular reference to *env* gene that are targets of neutralizing antibodies of globally circulating HIV-1 remains crucial for the identification and prioritization of combination of bnAbs that would best suit to provide maximal geography and population-specific neutralization coverage.

Materials and Methods:

Retrieval of gp120 sequences.

Sequences for the gp120 gene were retrieved from manually curated LANL HIV database (www.hiv.lanl.gov). Briefly, HIV-1 subtype C nucleotide sequences fully covering the genomic region 6225-7758 (as per HXB2 numbering) were retrieved and subsequently filtered with a one sequence per individual filter criterion. Sequence entries without any information regarding the sample source country were excluded. Multiple sequence alignment for the amino acid sequences along with HXB2 sequence (GenBank: K03455.1) was produced with Gene cutter ("Gene Cutter," LANL) . Gene Cutter clips the coding regions from unaligned nucleotide sequences and produces amino acid alignments based on Hmmer v 2.32 algorithm with a training set of the full-length genome alignment. Alignments were manually curated using Bioedit v7.2.5 (Hall 1999). Sequences with internal stop codons were discarded.

Phylogenetic analysis.

The number of amino acid differences per site were estimated by averaging over all sequence pairs between different countries using Molecular Evolutionary Genetics Analysis software (MEGA v.10) (S. Kumar et al. 2018b). The rate variation among sites was modelled with a gamma distribution (shape parameter = 1). This analysis involved 1814 amino acid sequences. Phylogenetic trees were generated for the amino acid alignments with iqtree under 'HIVb' model with estimated Y parameters and number of invariable sites (Nguyen et al. 2015). Robustness of the tree topology was further assessed by SH-aLRT as well as 1000 ultrafast bootstrap replicates implemented in iqtree. A subtree consisting of 251 sequences were again constructed as mentioned previously.

Estimation of the variable loop properties and potential N linked glycosylation sites.

Variable loop regions for V1 (131-157: HXB2 numbering), V2 (158-196), V3 (296-331), V4 (386-417) and V5 (460-469) were retrieved from amino acid alignments with Bioedit v7.2.5. Each of the loop datasets were then processed with custom bash/awk scripts to generate length statistics. The length distributions were further assessed and compared by Kruskal Wallis test followed by Dunn's multiple comparison test.

Cumulative variable loop charge values were predicted for each of the sequences with custom bash scripts wherein, Lysine (K), Arginine (R) and Histidine (H) residues were assigned +1 values each while Aspartic acid (D) and Glutamic acid (E) were assigned -1 values each. Potential N-linked glycosylation sites were predicted in amino acid sequence datasets with N-GlycoSite tool hosted at the HIV-LANL database (Zhang et al. 2004) . Prevalence of each of the pNLG sites under study were calculated using custom bash/awk scripts and were further assessed statistically using Fisher's exact test. Country-wise pNLGs abundance heatmap was plotted using 'pheatmap' package in R.

Entropy analysis.

Shannon entropy for selected sequence data sets was generated using Entropy-two tool following 100 randomizations with replacement (Efron and Tibshirani 1991; Gaschen et al. 2002) (www.hiv.lanl.gov. Key sites for interaction with bnAbs VRC01, VRC03, VRC07, VRC13, CAP256:VRC26.25, PGDM1400, PG9, PG16, PGT121 and PGT128 were derived from CATNAP database. HIV-1 clade C overall entropy bed graph was derived from Genome browser on HIV-LANL database.

Analysis of bnAb epitope contact sites.

Specific epitope contact sites/positions along with documented variants imparting sensitivity or resistance phenotype were retrieved from CATNAP database for bnAbs: VRC01, VRC07, PGT121, PGT128, PGT145, PG9, PG16, VRC26.25, 3BNC117, 10-1074 and N6. Frequency of such resistant variants were calculated across sequences from countries with more than 20 sequences available, using in-house bash scripts. Contact site resistance heatmaps were prepared with Circos v0.61 (Krzywinski et al. 2009).

Prediction of bnAb sensitivity with bNAb-ReP.

Sensitivity of contact sites to bnAbs (3BNC117, VRC01, VRC07, CAP256:VRC26.25, PGDM1400, PG9, PG16, PGT145, PGT121, PGT128 and 10-1074) were predicted with bNAb-ReP tool (Rawi et al. 2019). Neutralization data corresponding to the selected bnAbs was retrieved from CATNAP database (www.hiv.lanl.gov).

The temporal prediction data was stratified into approximately three decades as follows: 1986-2000 (N=244), 2001-2010 (N=1187) and 2011-2019 (N=333).

Statistical analyses and data presentation.

Statistical analyses for variable loop length distributions were performed using GraphPad Prism version 5.01 for Windows, GraphPad Software, San Diego California USA. Statistical Comparison of Fisher's test for pNLG sites as well as abundance of bnAb resistance associated residues was performed through R statistical computing software (v3.4.0) and R studio v1.0.143 (R. Team 2015; R. C. Team 2018). Phylogenetic trees were visualised and edited with the R package 'Graphlan' (Asnicar et al. 2015). Variable entropy positions were plotted on prefusion gp120 envelope model derived from PDB:5U7O in Chimera v1.14 (Pettersen et al. 2004). Plots depicting variable region characteristics, entropy differences and bNAb sensitivity predictions were prepared using 'ggplot2' package in R (Analysis.. 2016). Trend analysis for predicted bnAb sensitivity was performed by Jonckheere-Terpstra test implemented in R statistical software.

Acknowledgements.

We thank our laboratory members for providing valuable inputs and suggestions. The authors wish to acknowledge the funding support from the Wellcome Trust/DBT India Alliance Team Science Grant (IA/TSG/19/1/600019), Science & Engineering Research Board, Department of Science & Technology, Government of India (CRG/2019/002939) and Department of Biotechnology, Government of India (BT/PR24520/MED/29/1222/2017). ICMR-NIRRH is acknowledged for server support for computation analysis. IAVI's work was made possible by generous support from many donors, including the Bill & Melinda Gates Foundation, the Ministry of Foreign Affairs of Denmark, Irish Aid, the Ministry of Finance of Japan, the Ministry of Foreign Affairs of the Netherlands, the Norwegian Agency for Development Cooperation (NORAD), the United Kingdom Department for International Development (DFID), and the United States Agency for International Development (USAID). The full list of IAVI donors is available at www.iavi.org. The contents are the responsibility of the International AIDS Vaccine Initiative and do not

necessarily reflect the views of USAID or the United States Government. We thank Prof Gagandeep Kang, Translational Health Science & Technology Institute for support.

Table and Figure Legends:

Table 1: Details of Country-wise HIV-1 subtype C sequences retrieved from LANL-HIV database.

Table 2: Details of sequences analyzed, and their sources used for predication analysis of HIV-1 subtype C to different bnAbs.

Figure 1: Phylogenetic Analysis of HIV-1 subtype C gp120 amino acid sequence. A. A maximum likelihood tree depicting phylogenetic association of 1837 HIV-1 subtype C amino acid sequences depicted radially. The legend describes the country codes as well as the sequence distribution among the countries. **B.** A maximum likelihood subtree detailing phylogenetic association between sequences from South Africa and Those from India, China and Nepal. The black dots indicate nodes with corresponding SH-aLRT and bootstrap values.

Figure 2: Assessment of Variable region characteristics. A. variable region length: gp120 variable region (V1, V2, V1+V2, V4, V5) as well as entire gp120 lengths have been plotted on the Y axis against the Countries of origin indicated on the X axis. **B.** Variable region charge: gp120 variable region Charges for V1, V2 and V1+V2 as well as hypervariable regions within them have been plotted on the Y axis against the Countries of origin indicated on the X axis. P values have been indicated following a statistical analysis by Kruskal-Wallis test followed by Dunn's multiple comparison.

Figure 3: Assessment of potential N-linked glycosylation sites. A heatmap comparison of abundance of pNLG sites plotted on Y axis against countries plotted on X axis. Each pixel represents 1 pNLG site data from 1 country. Specific domains of pNLGs have been indicated along the Y axis. Color key represents correlation of color intensity with abundance of pNLGs ranging from 0 to 100.

Figure 4: Entropy analysis. A. Shannon Entropy difference (H(background)-H(query) (unit: bits) has been plotted on Y axis against each amino acid position on X axis, where ZA dataset was the background while IN dataset was the query. Different domains of gp120 have been indicated. Bars with red color indicate

positions with statistically significant entropy differences. Bars above 0 indicate higher entropy in South Africa while those below 0 indicate higher entropy in India. **B.** Variable entropy positions are plotted on prefusion gp120 envelope model derived from PDB:5U7O. gp120 domains have been color coded. Residue position highlighted in red indicate statistically significantly higher entropy in India compared to South Africa while those highlighted in blue indicate statistically significantly higher entropy in South Africa compared to India

(C) Shannon entropy (bits) at key sites for bnAbs VRC01, VRC03, VRC07, VRC13, CAP256:VRC26.25, PGDM1400, PG9, PG16, PGT121 and PGT128 excluding positions in hypervariable regions have been plotted for overall Subtype C, South Africa (ZA) and India (IN). Statistical comparison p values have been obtained following application of Mann-Whitney test.

Figure 5: Abundance of bnAb resistance associated residues. A circos heatmap depicting abundance of bnAb resistance associated residues was plotted for 11 bnAbs (VRC01, VRC07, PGT121, PGT128, PGT145, PG9, PG16, VRC26.25, 3BNC117, 10-1074 and N6) wherein each track indicates the country of origin. Each pixel on the circular track indicates a specific residue position colored as per abundance of resistance causing residues at that position as per the color key.

Figure 6: Prediction of bnAb sensitivity across different countries. Each panel indicates available country-wise CATNAP data plotted next to country-wise prediction data for available sequences for 3BNC117, VRC01, VRC03, VRC07, VRC13, CAP256:VRC26.25, PGDM1400, PG9, PG16, PGT145, PGT121, and PGT128 and 10-1074. In the CATNAP data panels, country-wise violin plots have been inlayed with boxplots against reported IC50 (µg/mL) values. Black dots indicate outliers while red dots indicate median values. Red background zone indicates bnAb resistance (IC50 >50 µg/mL) while green background zone indicates bnAb sensitivity (IC50 < 50 µg/mL). In the bNAb-ReP data panels, country-wise violin plots have been inlayed with boxplots against probability of neutralization as predicted by bNAb-ReP. Black dots indicate outliers while red dots indicate median values. Red background zone indicates probable bnAb resistance (Neutralization probability < 0.5) while green background zone indicates probable bnAb sensitivity (Neutralization probability > 0.5).

Figure 7: Assessment of predicted bnAb sensitivity over time. Each panel indicates cumulative prediction data for available sequences for 3BNC117, VRC01, VRC03, VRC07, VRC13, CAP256:VRC26.25, PGDM1400, PG9, PG16, PGT145, PGT121, and PGT128 and 10-1074 plotted against 3-time periods as follows: 1986-2000 (N=244), 2001-2010 (N=1187) and 2011-2019 (N=333). P values indicate trend analysis performed by Jonckheere-Terpstra test.

References

Analysis.., Wickham H (2016). ggplot2: Elegant Graphics for Data (2016), 'ggplot2: Elegant Graphics for Data Analysis.', *Springer-Verlag New York.* , (ISBN 978-3-319-24277-4).

Asnicar, F., et al. (2015), 'Compact graphical representation of phylogenetic data and metadata with GraPhlAn', *PeerJ,* 3, e1029.

Bai, H., et al. (2019), 'The breadth of HIV-1 neutralizing antibodies depends on the conservation of key sites in their epitopes', *PLoS Comput Biol,* 15 (6), e1007056.

Binley, J. M., T. Wrin, B. Korber, M. B. Zwick, M. Wang, C. Chappey, G. Stiegler, R. Kunert, S. Zolla-Pazner, H. Katinger, C. J. Petropoulos, and and Burton., D. R. (2004), 'Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies.', *J Virol,* 78, 13232–52.

Bouvin-Pley, M., et al. (2013), 'Evidence for a continuous drift of the HIV-1 species towards higher resistance to neutralizing antibodies over the course of the epidemic', *PLoS Pathog,* 9 (7), e1003477.

Bouvin-Pley, M., et al. (2014), 'Drift of the HIV-1 envelope glycoprotein gp120 toward increased neutralization resistance over the course of the epidemic: a comprehensive study using the most potent and broadly neutralizing monoclonal antibodies', *J Virol,* 88 (23), 13910-7.

Bunnik, E. M., et al. (2010), 'Adaptation of HIV-1 envelope gp120 to humoral immunity at a population level', *Nat Med,* 16 (9), 995-7.

Buonaguro, L., Tornesello, M. L., and Buonaguro, F. M. (2007), 'Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications', *J Virol,* 81 (19), 10209-19.

Bures, R., et al. (2002), 'Regional clustering of shared neutralization determinants on primary isolates of clade C human immunodeficiency virus type 1 from South Africa', *J Virol,* 76 (5), 2233-44.

Burnie, J. and Guzzo, C. (2019), 'The Incorporation of Host Proteins into the External HIV-1 Envelope', *Viruses,* 11 (1).

Caskey, M., Klein, F., and Nussenzweig, M. C. (2019), 'Broadly neutralizing anti-HIV-1 monoclonal antibodies in the clinic', *Nat Med,* 25 (4), 547-53.

Caskey, M., et al. (2015), 'Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117', *Nature*.

Caskey, M., et al. (2017), 'Antibody 10-1074 suppresses viremia in HIV-1-infected individuals', *Nat Med,* 23 (2), 185-91.

DeLeon, O., et al. (2017), 'Accurate predictions of population-level changes in sequence and structural properties of HIV-1 Env using a volatility-controlled diffusion model', *PLoS Biol,* 15 (4), e2001549.

Deshpande, S., et al. (2016), 'HIV-1 clade C escapes broadly neutralizing autologous antibodies with N332 glycan specificity by distinct mechanisms', *Retrovirology,* 13 (1), 60.

Doores, K. J., et al. (2015), 'Two classes of broadly neutralizing antibodies within a single lineage directed to the high-mannose patch of HIV envelope', *J Virol,* 89 (2), 1105-18.

Efron, B. and Tibshirani, R. (1991), 'Statistical data analysis in the computer age', *Science,* 253 (5018), 390-5.

Escolano, A., Dosenovic, P., and Nussenzweig, M. C. (2017), 'Progress toward active or passive HIV-1 vaccination', *J Exp Med,* 214 (1), 3-16.

Gaschen, B., et al. (2002), 'Diversity considerations in HIV-1 vaccine selection', *Science,* 296 (5577), 2354-60.

Gnanakaran, S., et al. (2007), 'Clade-specific differences between human immunodeficiency virus type 1 clades B and C: diversity and correlations in C3-V4 regions of gp120', *J Virol,* 81 (9), 4886-91.

Hake, A. and Pfeifer, N. (2017), 'Prediction of HIV-1 sensitivity to broadly neutralizing antibodies shows a trend towards resistance over time', *PLoS Comput Biol,* 13 (10), e1005789.

Hall, T. (1999), 'BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. ', *Nucleic Acids Symp. ,* (https://doi.org/citeulike-article-id:691774).

Han, C., et al. (2020), 'Key Positions of HIV-1 Env and Signatures of Vaccine Efficacy Show Gradual Reduction of Population Founder Effects at the Clade and Regional Levels', *mBio,* 11 (3).

Hemelaar, J., et al. (2019), 'Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis', *Lancet Infect Dis,* 19 (2), 143-55.

Hraber, P., et al. (2014), 'Impact of clade, geography, and age of the epidemic on HIV-1 neutralization by antibodies', *J Virol,* 88 (21), 12623-43.

Korber, B., et al. (2001), 'Evolutionary and immunological implications of contemporary HIV-1 variation', *Br Med Bull,* 58, 19-42.

Krzywinski, M., et al. (2009), 'Circos: an information aesthetic for comparative genomics', *Genome Res,* 19 (9), 1639-45.

Kulkarni, S. S., et al. (2009), 'Highly complex neutralization determinants on a monophyletic lineage of newly transmitted subtype C HIV-1 Env clones from India', *Virology,* 385 (2), 505-20.

Kumar, R., et al. (2018a), 'Broadly neutralizing antibodies in HIV-1 treatment and prevention', *Ther Adv Vaccines Immunother,* 6 (4), 61-68.

Kumar, S., et al. (2018b), 'MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms', *Mol Biol Evol,* 35 (6), 1547-49.

Li, M., 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 (23), 11776-90.

Lynch, R. M., et al. (2009), 'Appreciating HIV type 1 diversity: subtype differences in Env', *AIDS Res Hum Retroviruses,* 25 (3), 237-48.

Mendoza, P., et al. (2018), 'Combination therapy with anti-HIV-1 antibodies maintains viral suppression', *Nature,* 561 (7724), 479-84.

Nguyen, L. T., et al. (2015), 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Mol Biol Evol,* 32 (1), 268-74.

Nishimura, Y. and Martin, M. A. (2017), 'Of Mice, Macaques, and Men: Broadly Neutralizing Antibody Immunotherapy for HIV-1', *Cell Host Microbe,* 22 (2), 207-16.

Novitsky, V., et al. (2002), 'Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design?', *J Virol,* 76 (11), 5435-51.

Ogert, R. A., et al. (2001), 'N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism', *J Virol,* 75 (13), 5998-6006.

Pettersen, E. F., et al. (2004), 'UCSF Chimera--a visualization system for exploratory research and analysis', *J Comput Chem,* 25 (13), 1605-12.

Rawi, R., et al. (2019), 'Accurate Prediction for Antibody Resistance of Clinical HIV-1 Isolates', *Sci Rep,* 9 (1), 14696.

Ringe, R., Phogat, S., and Bhattacharya, J. (2012), 'Subtle alteration of residues including N-linked glycans in V2 loop modulate HIV-1 neutralization by PG9 and PG16 monoclonal antibodies', *Virology,* 426 (1), 34-41.

Seaman, M. S., et al. (2010), 'Tiered categorization of a diverse panel of HIV-1 Env pseudoviruses for assessment of neutralizing antibodies', *J Virol,* 84 (3), 1439-52.

Sok, D. and Burton, D. R. (2016), 'HIV Broadly Neutralizing Antibodies: Taking Good Care Of The 98', *Immunity,* 45 (5), 958-60.

--- (2018), 'Recent progress in broadly neutralizing antibodies to HIV', *Nat Immunol,* 19 (11), 1179-88.

Sutar, J., et al. (2019), 'Effect of diversity in gp41 membrane proximal external region of primary HIV-1 Indian subtype C sequences on interaction with broadly neutralizing antibodies 4E10 and 10E8', *Virus Res,* 273, 197763.

Taylor, B. S., et al. (2008), 'The challenge of HIV-1 subtype diversity', *N Engl J Med,* 358 (15), 1590-602.

Team, R Core (2018), 'R: A language and environment for statistical computing.'.

Team, RStudio (2015), 'RStudio: Integrated Development for R.'.

van Gils, M. J., et al. (2011), 'Longer V1V2 region with increased number of potential N-linked glycosylation sites in the HIV-1 envelope glycoprotein protects against HIV-specific neutralizing antibodies', *J Virol,* 85 (14), 6986-95.

Zhang, M., et al. (2004), 'Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin', *Glycobiology,* 14 (12), 1229-46.

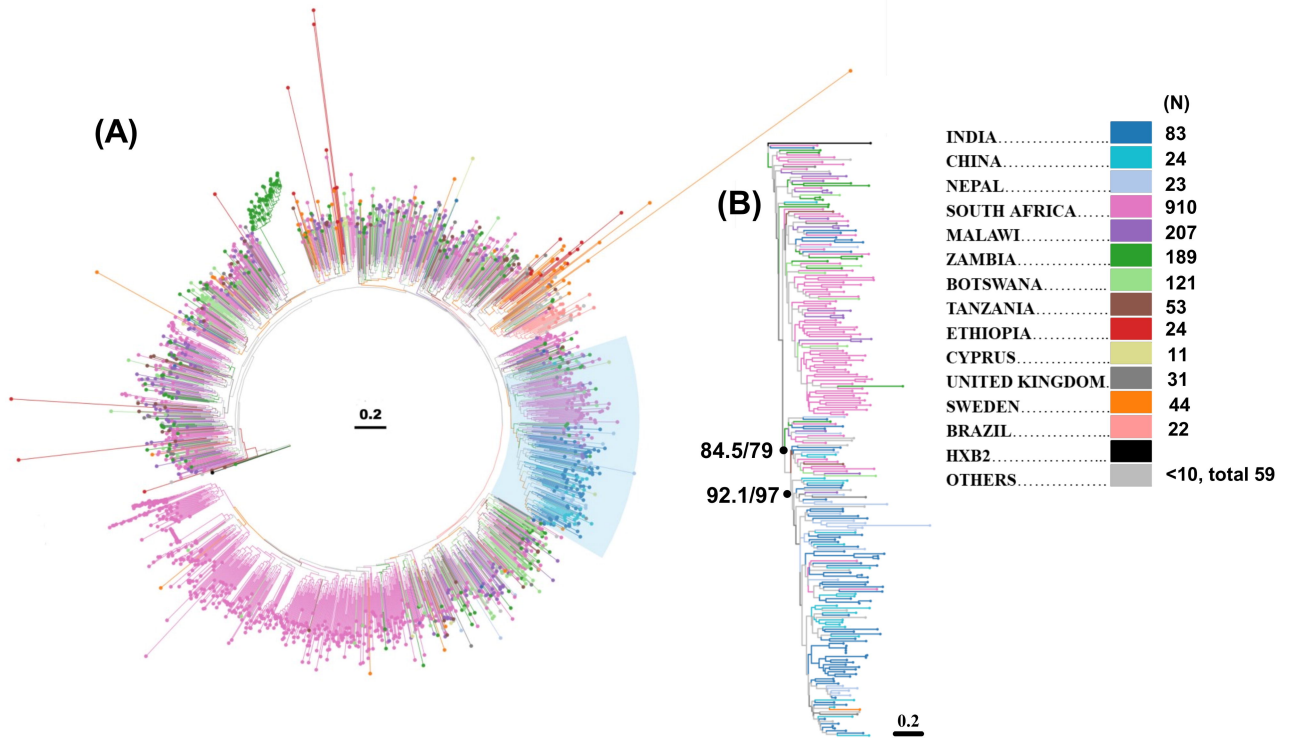
Table 1: Details of Country-wise sequences retrieved from LANL-HIV database

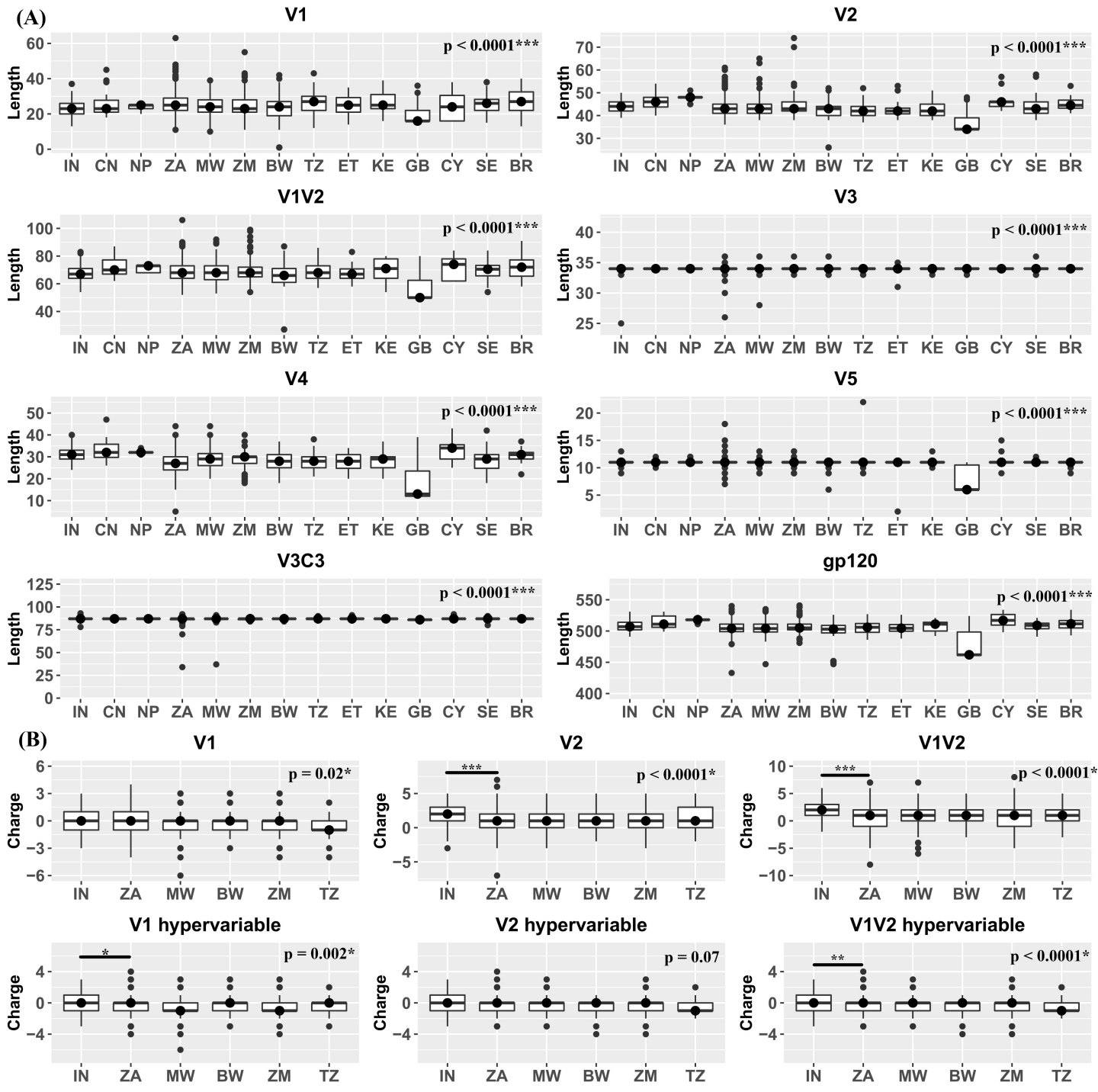
Country Name	Country code	No of sequences
South Africa	ZA	1015
Malawi	MW	210
Zambia	ZM	191
Botswana	BW	121
India	IN	84
United Republic of Tanzania	TZ	53
Sweden	SE	45
United Kingdom	GB	31
China	CN	24
Ethiopia	ET	24
Nepal	NP	23
Brazil	BR	22
Kenya	KE	13
Cyprus	CY	11
Others*	-	60

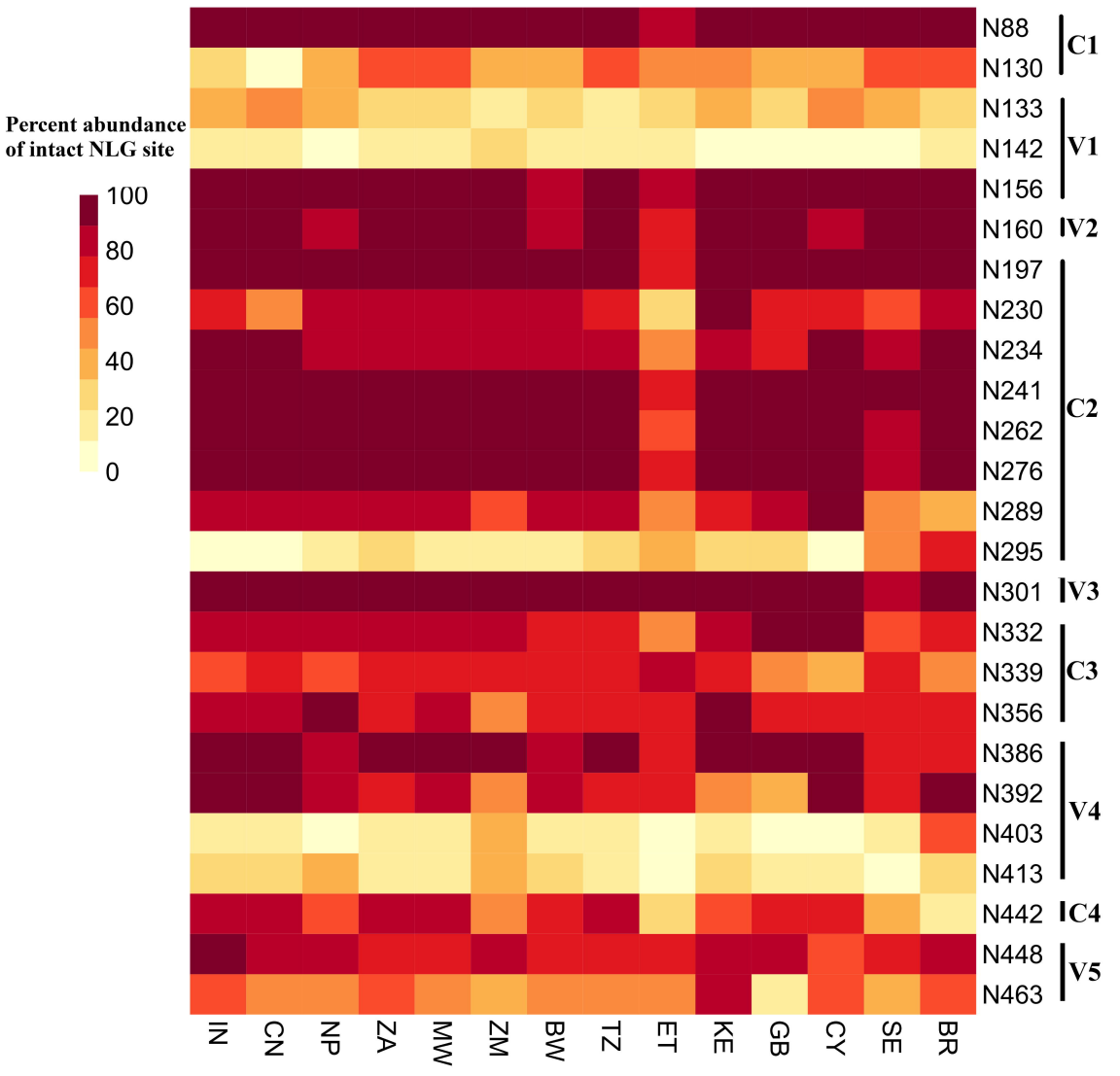
Footnotes: *- Republic of Angola (AO): 1, Belgium (BE): 2, Bulgaria (BG): 1, Burundi (BI): 8, Germany (DE): 1, Djibouti (DJ): 2, Denmark (DK): 1, Spain (ES): 6, Finland (FI): 5, France (FR): 3, Georgia (GE): 1, Gambia (GM): 2, Israel (IL): 4, Italy (IT): 1, Myanmar (MM): 1, Senegal (SN): 3, Somalia (SO): 1, Thailand (TH): 2, Uganda (UG): 1, United States (US): 7, Urugyuay (UY): 1, Yemen (YE): 1, Zimbabwe (ZW): 1, No Country: 4

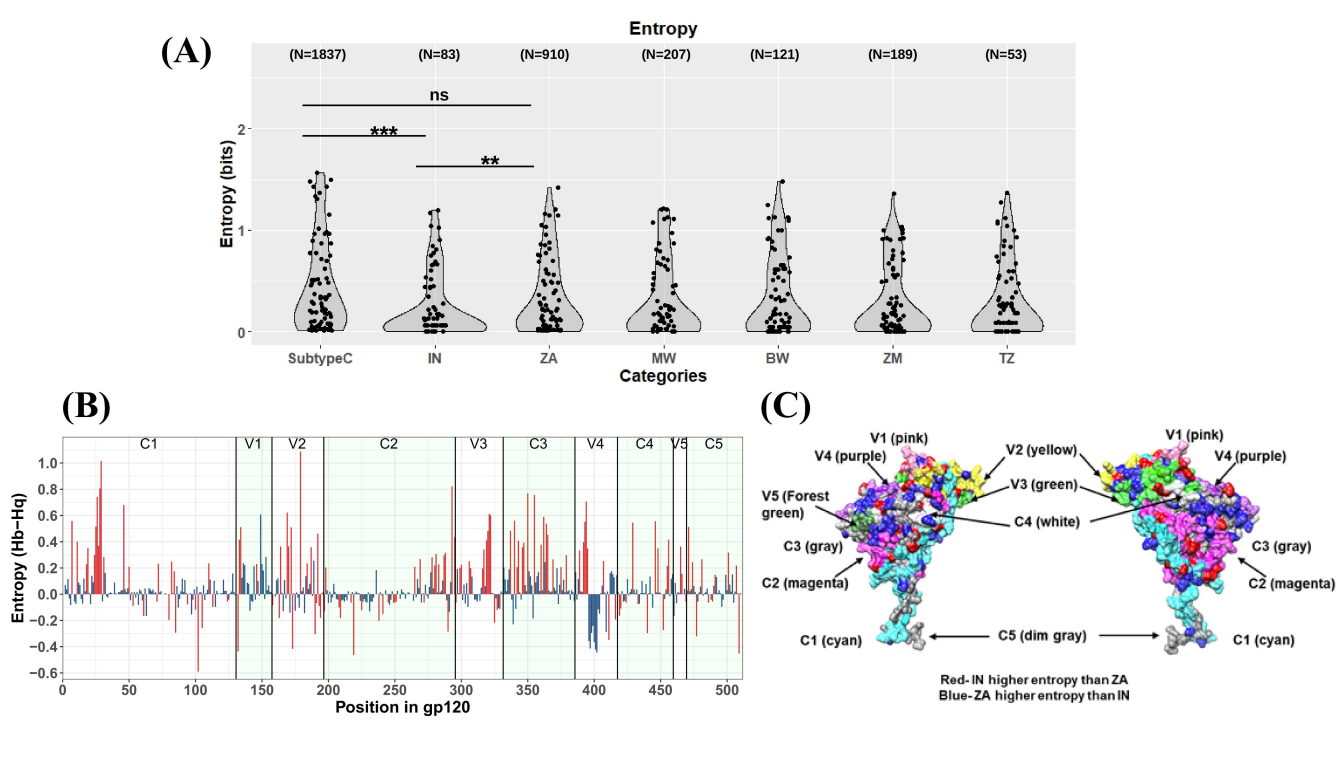
Table 2: Details of sequences analyzed, and their sources used for predication analysis of HIV-1 subtype C to different bnAbs.

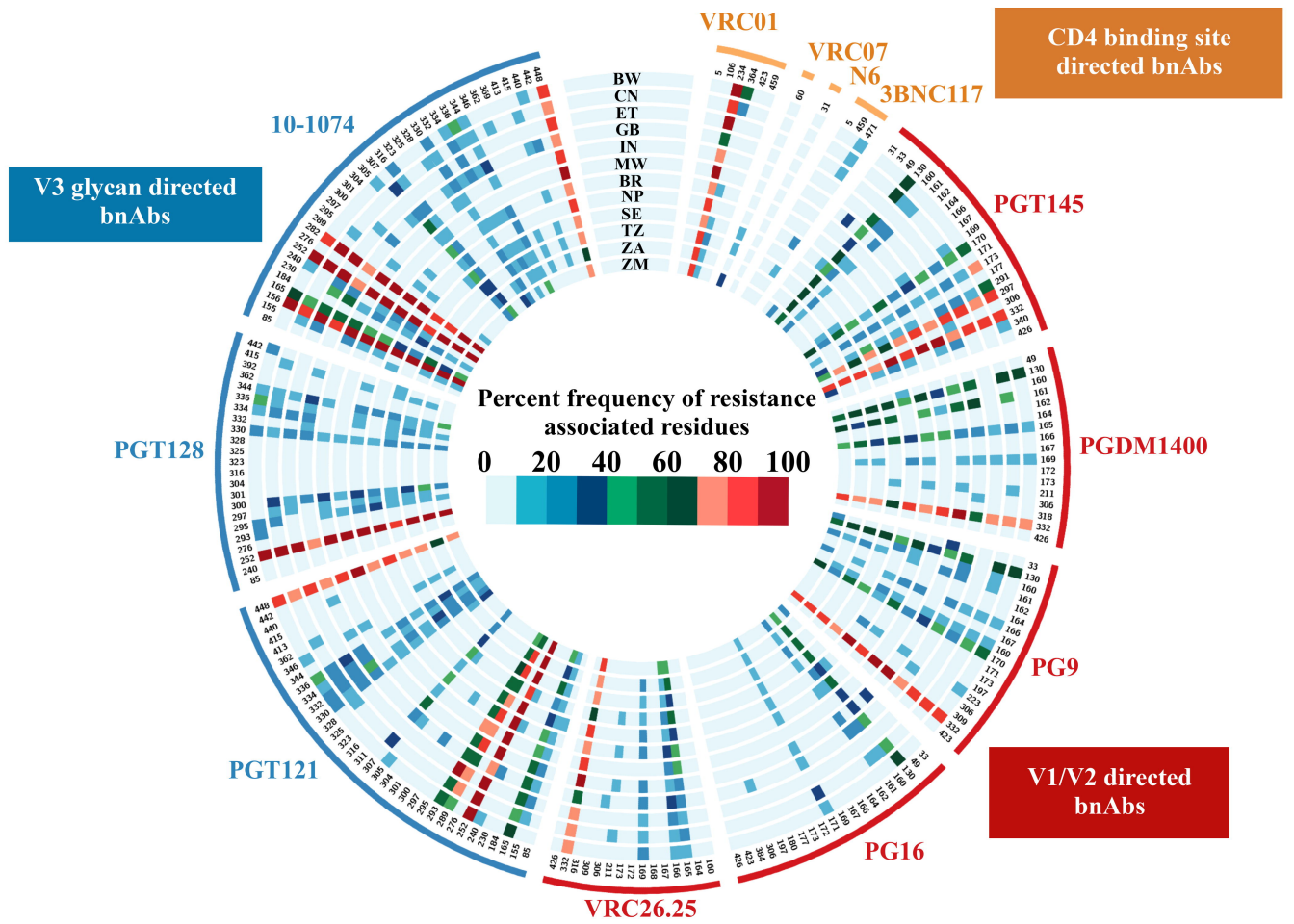
Country	Number of sequences	
	CATNAP	LANL-HIV (bNAb-ReP)
India	13	83
China	7	24
South Africa	146	910
Malawi	62	207
Zambia	23	189
Tanzania	32	53
Total	283	1466

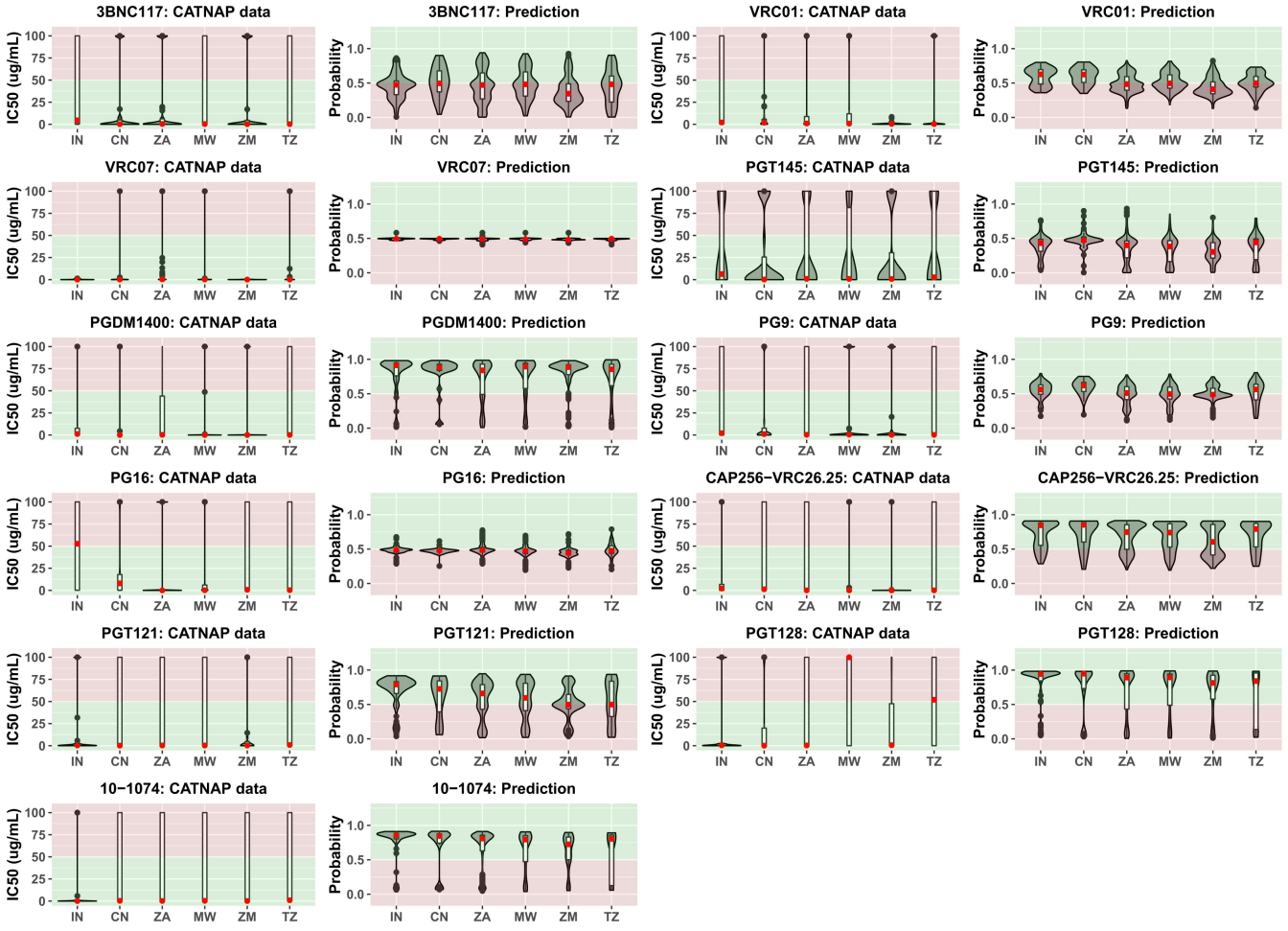

(A)
0.2
(B)
84.5/79
92.1/97
0.2
(N)
INDIA 83
CHINA 24
NEPAL 23
SOUTH AFRICA 910
MALAWI 207
ZAMBIA 189
BOTSWANA 121
TANZANIA 53
ETHIOPIA 24
CYPRUS 11
UNITED KINGDOM 31
SWEDEN 44
BRAZIL 22
HXB2
OTHERS <10, total 59

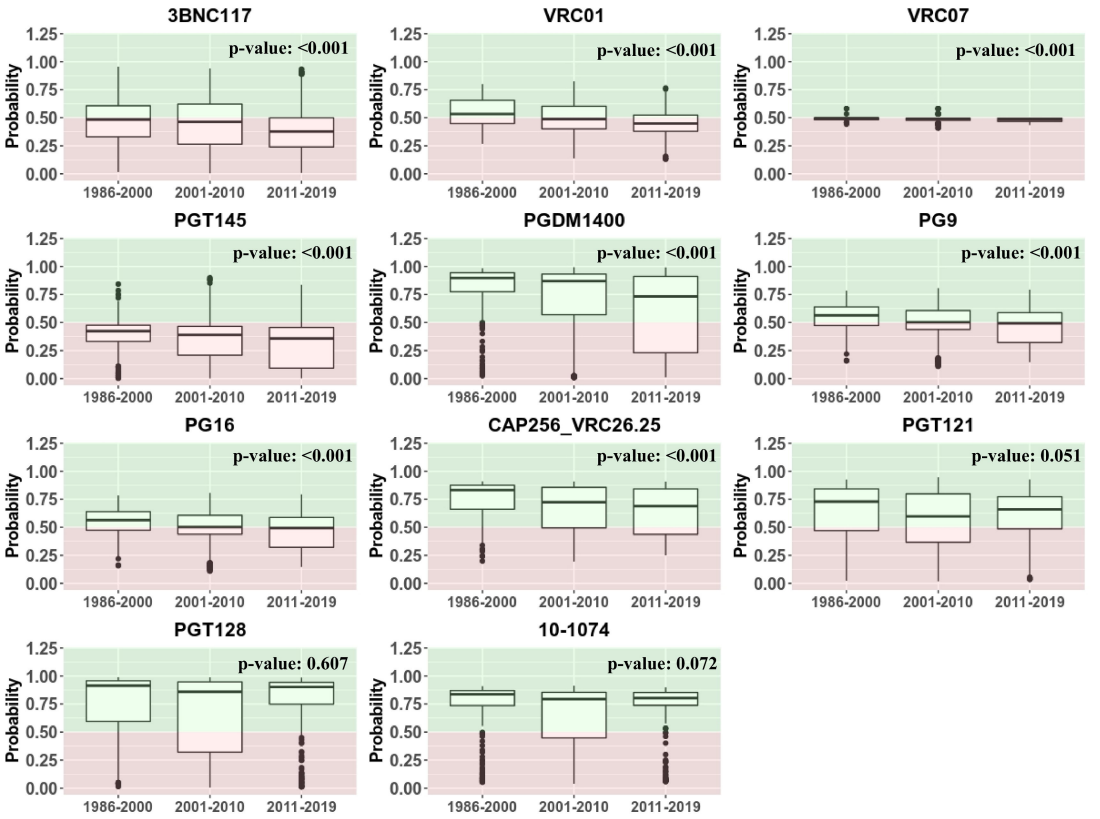

(A)
V1
V2
V1V2
V3
V4
V5
V3C3
gp120
Length
p < 0.0001***
IN CN NP ZA MW ZM BW TZ ET KE GB CY SE BR
(B)
V1
V2
V1V2
V1 hypervariable
V2 hypervariable
V1V2 hypervariable
Charge
p = 0.02*
p < 0.0001*
p = 0.002*
p = 0.07
IN ZA MW BW ZM TZ


Percent abundance of intact NLG site
100
80
60
40
20
0
N88
N130
N133
N142
N156
N160
N197
N230
N234
N241
N262
N276
N289
N295
N301
N332
N339
N356
N386
N392
N403
N413
N442
N448
N463
C1
V1
V2
C2
V3
C3
V4
C4
V5
IN
CN
NP
ZA
MW
ZM
BW
TZ
ET
KE
GB
CY
SE
BR


(A)
Entropy
(N=1837)
(N=83)
(N=910)
(N=207)
(N=121)
(N=189)
(N=53)
ns
***
**
Entropy (bits)
0
1
2
SubtypeC
IN
ZA
MW
BW
ZM
TZ
Categories
(B)
C1
V1
V2
C2
V3
C3
V4
C4
V5
C5
Entropy (Hb−Hq)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
50
100
150
200
250
300
350
400
450
500
Position in gp120
(C)
V1 (pink)
V4 (purple)
V5 (Forest green)
C3 (gray)
C2 (magenta)
C1 (cyan)
V2 (yellow)
V3 (green)
C4 (white)
C5 (dim gray)
Red- IN higher entropy than ZA
Blue- ZA higher entropy than IN


CD4 binding site directed bnAbs
VRC01
VRC07
N6
3BNC117
V3 glycan directed bnAbs
10-1074
PGT145
PGT128
PGDM1400
PG9
V1/V2 directed bnAbs
PGT121
PG16
VRC26.25
BW
CN
ET
GB
IN
MW
BR
NP
SE
TZ
ZA
ZM
Percent frequency of resistance associated residues
0 20 40 60 80 100


3BNC117: CATNAP data
3BNC117: Prediction
VRC01: CATNAP data
VRC01: Prediction
VRC07: CATNAP data
VRC07: Prediction
PGT145: CATNAP data
PGT145: Prediction
PGDM1400: CATNAP data
PGDM1400: Prediction
PG9: CATNAP data
PG9: Prediction
PG16: CATNAP data
PG16: Prediction
CAP256−VRC26.25: CATNAP data
CAP256−VRC26.25: Prediction
PGT121: CATNAP data
PGT121: Prediction
PGT128: CATNAP data
PGT128: Prediction
10−1074: CATNAP data
10−1074: Prediction
IC50 (ug/mL)
Probability
IN
CN
ZA
MW
ZM
TZ


3BNC117
p-value: <0.001
VRC01
p-value: <0.001
VRC07
p-value: <0.001
PGT145
p-value: <0.001
PGDM1400
p-value: <0.001
PG9
p-value: <0.001
PG16
p-value: <0.001
CAP256_VRC26.25
p-value: <0.001
PGT121
p-value: 0.051
PGT128
p-value: 0.607
10-1074
p-value: 0.072
Probability
0.00
0.25
0.50
0.75
1.00
1.25
1986-2000
2001-2010
2011-2019