Distinct antibody-mediated selection in a narrow-source HIV-1 outbreak among Chinese former plasma donors

Sophie M. ANDREWS, Yonghong ZHANG, Tao DONG, Sarah L. ROWLAND-JONES#, Sunetra GUPTA#, Joakim ESBJÖRNSSON**

Sophie M. ANDREWS - Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom. E-mail: sophie.andrews@ndm.ox.ac.uk
Yonghong ZHANG - Beijing You’an Hospital, Capital Medical University, Beijing, China. E-mail: 13810108505@163.com
Tao DONG - Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom. E-mail: tao.dong@imm.ox.ac.uk
Sarah L. ROWLAND-JONES - Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom. E-mail: sarah.rowland-jones@ndm.ox.ac.uk
Sunetra GUPTA - Department of Zoology, University of Oxford, Oxford, United Kingdom. E-mail: sunetra.gupta@zoo.ox.ac.uk
Joakim ESBJÖRNSSON - Nuffield Department of Medicine, University of Oxford, NDMRB, Old Road Campus, Roosevelt Drive, Headington, Oxfordshire, United Kingdom, OX3 7FZ. joakim.esbjornsson@ndm.ox.ac.uk

*Corresponding author
# Joint senior author

Abstract Word count: 169
Main body word count: 3,405 (2,424 excluding the Methods section)
ABSTRACT

The HIV-1 envelope protein mutates rapidly to evade recognition and killing, and is a major target of the humoral immune response and in vaccine development. Identification of common antibody epitopes for vaccine development have been complicated by large variations on both virus (different infecting founder strains) and host genetic levels. We studied HIV-1 envelope gp120 evolution in 12 Chinese former plasma donors infected with a purportedly single founder virus, with the aim of identifying common antibody epitopes under immune selection. We found five amino acid sites to be under significant positive selection in ≥50% of the patients, and 22 sites housing mutations consistent with antibody-mediated selection. Despite strong selection pressure, some sites housed a limited repertoire of amino acids. Structural modelling revealed that most sites were located on the exposed distal edge of the Gp120 trimer, whilst wholly invariant sites clustered within the centre of the protein complex. Four sites, flanking the V3 hypervariable loop of the Gp120, represent novel antibody epitopes that may be suitable as vaccine candidates.

KEYWORDS: HIV-1, gp120, evolution, antibody, epitope
INTRODUCTION

The human immunodeficiency virus type 1 (HIV-1) glycoprotein Gp120 is a 120 kDa surface-expressed protein that is essential for viral entry into the cell. It is encoded by the env gene, and consists of five variable regions (V1-V5) interspersed between five conserved regions (C1-C5). The Gp120 forms heterodimers with Gp41 which themselves trimerise, studding the viral membrane at a density of around fourteen copies per virion. Whilst the cellular immune response against HIV-1 targets epitopes dispersed throughout the viral genome, the accessibility of Gp120 on the cell surface makes it the major target of humoral responses.

The humoral response against HIV-1 Gp120 develops rapidly within around four weeks of detectable plasma viral loads, but neutralising antibodies (NAbs) typically only develop after several months of infection. Around two hundred antibodies have been described that recognise the Gp120 protein (LANL Immunology Database; http://www.hiv.lanl.gov/content/immunology), and many of the epitopes cluster within the V3 loop. However, the interplay between Gp120 and the adaptive immune response is complex, and the role that antibodies play in the control of infection is a contentious issue. The loss of neutralising activity has been associated with faster disease progression in some individuals. In addition, studies in macaques have indicated that B lymphocyte depletion-associated reductions in NAb titre inversely correlate with viral load, suggesting that the humoral response may contribute at least in part to the control of viral replication. However, whilst NAbs do exert selection pressure on the virus, the breadth of response does not correlate with or predict progression to AIDS.

It is commonly believed that the reason why antibody responses may play a limited role in the control of HIV-1 is because the virus can mutate easily to escape neutralization by these
responses: as one antibody is evaded, new antibodies are raised but these are evaded again in an on-going cycle\textsuperscript{9,12-14}. This view is supported by the observation that HIV-1 is rarely susceptible to neutralisation by contemporaneous antibodies in early infection\textsuperscript{15,16}, whilst the same antibodies are able to effectively neutralise historic virus\textsuperscript{9,12,17}. However, in clinically latent infection, viral variants evolve susceptibility to neutralisation by contemporaneous NAbs, or to sera sampled much earlier in infection\textsuperscript{18-20}. It is therefore possible that antibody responses do play an important role in controlling HIV-1, at least in the latent phase, with re-emergence of variants occurring periodically as the associated NAb responses fall below a certain threshold but then are restored by stimulation by the variant\textsuperscript{21}.

Indeed, several apparent paradoxes in HIV-1 pathogenesis and the genetics of host susceptibility can be resolved by assuming that NAb s play an important role in the control of infection, as shown by a recent modelling study\textsuperscript{21}. Non-neutralising responses with Fc-related activities – including antibody-dependent cellular cytotoxicity (ADCC) or antibody-mediated cellular viral inhibition (ADCVI) – directed at epitopes of intermediate variability, may also help maintain chronicity of infection. This is consistent with the findings of studies in rhesus macaques demonstrating that simian immunodeficiency virus isolated during clinically latent infection remains susceptible to ADCVI responses from earlier plasma, despite no detectable contemporaneous, autologous neutralising response\textsuperscript{22}. A potential therapeutic approach to preventing disease progression may therefore be to develop vaccines that boost and maintain such partially cross-protective responses.

Here we studied \textit{gp120} evolution in a narrow-source cohort of former plasma donors (FPDs) from Henan Province in China (SM cohort)\textsuperscript{23}. The FPDs of Henan were infected with a narrow-source of virus through exposure to contaminated equipment, and transfusion with pooled red
blood cells during an illegal paid plasma donation scheme in the mid-1990s\textsuperscript{24,25}. Owing to the unusually homogeneous route and source of infection, and the narrow timeframe during which patients were exposed, the infecting founder is relatively conserved. This provided us with the unique opportunity to investigate where and how immunological selection pressure drives mutation within $gp120$. Our aim was not only to comprehensively map where antibodies are able to bind in natural infection, but also to understand the constraints placed on mutation in these regions, thereby testing whether antibody escape is a continuous linear process or one in which a limited number of epitopes are re-cycled over time. The identification of such epitopes of limited variability, which are nonetheless targets of natural immunity, may assist in the development of vaccination strategies that prevent viral escape: for example, by incorporating all possible substitution forms into the vaccine.
METHODS

Cohort characteristics and sampling

The SM cohort comprises HIV-1 patients from a small rural community in Henan province, China, as described previously. Between 1993 and 1995, the patients were infected with a narrow-source of subtype B’ virus during an illegal paid plasma donation scheme. Few individuals knew that they were infected until HIV screening programmes were implemented in China in 2004. Patients were then recruited to the cohort and gave informed consent for their samples to be used.

Patients with samples collected longitudinally over the course of HIV-1 infection permit detailed evolutionary analysis and identification of sites under selection. Twelve patients from the SM cohort with multiple samples collected between 2010 and 2014 were available for this study (Table S1). Ethical approval was obtained from Beijing You’an Hospital and the University of Oxford Tropical Ethics Committee (OxTREC).

HIV-1 env gp120 sequencing and sequence assembly

Viral RNA was isolated from cryopreserved plasma samples and purified using the QIAamp Viral RNA Extraction Kit (Qiagen) followed by reverse transcription using the SuperScript III Reverse Transcriptase System (Invitrogen). Env gp120 C2-V5 was amplified by nested touchdown PCR (primers listed in Table S2). Amplified DNA was purified using the MinElute Gel Extraction Kit (Qiagen), and ligated into a pCR4-TOPO sequencing vector using the TOPO TA Cloning Kit for Sequencing (Invitrogen). Chemically competent One Shot MAX Efficiency DH5α E. coli (Invitrogen) were transformed with the prepared plasmids, and cultured overnight at 37°C. Eighteen colonies were selected for colony PCR (M13F and M13R primers, Table S2), and the resulting products were purified using ExoSAP-IT (Affymetrix) and
sequenced to generate forward and reverse reads (Source BioScience). Contigs were assembled and controlled manually using Geneious v9.0.5\textsuperscript{26} (HXB2 \textit{gp120} positions 661-1455, accession number K03455). Sequences were multiple aligned using MUSCLE\textsuperscript{27}, and then manually edited in MEGA v6.06\textsuperscript{28}. Sequences with premature stop codons or frame-shifts were excluded to control for intra-patient clustering of sequences.

Inference of infecting founder strain

A consensus sequence was generated for each patient for each time-point, with an ambiguity threshold of 10\%. One sequence was selected per patient to generate a dataset with sequences evenly distributed across the sampling period. The sequences were aligned and codon-stripped to a final alignment length of 759 nucleotides. Patient SM007 was excluded from this analysis because it was not possible to conclusively rule out dual- or superinfection as preliminary data exploration demonstrated that sequences from this patient did not resolve monophyletically.

Bayesian Markov Chain Monte Carlo phylogenetic inference - implemented through BEAST v1.8.2 (\texttt{http://beast.bio.ed.ac.uk/beast/})\textsuperscript{29} - was used to estimate the time to the most recent common ancestor (tMRCA). Divergence time was estimated using the SRD06 model\textsuperscript{30}, and an uncorrelated lognormal relaxed molecular clock\textsuperscript{31} with a rate prior of 0.001 substitutions per site per year. The Markov chain was run for 100,000,000 generations to allow for adequate mixing, with posterior samples extracted every 10,000 generations. A burn-in period of 10\% was applied, and convergence of posterior probabilities was assumed once the effective sample size (ESS) of each parameter exceeded 200, as determined in Tracer v.1.5 (\texttt{http://tree.bio.ed.ac.uk/software/tracer/}). Three runs were combined in LogCombiner v1.8.2 (\texttt{http://beast.bio.ed.ac.uk/LogCombiner/}) and the mean root height of the tree was calculated. Phylogenetic trees were annotated in FigTree v.1.4.2 (\texttt{http://beast.bio.ed.ac.uk/figtree}).
As the SM cohort patients were infected with a narrow source of virus, the sequence of the MRCA was used as a surrogate for the infecting founder. The sequence of the reconstructed ancestor at the tree root from each run following burn-in was extracted, and a consensus sequence was generated from an alignment of these sequences (26,000 sequences) with an ambiguity threshold of 10%.

Viral subtyping

An unambiguous consensus sequence was generated from the sequences of each time-point for each patient. The sequences were then aligned to the LANL 2005 gp120 reference dataset (http://www.hiv.lanl.gov/), and viral subtyping was performed by Bayesian phylogenetic inference (BEAST v1.8.2). The generalised time-reversible (GTR) nucleotide substitution model plus invariant sites and gamma-distributed rate heterogeneity (GTR+I+G) was used, with a constant size coalescent tree prior, estimated base frequencies, and a strict molecular clock. Following exclusion of burn-in, TreeAnnotator v1.8.2 was used to determine the maximum clade credibility tree (MCC).

Selection pressure in env gp120

The ratio of non-synonymous (dN) to synonymous (dS) mutations was estimated for each codon in patient-specific alignments. Renaissance counting\textsuperscript{32,33} was implemented through BEAST v1.8.2, and the HKY85 nucleotide substitution model\textsuperscript{34}, three-site codon partitioning, a strict molecular clock with tip-dating of time stamped sequences were applied. Significant selection was defined as a 95% higher posterior density (HPD) range that did not encompass 1. An alignment representing all patients was then constructed from these data, and the dN/dS
estimates were combined across each aligned position. The proportion of patients with virus showing evidence of significant selection pressure in each cohort was calculated.

4 **Variant characterisation**

A variant was defined as any amino acid in any position in the alignment that differed from that present in the inferred founder. Owing to the extensive degree of variation, the hypervariable loops were conservatively stripped from the alignment prior to analysis (final length 179 amino acids). Major variants were defined as variants found in greater than 15% of the sequences. Whilst major variants are canonically defined as those present at a frequency greater than 5%, this value was conservatively tripled as the amplicon was approximately three times more variable than the full-length HIV-1 genome35.

12

**Structural modelling**

Homology modelling was implemented through SWISS-MODEL36-39 to map the translated SM cohort Gp120 consensus sequence to a cryo-electron microscopy (cryo-EM) crystal structure of a glycosylated HIV-1 envelope trimer (RCSB PDB 3J5M)40. The sites of interest were annotated on the modelled structure in PyMOL v1.7.4 (Schrödinger, LLC.).

18

**Data analysis**

Epitope mapping, selection mapping, variant characterisation, biophysical properties, statistical analysis, and plotting were all performed in R (v3.2.3)41 via the RStudio (v.0.98.1103) integrated development environment (http://www.rstudio.com/). The following libraries were used: dplyr42; scales43; gridExtra44; ggplot245; reshape246. All scripts are available from the authors on request.
RESULTS

Cohort characteristics

HIV-1 env gp120 sequences were recovered from 12 patients in the SM cohort (seven female and five male). The HLA types of those were representative of cohort frequencies. Sequence recovery was 74% from the available specimens, with ten of the patients yielding sequences from two or more time points (Fig. S1A). PCR success was associated with the viral load of the sample. The final dataset of 575 sequences is available from Genbank under accession numbers MF078678-MF079252.

Across the specimens sampled, median viral load was 7,388 copies ml⁻¹ of plasma (interquartile range (IQR): 1,612-30,403); median absolute CD4⁺ lymphocyte count was 337 cells µl⁻¹ (IQR: 248-400); and median CD4 percentage of lymphocytes was 24% (IQR: 15-30). Demographic and clinical characteristics of the patients sampled are shown in Table S1.

Maximum likelihood phylogenetic analysis showed a star-like relationship between the sequences, with short internal branches between patients, consistent with a narrow-source outbreak (Fig. S1B). Subtype analysis showed that all patient sequences clustered with the CRF15_01B strain: a circulating recombinant virus initially reported in Thailand composed of CRF01_AE with the majority of envelope being subtype B (Figure S1C). Bayesian phylogenetic analysis placed the origin of the SM cohort cluster at around January 1999 (95% HPD October 1988 to November 2006).

Five positions in Gp120 were under significant positive selection pressure in at least half of the patients irrespective of HLA profile.
The median evolutionary rate ratio of positions 1+2 to position 3 among the participants was 0.861 (IQR: 0.779-0.892), indicating an overall purifying selection. The overall ratio was <1 in all patients except SM021 (1.306). Next, we evaluated the intrapatient dN/dS ratio of each codon by renaissance counting. The majority of codons in Gp120 were under either significant positive or negative selection in seven of the ten patients with longitudinal samples (Fig. 1A), and neutral evolution was comparatively rare. Consistent with the 1+2:3 codon rate ratios, only one patient (SM021) appeared to have more residues under significant positive than negative selection pressure.

Within the variable loops, substantial negative selection could be seen in V3, but not in V4 and V5 (Fig. 1B). The negative selection in V3 corresponded with a marked increase in the density of neutralising antibody epitopes. Five codons, corresponding to positions T297, A337, S348, D415 and S468 in the HXB2 Gp120 (accession number K03455), were under significant positive selection in 50% or more of the patients irrespective of HLA profile (Figure 1A and Table S1). These sites were further mapped to a homology-modelled structure of the SM cohort Gp120 consensus, and clustered either within or immediately flanking the variable loops on the distal exposed edge of the protein complex (Fig. 1C).

**Significant selection pressure was exerted by the humoral response**

We next considered how the virus evolves in response to significant positive selection pressure. For each position in the partial Gp120 sequence, deviations from the inferred founder were detected and 288 unique amino acid variants were recorded across 128 variable sites (Fig. 2A). The remaining 51 positions were completely invariant (29%). Of the variants identified, many were present in a single or small collection of sequences and are likely of limited biological relevance. Major variants were therefore resolved, and to prevent overrepresentation by
particular patients, a single time-point was selected for each. Thirty-one major variants were detected in total, across 29 sites (Fig. 2A).

In patients with longitudinally-sampled sequences, the presence or absence of each major variant was recorded at each time point. Whether these sites were under significant positive selection pressure in that patient was also determined (Fig. 2B). Twenty-four of the 29 sites housing major variants were under significant positive selection pressure in at least one patient, and of these sites, a higher proportion exhibited fluctuating patterns of variant emergence than a consistent pattern wherein the variant was present at all time points sampled ($p<0.01$, two-tailed Fisher’s Exact Test).

Mapping these 24 sites to the homology-modelled structure of the SM cohort Gp120 consensus revealed that all but two were visible on the surface of the protein and likely accessible to antibodies (Fig. 2C). The exceptions were positions 345 and 424. Position 345 houses the major variant I345V, and is contained within the location of a known HLA-A11-restricted CTL epitope. This position was found only to be under significant positive selection pressure in patient SM176, who also expresses HLA-A11. Position 424 houses the CD4 binding site. The wholly invariant sites were similarly mapped, and the overwhelming majority clustered on the inner face of the quaternary structure.

Despite significant antibody-mediated selection pressure, some sites housed a limited repertoire of amino acids

We next considered the biophysical diversity of amino acids in each of the 22 sites of interest (Fig. 3). In all positions, the biophysical properties of the amino acid in the inferred founder were preserved in approximately 50% or more of the sequences. Little divergence was seen in
sites where the inferred founder residue was hydrophobic, with other properties being somewhat more variable. Tabulating the amino acids present in each site also reveals that seven positions flick back and forwards between just two or three amino acids.

**Four novel sites were identified and consistent with antibody activity**

The 22 sites were also cross-checked for existing antibody epitopes in the LANL Immunology Database (Table 1). Eighteen of the 22 sites were part of previously described antibody epitopes from different sources (human, mice, or both). In detail, five positions (T236, Q344, I345, V424, and D474) were contained within known human antibody epitopes, whilst 16 positions were contained within epitopes reported in mice. Four sites (V283, E290, L333, S334) flanking the V3 hypervariable loop – which itself houses the majority of antibody epitopes – were identified against which no antibodies have yet been reported. Two of these four sites (V283 and L333) primarily switch between I/V and I/L, respectively.
DISCUSSION

In line with previous observations\textsuperscript{48}, mapping the ratio of non-synonymous to synonymous substitutions showed that the majority of sites within the C2-V5 region of Env Gp120 were under negative selection in all but one patient out of ten. Whilst V4 and V5 loops exhibited a dearth of negative selection, the V3 hypervariable loop contained substantial negative selection. Of the five variable loops in Gp120, V3 is the most conserved with amino acid variation restricted to approximately 20\% of the loop’s residues\textsuperscript{49}. It is also likely that V3 is subject to stronger functional constraints due to its important role in co-receptor binding\textsuperscript{50-53}. Moreover, it has been shown that deletion of V3 abrogates viral infectivity\textsuperscript{54}.

Several sites within each patient showed evidence of significant positive selection, and five of these were common to at least half of the patients sampled. Structural modelling demonstrated that all but two of the 24 positively selected sites were found on exposed regions of the outer face of the protein complex. CTL epitopes in the HIV-1 Nef protein have been reported to cluster in hydrophobic regions\textsuperscript{55}, whilst more recent evidence suggests that their distribution may be random across the genome\textsuperscript{56}. Such strong clustering on the surface of the protein is therefore more consistent with antibody-mediated than CTL-mediated selection pressure. The exceptions in terms of surface exposure were positions 345 and 424, which were buried within the protein. Notably, position 345 was found to be under significant positive selection pressure in only one patient, SM176. This position is contained within a known HLA-A11-restricted CTL epitope, which is one of the HLA alleles expressed by patient SM176. It is therefore feasible that this variant has emerged in response to CTL-mediated selection pressure in this patient. Conversely, position 424 is important in CD4 binding\textsuperscript{57}, and is contained within a known human antibody epitope. Mutation of this residue to methionine has been shown to increase susceptibility to neutralisation\textsuperscript{58}.
We were able to assign most positions to known antibody epitopes in humans and mice. We also identified four novel sites consistent with humoral activity, which are not contained within any known antibody epitopes reported in the literature. These sites flank the V3 hypervariable loop, which is the most epitope dense region of Gp120 (LANL Immunology Database; https://www.hiv.lanl.gov/content/immunology/), although this may be due to a bias in reporting stemming from the extensive study of V3 in vaccine design rather than a genuine increase in immune activity. Whilst numerous antibodies against V3 have been described, the cross-neutralisation potential of these antibodies is generally low, reviewed by Hartley et al. Glycosylation, sequence variation, masking by V1-V2, and the specific amino acid make-up of the loop may contribute to this, reviewed by Pantophlet et al. However, some monoclonal and polyclonal antibodies specific to epitopes within V3 have been demonstrated to neutralise diverse HIV-1 strains in vitro. Two of the new sites exhibit particularly limited amino acid diversity and as such may warrant further investigation as potential components of vaccines targeting shared epitopes of very low diversity within V3.

Indeed, despite evidence for significant antibody-mediated selection pressure, some sites were relatively conserved in terms of their composition, containing only a limited number of biophysically similar amino acids. This is likely due to functional or structural constraints on the protein, and may reduce the ability of the virus to successfully escape antibodies targeting these regions. We also identified sites containing biophysically diverse amino acids that may be contained within epitopes eliciting highly effective antibody responses which cycle continuously between a limited number of biophysically distinct forms throughout chronic infection, as predicted by a previous modelling study. Consistent with this model, we found evidence within individuals that some sites contain major variants that appear and disappear over time, such as proline to leucine in position 369 and arginine to serine in position 444.
In summary, our data provide insight into how and where the surface of Gp120 is mutating over the course of clinically latent infection, and indicate that humoral immunity is the likely driving factor of such changes. Our detailed analysis of HIV-1 evolution and selection within hosts infected with a narrow-source virus allowed us to identify amino acid sites under positive selection that was likely attributed to host factors. Importantly, this comprehensive mapping resulted in the identification of both previously described and novel constrained antibody and T-cell epitopes. These sites are likely crucial to viral envelope function, and may aid in the development of future drugs and vaccines.
REFERENCES


Auguie, B. (2016).


Lesk, A. M. *Introduction to Bioinformatics.* (Oxford University Press, 2002).
ACKNOWLEDGEMENTS

We would like to thank the participants of the SM cohort who contributed samples towards this study. We would like to acknowledge the following funding bodies: SMA has received funding from the Wellcome Trust (099815). JE has received funding from the Swedish Research Council (350-2012-6628, 2016-01417) and the Swedish Society for Medical Research (SA-2016). SG has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 268904 – DIVERSITY. YZ’s work has been supported by the National Natural Science Foundation of China (81271842, 81320108017), Beijing Natural Science Foundation (7132098), Beijing Municipal Administration of Hospitals (XMLX201411), Beijing Key Laboratory (BZ0373), and Capital Health Development (TG-2015-19). Initial sample collection was funded by the Li Ka Shing Foundation.

AUTHOR CONTRIBUTIONS

S.M.A performed the experiments. S.M.A and J.E analysed the data. S.M.A, J.E, S.R-J, T.D, and S.G. contributed substantially to the conception and design of the study. Y.Z. sourced the samples used in the study. S.M.A wrote the manuscript, S.M.A, J.E, S.G. and S.R-J edited the manuscript, and all authors gave approval of the manuscript for submission.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing interests.

DECLARATIONS

None.
ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethical approval was obtained for this study from Beijing You'an Hospital and the University of Oxford Tropical Ethics Committee (OxTREC).

CONSENT FOR PUBLICATION

Not applicable

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and analysed during the current study are available in the Genbank repository (accession numbers: MF078678-MF079252). Custom R scripts used in the analysis of these data are available from the authors on request.
1 FIGURES

2 Figure 1

[Image of Figures 1A]
Figure 1. Selection pressure within the HIV-1 envelope Gp120 protein. A) Mean dN/dS ratios for each codon within each patient. A dN/dS estimate greater than 1 indicates positive selection whilst an estimate of less than 1 indicates negative selection. Sites under significant selection shown in blue whilst sites that have not reached statistical significance are shown in red and are assumed neutral. Differences in the number of codons across patients are the result of length variation in the V4 and V5 hypervariable loops. B) The proportion (absolute) of patients showing evidence of significant positive or negative selection across each aligned codon of the amplicon. Positive selection is shown in dark blue, whilst negative selection is shown in light blue. Sites under significant positive selection in at least 50% of patients are shown in purple. Variable loops V3, V4 and V5 are contained within the three boxes, respectively. The dotted lines denote 50% of patients. Antibody epitope clustering is shown in grey, whereby intensity denotes number of epitopes spanning that residue as reported in the
LANL Immunology Database (http://www.hiv.lanl.gov/content/immunology). Sequences have been aligned to the HXB2 Gp120 reference sequence (accession number K03455), and position is relative to this alignment. C) Homology-modelled structure of the SM cohort consensus Gp120 sequence in surface representation. Variable loops V3, V4 and V5 are shown in grey and sites under significant positive selection in 50% or more patients are shown in purple. Structure has been modelled on a glycosylated HIV-1 Gp120 trimer (RCSB PDB 3J5M)⁴⁰. For clarity, two molecules in the trimer are shown in line representation in grey.
Figure 2

A

[Bar chart showing frequency distribution of unique variants]
Figure 2. Variants mapping. A) Unique variants identified in the SM cohort envelope Gp120 sequences. Major variants present in greater than 15% of the sequences are shown in purple whilst minor variants (<15%) are shown in grey. Position is relative to HXB2 Gp120 (accession number K03455). B) Presence of major variants in individual patients of the SM cohort at each
time point sampled. Those shown in dark purple are also under significant positive selection
pressure within the specified individual, whilst those in light purple are either under significant
negative selection or under no significant selection pressure. C) Homology modelled structure
of the SM cohort consensus Gp120 sequence, showing the variants under significant positive
selection in 50% or more of the patients in purple, and all invariant sites in orange. All variants
except positions 345 and 424 are visible on the surface of the protein. Variable loops V3, V4
and V5 are shown in grey. Structure has been modelled on a glycosylated HIV-1 Gp120 trimer
(RCSB PDB 3J5M)\(^{40}\). For clarity, two molecules in the trimer are shown in line representation
in grey.
Figure 3. Biophysical properties of amino acids found in sites housing major variants.

The biophysical properties of the amino acids found in each site, coloured according to the Lesk format\textsuperscript{65}. The table shows the composition of each position as the total number of each amino acid found at that site.
Table 1. Antibody epitope sequences corresponding to the sites of interest identified in this study. Data made available by the Los Alamos National Laboratory (LANL) Immunology Database (http://www.hiv.lanl.gov/content/immunology). Position is relative to HXB2 Gp120 (accession number K03455).

<table>
<thead>
<tr>
<th>Variant</th>
<th>Position</th>
<th>Antibody</th>
<th>Sequence</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>F223Y</td>
<td>223</td>
<td>GV4H3</td>
<td>APAGFAIL</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>493-156</td>
<td></td>
<td>EPIPIHYCAPAGFAILKCNN</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>J1</td>
<td></td>
<td>GFAILKCNNK</td>
<td>mouse</td>
</tr>
<tr>
<td>T236S</td>
<td>236</td>
<td></td>
<td>KGSDKNVSTV</td>
<td>human</td>
</tr>
<tr>
<td>E268K</td>
<td>268</td>
<td>B12</td>
<td>RPVVSTQLLLNGLAEVEVV</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>110.E</td>
<td></td>
<td>NGSLAEEEVIRSVNFTDNA</td>
<td>mouse</td>
</tr>
<tr>
<td>V270I</td>
<td>270</td>
<td>B12</td>
<td>RPVVSTQLLLNGLAEVEVV</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>110.E</td>
<td></td>
<td>NGSLAEEEVIRSVNFTDNA</td>
<td>mouse</td>
</tr>
<tr>
<td>T278S</td>
<td>278</td>
<td>110.E</td>
<td>NGSLAEEEVIRSVNFTDNA</td>
<td>mouse</td>
</tr>
<tr>
<td>D279N</td>
<td>279</td>
<td>110.E</td>
<td>NGSLAEEEVIRSVNFTDNA</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>110.C</td>
<td></td>
<td>VIRSVNFTDN</td>
<td>mouse</td>
</tr>
<tr>
<td>V283I*</td>
<td>283</td>
<td>NA#</td>
<td>NA#</td>
<td>NA#</td>
</tr>
<tr>
<td>E290K*</td>
<td>290</td>
<td>NA#</td>
<td>NA#</td>
<td>NA#</td>
</tr>
<tr>
<td>E293G</td>
<td>293</td>
<td>IIB-V3-26</td>
<td>SVEINCTRPNNTRKS</td>
<td>mouse</td>
</tr>
<tr>
<td>L333I*</td>
<td>333</td>
<td>NA#</td>
<td>NA#</td>
<td>NA#</td>
</tr>
<tr>
<td>S334N*</td>
<td>334</td>
<td>NA#</td>
<td>NA#</td>
<td>NA#</td>
</tr>
<tr>
<td>R335S/K</td>
<td>335</td>
<td>P1H6</td>
<td>SSNWKE</td>
<td>mouse</td>
</tr>
<tr>
<td>T336A</td>
<td>336</td>
<td>P1H6</td>
<td>SSNWKE</td>
<td>mouse</td>
</tr>
<tr>
<td>N340K</td>
<td>340</td>
<td>P1H6</td>
<td>SSNWKE</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>P3B2</td>
<td></td>
<td>WKEM(D/N)R</td>
<td>mouse</td>
</tr>
<tr>
<td>Q344R</td>
<td>344</td>
<td>P3B2</td>
<td>WKEM(D/N)R</td>
<td>mouse</td>
</tr>
<tr>
<td></td>
<td>838-D</td>
<td></td>
<td>KSITK</td>
<td>human</td>
</tr>
<tr>
<td></td>
<td>1006-15D</td>
<td></td>
<td>KSITKG</td>
<td>human</td>
</tr>
<tr>
<td></td>
<td>1027-15D</td>
<td></td>
<td>KSITKGP</td>
<td>human</td>
</tr>
<tr>
<td>Variant</td>
<td>Position</td>
<td>Sequence</td>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>--------</td>
<td></td>
</tr>
<tr>
<td>I345V</td>
<td>345</td>
<td>838-D</td>
<td>KSITK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1006-15D</td>
<td>KSITKG</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1027-15D</td>
<td>KSITKGP</td>
<td></td>
</tr>
<tr>
<td>P369L</td>
<td>369</td>
<td>4D7/4</td>
<td>IFKQSSGGDPEIVTHSFNCGG</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.1(ARP 329)</td>
<td>FKQSSGGDPEIVTHSFNCGGGE</td>
<td></td>
</tr>
<tr>
<td>V424I</td>
<td>424</td>
<td>2D3</td>
<td>RIKQIINMWQEVGKAMYAPPI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JL413</td>
<td>KQIINMWQEVGKAMYA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5C2E5</td>
<td>QFINMWQEVK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>G3-211</td>
<td>IINMWQKVGGKAMYAP</td>
<td></td>
</tr>
<tr>
<td>R444S</td>
<td>444</td>
<td>polyclonal</td>
<td>KAMYAPISGQIRCSSNITG</td>
<td></td>
</tr>
<tr>
<td>N448K</td>
<td>448</td>
<td>polyclonal</td>
<td>KAMYAPISGQIRCSSNITG</td>
<td></td>
</tr>
<tr>
<td>D474N</td>
<td>474</td>
<td>polyclonal</td>
<td>LTRDGNNNNESEIFRPGG</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9201</td>
<td>GGGDMRDNWRSE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1C1</td>
<td>GGGDMRDNWRSELYKYKVK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H11</td>
<td>GGGDMRDNWRSELYKYKVK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>W2</td>
<td>GGGDMRDNWRSELYKYKVK</td>
<td></td>
</tr>
<tr>
<td>D476K</td>
<td>476</td>
<td>9201</td>
<td>GGGDMRDNWRSE</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1C1</td>
<td>GGGDMRDNWRSELYKYKVK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H11</td>
<td>GGGDMRDNWRSELYKYKVK</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>W2</td>
<td>GGGDMRDNWRSELYKYKVK</td>
<td></td>
</tr>
</tbody>
</table>

1 *Novel site identified in the current study.
2 #NA = Not available.