Analysis of HIV-1 envelope evolution suggests antibody-mediated selection of common epitopes among Chinese former plasma donors from a narrow-source outbreak

Sophie M. Andrews¹, Yonghong Zhang², Tao Dong³, Sarah L. Rowland-Jones¹, Sunetra Gupta⁴ & Joakim Esbjörnsson¹,⁵

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)¹. The Gp120 forms heterodimers with Gp41 which themselves trimerise, studding the viral membrane at a density of around fourteen copies per virion². 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 and development of HIV vaccines and antibody-based immunotherapy.

The humoral response against HIV-1 Gp120 develops rapidly within around four weeks of detectable plasma viral loads³, but neutralising antibodies (NAbs) typically only develop after several months of infection⁴. 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. 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. In addition, the loss of neutralising activity has been associated with faster disease progression in some individuals. 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 neutralisation by these responses: as one antibody is evaded, new antibodies arise and are evaded in a continuous cycle. This view is supported by the observation that HIV-1 is rarely susceptible to neutralisation by contemporaneous antibodies in early infection, whilst the same antibodies are able to effectively neutralise historic virus variants. However, in clinically latent infection, viral variants evolve susceptibility to neutralisation by contemporaneous NAbS, or to sera sampled much earlier in infection. 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.

Indeed, several apparent paradoxes in HIV-1 pathogenesis and the genetics of host susceptibility can be resolved by assuming that NAbS play an important role in the control of infection, as shown by a recent modelling study. 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 responses. A potential therapeutic approach to preventing disease progression may therefore be to develop vaccines that boost and maintain such partially cross-protective responses.

HIV-1 is one of the fastest evolving pathogens known to science due to extremely high mutation, recombination and replication rates. This leads to vast genetic diversity, and HIV-1 variants can differ genetically by >5% in an infected individual at a single time-point. The transmission of HIV-1 is associated with a major bottleneck, and in most cases, new infections are the result of the outgrowth of one single transmitted founder virus. During infection, HIV-1 continues to evolve at a high rate with diversification driven to a large extent by adaptive immune responses. Mutations that facilitate immune evasion are positively selected and becomes dominant in the viral population. Several studies have also shown a positive correlation between HIV-1 evolutionary rate and disease progression.

Here we studied gp120 evolution in a narrow-source cohort of former plasma donors (FPDs) from Henan Province in China (SM cohort). 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. Owing to the unusually homogeneous route and source of infection, and the narrow time frame during which study participants were exposed, the infecting founder was relatively conserved. To our knowledge, this is the first in-depth longitudinal study of HIV-1 envelope evolution in a population infected from a single source. 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 and identify potential antibody-restricted epitopes in natural infection, but also to understand the constraints placed on substitution in these regions, thereby testing whether positive selection exhibited fluctuating or consistent patterns of immune escape 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.

Results

Cohort characteristics. HIV-1 env gp120 sequences were recovered from 12 study participants in the SM cohort (Table 1). The HLA types of those were representative of cohort frequencies. Sequence recovery was 74% from the available specimens (Fig. S1A). PCR success was associated with the viral load of the sample. The final dataset consisted of 575 sequences. 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 study participants sampled are shown in Table 1.

Maximum likelihood phylogenetic analysis showed a star-like relationship between the sequences, with short internal branches between study participants, consistent with a narrow-source outbreak (Fig. S1B). Subtype analysis showed that all study participant 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 (Fig. S2). 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 study participants irrespective of HLA profile. Ten of the 12 study participants yielded sequences from two or more time-points and were subjected to evolutionary analyses (Fig. S1A). All time-points with available sequences for the ten study participants were included in the analyses. The median evolutionary rate ratio of positions 1 + 2 to position 3 among the participants was 0.861 (IQR: 0.779–0.892), indicating a general purifying selection over the HIV-1 env gp120 C2-V5 region in the majority of the study participants (the overall ratio was <1 in all study participants 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 study participants with longitudinal samples (Fig. 1), and neutral evolution was
Within the variable loops, substantial negative selection could be seen in V3, but not in V4 and V5 (Fig. 2A). 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 study participants irrespec-
tive of HLA profile (Fig. 1 and Table 1). 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. 2B).

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. 3). 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 vari-
ants were therefore resolved, and to prevent overrepresentation by particular study participants, a single time-point was selected for each study participant (the selection aimed to get an equal distribution of samples from each year of cohort follow-up). Thirty-one major variants were detected in total, across 29 sites (Fig. 3).

In study participants 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 study participant was also determined (Fig. 4A). Twenty-four of the 29 sites housing major variants were under significant positive selection pressure in at least one study participant, 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. 4B). 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 study participant 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 under significant positive selection pressure and visible on the surface of the protein (Fig. 5). 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 under significant positive selection pressure and visible on the surface of the protein were also cross-checked for existing antibody epitopes in the LANL Immunology Database and Genome Browser (Table 2). Twenty of these 22 sites had

| Patient | Sex | A1 | A2 | B1 | B2 | C1 | C2 | VL | CD4 abs | CD4% | ART | VL | CD4 abs | CD4% | ART | VL | CD4 abs | CD4% | ART |
|---------|-----|----|----|----|----|----|----|----|--------|-------|----|----|--------|-------|----|----|--------|-------|----|
| SM007   | F   | 1  | 2  | 40 | 57 | 6  | 8  | 7388| 326    | 23   | N  | 1445   | 300   | 24 | N  | 2884   | 253   | 16 | N  |
| SM021   | F   | 2  | 29 | 44 | 38 | 4  | 7  | 35827| 266    | 18   | N  | N/R   | N/R   | N/R | —  | 14782  | 206   | 20 | N  |
| SM039   | F   | 2  | 11 | 49 | 39 | 7  | 7  | 1612 | 492    | 30   | N  | 791   | 346   | 28 | N  | 1732   | 319   | 18 | Y  |
| SM176   | F   | 11 | 32 | 38 | 51 | 7  | 14 | 45569| 429    | 30   | N  | 43569 | N/R   | N/R | —  | 27348  | 95    | 3  | N  |
| SM209   | M   | 24 | 31 | 51 | 40 | 3  | 3  | 44584| 324    | 33   | N  | 1195  | 354   | 32 | N  | 4034   | 478   | 22 | N  |
| SM335   | M   | 30 | 31 | 13 | 40 | 6  | 3  | 7751 | 294    | 36   | N  | N/R   | 294   | 36 | N  | 71263  | 228   | 16 | Y  |
| SM358   | F   | 2  | 11 | 49 | 50 | 3  | 6  | 50   | 469    | 25   | N  | 1025  | 390   | 27 | N  | 1716   | 387   | 25 | N  |
| SM446   | M   | 2  | 26 | 38 | 51 | 12 | 14 | N/R  | 255    | 27   | N  | 7226  | 374   | 31 | N  | 3217   | 187   | 25 | N  |
| SM505   | M   | 11 | 2  | 35 | 52 | 12 | 12 | 8871 | 478    | 19   | N  | 17551 | 337   | 30 | N  | 210735 | 402   | 24 | N  |
| SM514   | F   | 2  | 31 | 40 | 40 | 3  | 15 | 30403| 389    | 11   | N  | N/R   | 1233  | 66 | N  | 5166   | 216   | 7  | N  |
| SM536   | F   | 11 | 30 | 13 | 15 | 6  | 7  | 4812 | 567    | 31   | N  | 23956 | 397   | 24 | N  | 12637  | 351   | 13 | N  |
| SM538   | M   | 2  | 24 | 13 | 40 | 3  | 3  | 409157| 92     | 10   | N  | 50    | 73    | 12 | N  | 50     | 52    | 9  | Y  |

Table 1. Demographic and clinical characteristics of the SM cohort study participants sampled. Sex: F = Female; M = Male. VL = Viral load in RNA copies per ml. CD4 abs = CD4 count in cells per µl. CD4% = CD4 cells as percentage of lymphocytes. ART = On antiretroviral treatment: Y = Yes; N = No. Data not recorded or available.
previously been associated with neutralising antibody activity, whereof 18 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. Two sites flanking the V3 hypervariable loop (E290 and L333) – which itself houses the majority of antibody epitopes – were identified against which no antibodies have yet been reported. The L333 site primarily switched between I/L.

Discussion

In line with previous observations, 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 study participant 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. It is also likely that V3 is subject to stronger functional constraints due to its important role in co-receptor binding. Moreover, it has been shown that deletion of V3 abrogates viral infectivity.

Several sites within each study participant showed evidence of significant positive selection, and five of these were common to at least half of the study participants 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, whilst more recent evidence suggests that their distribution may be random across the genome. 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 study participant, SM176. This position is contained within a known HLA-A11-restricted CTL epitope, which is one of the HLA alleles expressed by study participant SM176. It is therefore feasible that this variant has emerged in response to CTL-mediated selection pressure in this study participant. Conversely, position 424 is important in...
CD4 binding, and is contained within a known human antibody epitope. Mutation of this residue to methionine has been shown to increase susceptibility to neutralisation.

We were able to assign most positions to known antibody epitopes in humans and mice. We also identified two novel sites, which were not contained within any known antibody epitopes reported in the literature. Position 290 has, however, been associated with the CTL-restricted epitope AKTIIVQLTEPVE in the HIV-1 CRF02_AG lineage (https://www.hiv.lanl.gov/content/sequence/genome_browser/browser.html). Moreover, 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. Further characterisation by neutralisation experiments could help to confirm these novel surface-exposed epitopes as targets by humoral or cellular immune responses.

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. In addition, more complex causes of mutations, such as blockade of the accessibility of antibodies to the real epitopes or compensatory mutations to primary changes induced by antibodies, cannot be excluded. 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 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 could be 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 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.

Figure 2. Summary of the selection pressure within the HIV-1 envelope Gp120 protein among the study participants. (A) The proportion (absolute) of study participants 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 study participants are shown in purple. Variable loops V3, V4 and V5 are contained within the three boxes, respectively. The dotted lines denote 50% of study participants. 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. (B) Homology-modeled 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 study participants are shown in purple. Structure has been modelled on a glycosylated HIV-1 Gp120 trimer (RCSB PDB 3J5M). For clarity, two molecules in the trimer are shown in line representation in grey.
In summary, our data provide insight into how and where the surface of Gp120 is mutating over the course of clinically latent infection. Our detailed analysis of HIV-1 evolution and selection allowed us to identify amino acid sites under positive selection that was likely attributed to host factors (since the study participants were infected with a purportedly identical founder virus). In line with this, it is also possible that site-specific selection may be virus-specific. Importantly, this comprehensive mapping resulted in the identification of both previously described and novel constrained antibody and T-cell epitopes. These sites may be crucial to viral envelope function, and if verified in functional studies, provide suitable targets for future drugs and vaccines.

Methods

Cohort characteristics and sampling. The SM cohort comprises HIV-1 infected study participants from a small rural community in Henan province, China, as described previously. Between 1993 and 1995, the participants were infected with a narrow-source HIV-1 subtype B virus during a blood plasma donation scheme. Few individuals knew that they were infected until HIV screening programmes were implemented in China in 2004. Study participants were then recruited to the cohort.

Study participants with samples collected longitudinally over the course of HIV-1 infection permit detailed evolutionary analysis and identification of sites under selection. Twelve study participants from the SM cohort...
with multiple samples collected between 2010 and 2014 were available for this study (Table 1). All study participants gave informed consent for their samples to be used for research purposes.

**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). The Env gp120 C2-V5 (approximately 799 bp, HXB2 [accession number: K03455] positions 6883–7681) was amplified by nested touchdown PCR (primers listed in Table S1). 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 S1), 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 (HXB2 gp120 positions 661–1455, accession number K03455). Sequences were multiple aligned using

---

**Figure 4.** Overall presence of major variants. (A) Presence of major variants in individual study participants 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. (B) Homology modelled structure of the SM cohort consensus Gp120 sequence, showing the variants under significant positive selection in 50% or more of the study participants 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). For clarity, two molecules in the trimer are shown in line representation in grey.
MUSCLE\(^{62}\), and then manually edited in MEGA v6.06\(^{63}\). Sequences were controlled for intra-patient clustering by maximum-likelihood phylogenetics.

**Inference of infecting founder strain.** To infer the founder HIV-1 strain of the infected study participants, a consensus sequence was generated for each study participant for each time-point, with an ambiguity threshold of 10%. One sequence was selected per study participant 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. Study participant 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 study participant did not resolve monophyletically.

Bayesian Markov Chain Monte Carlo phylogenetic inference - implemented through BEAST v1.8.2 (http://beast.bio.ed.ac.uk/beast/)\(^{64}\) - was used to estimate the time to the most recent common ancestor (tMRCA). Divergence time was estimated using the SRD06 model\(^{65}\), and an uncorrelated lognormal relaxed molecular clock\(^{66}\) 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 (http://tree.bio.ed.ac.uk/software/tracer/). Three runs were combined in LogCombiner v1.8.2 (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 (http://beast.bio.ed.ac.uk/figtree).

As the SM cohort study participants 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 study participant. 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).
| Variant | Position | MAb ID | Sequence of antibody epitope | Source | Neutralising Antibody position | Neutralising antibody context |
|---------|----------|--------|----------------------------|--------|------------------------------|------------------------------|
| F223Y  | 223      | GY4H3  | APAGFAIL                   | mouse  | N                            | This position has a strong co-variation with potency and structural proximity. This position is part of a common glycosylation site at position 234. |
|         | 493-156  |        | EPIPHYCAPAGFAILKCNN        | mouse  | N                            |                              |
|         | J1       |        | GFAILKCNNK                 | mouse  | N                            |                              |
| T236S  | 236      | 1006-30-D | KGCKKNVSTV            | human  | Y                            |                              |
| E268K  | 268      | B12    | RPVSTQLLLNGSLAEEEV        | mouse  | Y                            | This is an antibody b12 signature site in which E or S has been associated with sensitivity, and K or R has been associated with resistance. Other antibodies associated with this site include VRC01. |
|         | 110E     |        | NGSLAEEEVIRSVNFTDNA       | mouse  | Y                            |                              |
| V270I  | 270      | B12    | RPVSTQLLLNGSLAEEEV        | mouse  | N                            |                              |
|         | 110E     |        | NGSLAEEEVIRSVNFTDNA       | mouse  | N                            |                              |
| T278S  | 278      | 110E   | NGSLAEEEVIRSVNFTDNA       | mouse  | Y                            | Antibodies associated with this site include: 1R2530, 3BNC60, 3BNC117, 8ANC131, 8ANC195, N6, NIH45-46, VRC01, VRC03, and VRC-PG04. |
|         | 110C     |        | VIRSVNFTDN                | mouse  | Y                            |                              |
| D279N  | 279      | 110E   | NGSLAEEEVIRSVNFTDNA       | mouse  | Y                            | This site has been associated with the loss of transmitted-founder sequence, suggested to represent antibody-driven selection (Hraber et al.82). Antibodies associated with this site include: 1B2530, 3BNC60, 3BNC117, A16, CH103, CH235, N6, NIH45-46, VRC01, VRC03, VRC18, VRC27, and VRC-PG04. |
|         | 110C     |        | VIRSVNFTDN                | mouse  | Y                            |                              |
| V283I  | 283      | NA     | NA                         | NA     | Y                            | Antibodies associated with this site include: HJ16, N6, VRC01, VRC03, VRC13, VRC16, VRC27, and VRC-PG04. |
| E290K4 | 290      | NA     | NA                         | NA     | N                            |                              |
| E293G  | 293      | IIIB-V3-26 | SVEINCTRPNNTRKSI       | mouse  | N                            | Antibodies associated with this site include: 1B2530, 3BNC55, 3BNC117, A16, CH103, CH235, N6, NIH45-46, VRC01, VRC03, VRC18, VRC27, and VRC-PG04. |
| L331P  | 333      | NA     | NA                         | NA     | N                            | Antibodies associated with this site include: 1B2530, 3BNC55, 3BNC117, A16, CH103, CH235, N6, NIH45-46, VRC01, VRC03, VRC18, VRC27, and VRC-PG04. |
| S334N  | 334      | NA     | NA                         | NA     | Y                            | This site has been described as a supersite of vulnerability for antibody neutralisation83. The site has also been associated with the loss of transmitted-founder sequence, suggested to represent antibody-driven selection (Hraber et al.82). Antibodies associated with this site include: 2G12, PCDN33A, PCDN38A, and PCDN38B. |
| R335S/K| 335      | P1H6   | SSNWKE                     | mouse  | Y                            | Antibodies associated with this site include the PCDN antibodies. |
| T336A  | 336      | P1H6   | SSNWKE                     | mouse  | N                            |                              |
| N340K  | 340      | P1H6   | SSNWKE                     | mouse  | N                            |                              |
|         | P5B2     | WKEM(D/N)R | mouse                    | Y      | Site shown to be under significant selection following 3BNC117 immunotherapy. |
| Q344R  | 344      | P5B2   | WKEM(D/N)R                | mouse  |                              |                              |
|         | 838-D    | KSKIK  | human                     |        |                              |                              |
|         | 1006-15D | KSKIKG | human                     |        |                              |                              |
|         | 1027-15D | KSKIKGP| human                     |        |                              |                              |
| I345V  | 345      | 838-D  | KSKIK                      | human  | N                            |                              |
|         | 1006-15D | KSKIKG | human                     |        |                              |                              |
|         | 1027-15D | KSKIKGP| human                     |        |                              |                              |
| F369L  | 369      | 4D7/4  | IFKQSSGGDPEIVTHSFNCGG     | mouse  | Y                            | This is an antibody b12 signature site in which A or P has been associated with sensitivity, and I, L, or Q has been associated with resistance. Other antibodies associated with this site include: CH103, and IgG1b12. |
|         | 36.1(ARP | 529    | FKQSSGGDPEIVTHSFNCGG     | mouse  |                              |                              |
| V424I  | 424      | 2D3    | RIKQINMWQEVGKAMYAPP       | mouse  | N                            |                              |
|         | IL413    | KQINMWQEVGKAMYA           | human  |                              |                              |
|         | 5C2E5    | QFINMWQEVK              | mouse  |                              |                              |
|         | G1-211   | JINMWQEVKJMYA           | mouse  |                              |                              |
| R44S4  | 444      | polyclonal               | KAMYAPPISGGQIRCSSNITG    | mouse  | Y                            | The S444 has been associated with increased susceptibility to neutralisation by 10-996. |
| N448K  | 448      | polyclonal               | KAMYAPPISGGQIRCSSNITG    | mouse  | Y                            | Antibodies associated with this site include: 2G12, 3BC176, 3BC315, and PGT151-PGT158. |

Continued
Table 2. Antibody epitope sequences corresponding to the sites identified on the surface of the HIV-1 Envelope. 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). Additional information about the neutralising antibody positions and associations can be found with references in the LANL HIV Genome Browser (https://www.hiv.lanl.gov/content/sequence/Aidsseqbrowser/browser.html). *Name of monoclonal antibody or “polyclonal” (if a general response is being studied) as listed in the LANL Immunology database. Y = Yes; N = No. ‘NA = Not available. ‘Novel site identified in the current study.

| Variant | Position | MAb ID* | Sequence of antibody epitope | Source | Neutralising Antibody position* | Neutralising antibody context |
|---------|----------|---------|-------------------------------|--------|---------------------------------|------------------------------|
| D474N   | 474      | polyclonal | LTRGLGGN睢NLEIFRPGGGD | human | Antibodies associated with this site include: 12A21, 8ANC131, 8ANC134, HJ16, N6, NIH45–46, VRC01, VRC03, VRC16, VRC27, and VRC-PG04. |
| 9201    |          |         | GGGDMRDNWRSY | mouse |                                |                              |
| 1C1     |          |         | GGGDMRDNWRLYKVVK | mouse | Y                              |                              |
| H1I     |          |         | GGGDMRD | mouse |                                |                              |
| W2      |          |         | GGGDMRDNNWRSYKVKY | mouse |                                |                              |
| D476K   | 476      |         |                                |        |                                 |                              |
| 9201    |          |         | GGGDMRDNWRSY | mouse | Y                              | Antibodies associated with this site include: 8ANC131, 8ANC134, A16, N6, NIH45–46, VRC01, VRC27, and VRC-PG04. |
| 1C1     |          |         | GGGDMRDNWRLYKVVK | mouse |                                |                              |
| H1I     |          |         | GGGDMRD | mouse |                                |                              |
| W2      |          |         | GGGDMRDNNWRSYKVKY | mouse |                                |                              |

Selection pressure in env gp120. The ratio of non-synonymous (dN) to synonymous (dS) mutations was estimated for each codon in study participant-specific alignments. Renaissance counting\(^{61,67}\) was implemented through BEAST v1.8.2, and the HK85 nucleotide substitution model\(^{68}\), 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 study participants was then constructed from these data, and the dN/dS estimates were combined across each aligned position. The proportion of study participants with virus showing evidence of significant selection pressure in each cohort was calculated.

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 genome\(^{69}\).

Structural modelling. Homology modelling was implemented through SWISS-MODEL\(^{70–73}\) 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)\(^{74}\). The sites of interest were annotated on the modelled structure in PyMOL v1.7.4 (Schrödinger, LLC.).

Data analysis. Epitope mapping, selection mapping, variant characterisation, biophysical properties, statistical analysis, and plotting were all performed in R (v3.2.3)\(^{75}\) via the RStudio (v.0.98.1103) integrated development environment (http://www.rstudio.com/). The following libraries were used: dplyr\(^{76}\); scales\(^{77}\); gridExtra\(^{78}\); ggplot2\(^{79}\); reshape2\(^{80}\).

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).

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.

References
1. Wyatt, R. et al. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**, 705–711 (1998).
2. Zhu, P. et al. Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* **441**, 847–852 (2006).
3. Tomaras, G. D. et al. Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *J. Virol.* **82**, 12449–12463 (2008).
4. Legrand, E. et al. Course of specific T lymphocyte cytotoxicity, plasma and cellular viral loads, and neutralizing antibody titers in 17 recently seroconverted HIV type 1-infected patients. *AIDS Res. Hum. Retroviruses* **13**, 1383–1394 (1997).
5. Schmitz, J. E. et al. Effect of humoral immune responses on controlling viremia during primary infection of rhesus monkeys with simian immunodeficiency virus. *J. Virol.* **77**, 2165–2173 (2003).
6. Miller, C. J. et al. Antiviral antibodies are necessary for control of simian immunodeficiency virus replication. *J. Virol.* **81**, 5024–5035 (2007).
A longitudinal assessment of autologous neutralizing antibodies in children in vivo following recent HIV infection.
46. Shioda, T., Levy, J. A. & Cheng-Mayer, C. Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1. Proc. Natl. Acad. Sci. USA 89, 9434–9438 (1992).
47. Cardozo, T. et al. Structural basis for coreceptor selectivity by the HIV type 1 V3 loop. AIDS Res. Hum. Retroviruses 23, 415–426 (2007).
48. Fenyo, E. M., Esbjörnsson, J., Medstrand, P. & Jansson, M. Human immunodeficiency virus type 1 biological variation and coreceptor use: from concept to clinical significance. Journal of internal medicine 270, 520–531, https://doi.org/10.1111/j.1365-2976.2011.02455.x (2011).
49. Cao, J. et al. Replication and neutralization of human immunodeficiency virus type 1 lacking the V1 and V2 variable loops of the gp120 envelope glycoprotein. J. Virol. 71, 9808–9812 (1997).
50. Lucchiar-Hartz, M. et al. Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic T lymphocyte epitope clustering in HIV-1. Proc. Natl. Acad. Sci. USA 100, 7755–7760 (2003).
51. Schmid, B. V., Keşmir, C. & de Boer, R. J. The distribution of CTL epitopes in HIV-1 appears to be random, and similar to that of other pathogens. BMC Evol. Biol. 9, 1–15 (2009).
52. Joos, B. et al. Positive in vivo selection of the HIV-1 envelope protein gp120 occurs at surface-exposed regions. The Journal of infectious diseases 196, 313–320, https://doi.org/10.1086/518935 (2007).
53. Kwong, P. D. et al. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393, 648–659 (1998).
54. Kiefer, F. et al. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-Model. Trends Biochem. Sci. 36, 833–841 (2011).
55. Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 26, 801–814 (2009).
56. Snoeck, J., Fellay, J., Bartha, I., Douek, D. C. & Telenti, A. Mapping of positive selection sites in the HIV-1 genome in the context of sample collection was funded by the Li Ka Shing Foundation.
57. Auguie, B. & Antonov A. gridExtra: Miscellaneous Functions for “Grid” Graphics https://CRAN.R-project.org/package=gridExtra (2016).
58. Wickham, H. Reshaping Data with the reshape Package. Journal of Statistical Software 21, 1–20 (2007).
59. Lesk, A. M. Introduction to Bioinformatics. (Oxford University Press, 2002).
60. Ho, C. E. et al. Anti-V3 monoclonal antibodies display broad neutralizing activities against multiple HIV-1 subtypes. PLoS One 5, e10254 (2010).
61. Kearse, M. et al. Automated sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
62. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Mol. Biol. Evol. 21, 1602–1607 (2004).
63. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729 (2013).
64. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214 (2007).
65. Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23, 7–9 (2006).
66. Drummond, A. J., Ho, S., Phillips, M. & Rambaut, A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88 (2006).
67. O’Brien, J. D., Minin, V. N. & Suchard, M. A. Learning to count: robust estimates for labeled distances between molecular sequences. Mol. Biol. Evol. 26, 1–15 (2009).
68. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
69. Schmid, B. V., Keşmir, C. & de Boer, R. J. The distribution of CTL epitopes in HIV-1 appears to be random, and similar to that of other pathogens. BMC Evol. Biol. 9, 1–15 (2009).
70. Joos, B. et al. Positive in vivo selection of the HIV-1 envelope protein gp120 occurs at surface-exposed regions. The Journal of infectious diseases 196, 313–320, https://doi.org/10.1086/518935 (2007).
71. Kwong, P. D. et al. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393, 648–659 (1998).
72. Kiefer, F. et al. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-Model. Trends Biochem. Sci. 36, 833–841 (2011).
73. Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 26, 801–814 (2009).
74. Snoeck, J., Fellay, J., Bartha, I., Douek, D. C. & Telenti, A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology 8, 1–8 (2011).
75. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22, 195–201 (2006).
76. Auguie, B. & Antonov A. gridExtra: Miscellaneous Functions for “Grid” Graphics https://CRAN.R-project.org/package=gridExtra (2016).
77. Wickham, H. scales: Scale Functions for Visualisation. https://CRAN.R-project.org/package=scales (2016).
78. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York 2009).
79. Wickham, H. Reshaping Data with the reshape Package. Journal of Statistical Software 21, 1–20 (2007).
80. Lesk, A. M. Introduction to Bioinformatics. (Oxford University Press, 2002).
81. Auguie, B. & Antonov A. gridExtra: Miscellaneous Functions for “Grid” Graphics https://CRAN.R-project.org/package=gridExtra (2016).
82. Auguie, B. & Antonov A. gridExtra: Miscellaneous Functions for “Grid” Graphics https://CRAN.R-project.org/package=gridExtra (2016).
83. Kong, L. et al. Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. Nature structural & molecular biology 20, 796–803, https://doi.org/10.1038/nomb.2014.200 (2013).

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.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-23913-2.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2018

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018