## SUPPLEMENT

## Extended Discussion

**Convergence in the Indel mutations in the hypervariable V1, V4 and V5 loops of CH505 and CH848 infected humans and rhesus macaques.** In functional sequences, insertions and deletions (Indels) occur in multiples of three nucleotides so as to retain the integrity of the Env open reading frame. Insertions in the hypervariable regions primarily occur by duplications, which result in perfect or imperfect direct repeats (*47*). In the CH505 infected human and rhesus animals, there was a striking recurrence of precise Indel events (**Fig. S9**). In V1, for example, *env* sequences from the human and all six animals contained an identical 3 nucleotide deletion. The human and three RMs contained an identical 12 nucleotide perfect direct repeat insertion. Five additional distinct insertions of between 6 and 24 nucleotides were shared among between the human sequences and a subset of rhesus macaque sequences. The most striking set of shared Indel patterns overall was observed in subject CH505 and RM6069, where the two shared five unique insertions and one deletion. The other five animals shared between two and three Indels identically with the human subject (**Fig. S9**). To estimate the probability that these shared Indels could have occurred by chance, we subjected sequences from the human subject CH505 and from three animals (RMs 6069, 6070, 6072) to rigorous statistical analysis (see below for statistical methods). First, we considered the natural distribution of V1 loop lengths found in the LANL database (www.hiv.lanl.gov) and used a comparable window in time from the longitudinal sampling (~1 year from infection) to study the evolution of the hypervariable loops in subject CH505 and the RMs. We estimated the probability of 5 precisely shared Indels in both the human CH505 and RM6069 at $p<3.3x10^{-9}$ by chance alone, and the probability of all three monkeys sharing as many precise Indels with CH505 as $p<3.0x10^{-13}$. Another approach to estimating the likelihood of seeing identical Indels is to focus on the 3 base deletion in V1 shaded in red (**Fig. S9**) that was found in the human CH505 and in all 6 RMs infected with SHIV CH505. The probability of an identical deletion occurring in all 7 hosts by chance was estimated to be $p<6.2 \times 10^{-7}$. Seven distinct insertions of 3 to 9 nucleotides were identified in the hypervariable region of V5 in CH505 infected human or macaque sequences. Each of these insertions represented perfect direct repeats. One of these Indels was shared between the human and RM6069, and five others were shared among different animals. Two direct repeats were shared among three animals. The likelihood of this happening by chance was estimated to be $p < 1.5 \times 10^{-5}$. Additional Indels in the V4 region of CH505 Envs were shared between human and rhesus (**Fig. S9**), and still other Indels in V1, V4 and V5 were shared between the CH848 infected human and rhesus animals (**Fig. S10**). The statistical likelihood that these reiterations occurred by chance is vanishingly small. Instead, the repetition of Indels in human and rhesus variable loop sequences attests to the fitness advantage associated with specific shared patterns of convergent or parallel evolution.

*Statistical methods to analyze repeated Indels.* We first collected all viable V1 loops, Cys to Cys, from the LANL database (www.hiv.lanl.gov) of 2546 full HIV-1 genome DNA alignments to get a baseline assessment of what is possible and common for the virus in terms of V1 loop lengths. We assumed that this distribution reflects how readily HIV-1 tolerates variation in V1 length and how likely V1 is to take on a particular size. We used this distribution as an *a priori* baseline rather than assuming all sizes are equally probable. We also assumed that the strand switching that gives rise to the Indels is not impacted so much by length of the Indel but rather by what sequences are observed and by how well the length of the V1 is tolerated after the primary Indel-mutation event. CH505 T/F Env has a relatively short V1 length of 69 nucleotides. One of the V1 insertions was a perfect direct repeat of 15 nucleotides (AATGCTACTGCCAGCAATGCTACTGCCAGC), which gave rise to a V1 length of 69+15=84 nucleotides. In the HIV-1 database, the frequency of 84 base V1 segments is 195/2546=0.0766, so we made the assumption that random Indels would create a length of 84 bases 0.077% of the time. The hypervariable part of V1 is 51 bases long (this is the stretch within V1 that is unalignable in the database) and we assume an Indel that retains the correct reading frame can happen anywhere in this 51 base long stretch, with the probability of this particular length insertion at the exact location estimated at 0.0766/51=0.00150. We then derive:

$$P_{matched} = \binom{S}{M} * \sum_{i,j,k,\ldots,m} M! * P_i * \; P_j * \cdots * P_m \leq \binom{S}{M} * \binom{H}{M} * \mathrm{M}! * P_1^* * P_2^* \; * \; \cdots * P_M^*$$

where
$H$ = number of human Indels
$M$ = number of matched Indels found in monkeys and human
$S$ = number of monkey Indels
$P_i$ = probability of seeing Indel *i (length*location)*
Let $P_1^*, P_2^*, \ldots, P_M^*$ be the M largest values of $P_i$

There are many different ways (*H* choose *M*) of matching *M* out of *H* Indels, but the probability is different for each choice of *M*, and depends on the probabilities $P_i$ , $P_j$ , $P_k$ , $P_l$ , $P_m$ associated with the particular *i,j,k,l,m* in each choice. Since order doesn't matter, there is an extra factor of *M!* in the probability. The separate computation for each combination of M out of H would give the probability for that particular combination; we add those probabilities to get an overall probability of matching M out of H. To simplify this, we can place an upper-bound on the probability that a set of M Indels would exactly be matched in a monkey by recognizing that there are (H choose M) terms, and that all the terms are less than or equal to the maximum term. The S choose M prefactor takes into account that there were more Indels in the monkey than matched the human. This maximum is computed by taking the product of the M largest probabilities; in this study the upperbound was extremely small so we didn't take this further.

The probability of the number of 5 shared Indels between
CH505 and RM6069 happening by chance alone is $<3.3\text{x}10^{-9}$

10 choose 5 = 252, 5! = 120
10 chose 1 = 10
10 choose 2 = 45, 2! = 1
7 choose 5 = 21

$P_{RM6069}$ = 21*252*120*(0.0015*0.0015*0.0015*0.0013*0.0012) = $3.3\text{x}10^{-9}$
$P_{RM6070}$ = 3*10*0.0015 = 0.045
$P_{RM6072}$ = 1*45*2*(0.0015*0.0015) = 0.0002

The probability of all three monkeys having the shared Indels that were observed occurring by chance alone is $<3.0\text{x}10^{-13}$.

We can also ask how likely it would be for all four hosts (one human and three monkeys) to share the same Indel; we ask this without specifying which Indel is shared. The question is: what is the chance that any Indel is shared among all four? Let $P_1$, $P_2$, …, $P_K$ correspond to the K possible Indels. Let A1, A2, A3, A4 be the number of Indels observed in the four animals (for in our study, A1=10, A2=7, A3=3, A4=2); then $A_iP_k$ is the probability of observing the k'th Indel in the i'th animal. And $A1A2A3A4Pk^4$ is the probability of observing a specific Indel k in all animals. Thus we can write

$$A_1A_2A_3A_4 \sum_{k=1}^{K} P_k^4$$

as the probability of observing an unspecified Indel in all four animals. This probability is: 10*7*3*2*(1.47975 x $10^{-9}$) = 6.2 x $10^{-7}$,
as $\sum_{k=1}^{K} P_k^4 = 1.47975\mathrm{e}-09$.

The length of the hypervariable region of V5 in CH505 T/F is 30 bases, which is short compared to global group M sequences with a median of 39. Selective pressures in the human subject CH505 and in 4 of 6 monkeys drove V5 to get longer throughout the course of infection. Given that identical insertions showed up in multiple individuals, we asked how likely it would be for three individuals to share the same Indel. Again, we asked this without specifying which Indel is shared. Let $P_1$, $P_2$, …, $P_K$ correspond to the K possible Indels. Let A1, A2, A3 be the number of Indels observed in the 3 animals (for in our study, A1=5, A2=4, A3=5,); then $A_iP_k$ is the probability of observing the k'th Indel in the i'th animal. And A1A2A3Pk$^3$ is the probability of observing a specific Indel k in all animals. Thus we can write

$$A_1A_2A_3 \sum_{k=1}^{K} P_k^3$$

as the probability of observing an identical Indel in all three animals. This probability is 5*4*5*(3.89133e-05)= 0.0039 for a single event. The likelihood that two unique Indels would occur in three animals, as was the case for RMs 6072, 6069 and 5695, is 0.00389^2 = 1.5 x $10^{-5}$.

**Enhanced N160K responses.** In **Figs. 1**, **3** and **S12**, we showed evidence of V2 apex, C-strand targeted bNAb activity in the plasma of RMs 5695, 6070, 42056 and 40591. These data included broad neutralizing activity against multiple heterologous HIV-1 strains; loss of neutralizing activity when key residues 166 and 169 were mutated to eliminate positively charged arginines or lysines; and detection of bNAb escape mutations at these same 166 and 169 residues in vRNA sequences from serially collected plasma specimens. In animal RM5695, we isolated four bNAb mAbs (RHA1.V2.01-04) that exhibited these properties and accounted for most of the neutralization breadth observed in the animal's plasma. We then showed by Cryo-EM that the mAb RHA1.V2.01 binds to the trimer apex-hole residues with contacts to C-strand residues and N160 glycan of the BG505 DS-SOSIP trimer (**Figs. 6, S19,27**). A key feature of the RHA1.V2.01 mAb was its strict dependence on N160 for binding and neutralization. This is an expected property since nearly all reported human V2 apex targeted bNAbs (e.g., PG9, PGT145, CH01, PCT64) depend on interactions with N160 for their binding and neutralization activity. The exceptions are some members of the VRC26 lineage of V2 apex bNAbs that do not require N160 for their activity (*3, 49*). A surprising finding in our study was that the plasma from RMs 5695, 6070, 42056 and 40591 variably neutralized heterologous viruses strains with N160K substitutions, depending on the Env background (**Figs. 3**, **S13,S23**). This included animal RM5695 whose RHA1 broadly neutralizing mAbs were strictly N160 dependent. Plasma from RM5695 showed enhanced neutralization of Q23.17.N160K, no difference in neutralization of T250 wt compared with T250.N160K, and failed to neutralize BG505.N160K (**Fig. S13A**). Plasma from animals RM6070 (**Fig. S13B**) and RM42056 (**Fig. S13D**) exhibited plasma neutralization titers against heterologous viruses with N160K substitutions that were extraordinarily high ($ID_{50}$ >1:100,000). This potent neutralization was directed entirely to the V2 apex, since N160K.K169E double mutants eliminated or drastically reduced the N160K enhanced neutralization (**Figs S13B,D**). We considered two possible explanations for these findings: (i) In RMs 5695, 6070 and 42056, strain-specific 2909-like antibodies (*83, 84*) were elicited in addition to prototypical N160-dependent bNAbs; or (ii) C-strand targeted V2 apex bNAbs were elicited that do not require binding to N160 for their activity but instead are enhanced in potency when the protective shielding afforded by N160 is lost. Recently, we isolated rhesus V2 apex C-strand targeted broadly neutralizing mAbs from a SHIV.CH505 infected RM that was not part of the current study and found that it exhibits the latter features (R. Roark and G. Shaw, unpublished).