Infection pressure in apes has driven selection for CD4 alleles that resist lentivirus (HIV/SIV) infection

Cody J. Warren[1]†, Arturo Barbachano-Guerrero[1]†, Vanessa L. Bauer[1], Alex C. Stabell[1], Obaiah Dirasantha[1], Qing Yang[1], Sara L. Sawyer[1*]

[1]BioFrontiers Institute, Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado, USA.

†These authors contributed equally to this work

*Correspondence to: ssawyer@colorado.edu

Abstract

Simian immunodeficiency viruses (SIVs) comprise a large group of primate lentiviruses that endemically infect African monkeys. HIV-1 spilled over to humans from this viral reservoir, but the spillover did not occur directly from monkeys to humans. Instead, a key event was the introduction of SIVs into great apes, which then set the stage for infection of humans. Here, we investigate the role of the lentiviral entry receptor, CD4, in this key and fateful event in the history of SIV/HIV emergence. First, we reconstructed and tested ancient forms of CD4 at two important nodes in ape speciation, prior to the infection of chimpanzees and gorillas with these viruses. These ancestral CD4s fully supported entry of diverse SIV isolates related to the virus(es) that made this initial jump to apes. In stark contrast, modern chimpanzee and gorilla CD4s are more resistant to these viruses. To investigate how this resistance in CD4 was gained, we acquired CD4 sequences from 32 gorilla individuals of 2 species, and identified alleles that encode 8 unique CD4 proteins. Function testing of these identified allele-specific CD4 differences in susceptibility to virus entry. By engineering single point mutations from gorilla CD4 alleles into a permissive human CD4 receptor, we demonstrate that acquired SNPs in gorilla CD4 did convey resistance to virus entry. We provide a population genetic analysis to support the theory that selection is acting in favor of more and more resistant CD4 alleles in apes with endemic SIV infection (gorillas and chimpanzees), but not in other ape species (bonobo and orangutan) that lack SIV infections. Taken together, our results show that SIV has placed intense selective pressure on ape CD4, acting to drive the generation of SIV-resistant CD4 alleles in chimpanzees and gorillas.

INTRODUCTION

Simian immunodeficiency viruses (SIVs) comprise a large group of primate lentiviruses that infect African monkey species (Klatt et al., 2012; Sharp and Hahn, 2011). HIV-1 emerged into humans from this diverse viral reservoir, but was not a spillover of virus directly from monkeys to humans. Instead, a key transition was the spillover of SIVs into great apes, which then set the stage for infection of humans **(Figure 1)**. First, SIV of chimpanzees (SIVcpz) arose following the cross-species transmission and recombination of multiple SIVs from infected monkeys upon which chimpanzees predate (Bailes et al., 2003; Sharp et al., 2005). It is unknown if this virus recombination event occurred in the monkey reservoir before the first chimpanzee was infected, or if it occurred within chimpanzee populations. Subsequently, SIVcpz transmitted to gorillas (giving rise to SIVgor) (Heuverswyn et al., 2006; Takehisa et al., 2009). Chimpanzees and gorillas have been endemically infected with SIVcpz and SIVgor since those spillover events (Sharp and Hahn, 2011). Spillover to humans from both chimpanzees and gorillas subsequently occurred on multiple occasions (Heuverswyn et al., 2006; Keele et al., 2006; Plantier et al., 2009). One of these spillovers yielded HIV-1 "group M" – the pandemic virus that has swept the globe and which has infected over 80 million people. A third great ape species native to Africa, the bonobo, remains uninfected with SIV. Orangutans, the final great ape species, are native to Asia and also remain uninfected.

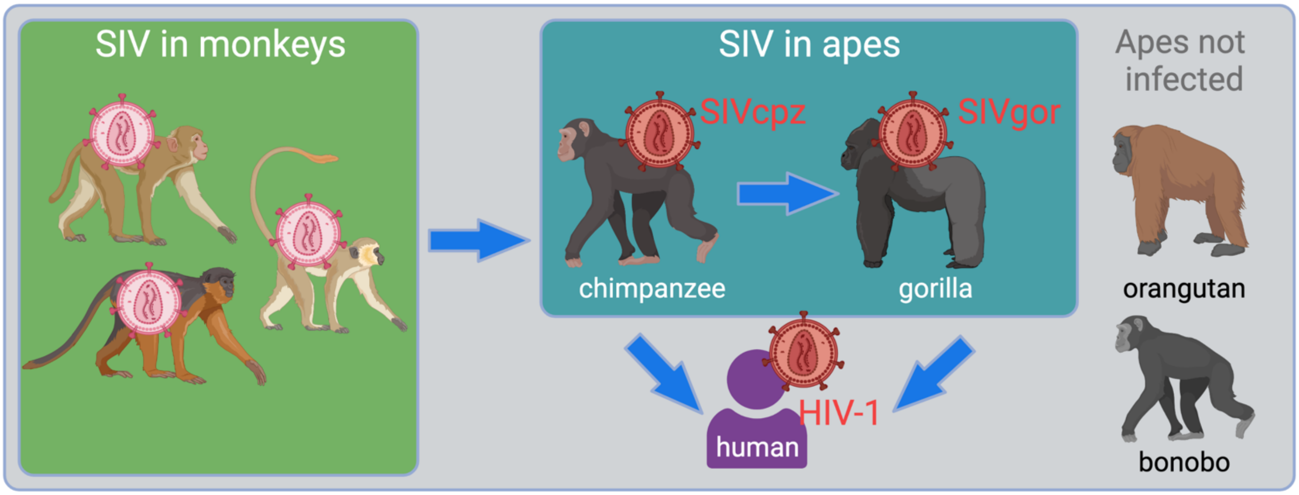


Figure 1. Overview of the emergence of SIV into apes, ultimately giving rise to several strains of HIV-1. The figure shows, in the green box, the SIV reservoir that exists in African monkeys. Chimpanzees became infected from this reservoir and new virus was created, SIVcpz (Bailes et al., 2003; Sharp et al., 2005). From there, chimpanzees infected both gorillas and humans. The final two great ape species, orangutans and bonobos, are not known to harbor any form of SIV.

CD4 is the main viral entry receptor for primate lentiviruses (SIV and HIV). CD4 is a surface protein expressed on T cells, where it is bound by the viral envelope (Env) glycoprotein to begin viral entry. To understand the role that CD4 plays in dictating the host tropism of SIVs, one must first appreciate the remarkable evolutionary signatures contained in the CD4 gene. CD4 has evolved under positive natural selection over the course of primate evolution (Meyerson et al., 2014; Zhang et al., 2008). This type of selection operates in favor of new alleles of CD4 that have better resistance to virus entry (Meyerson and Sawyer, 2011). As such, it has been noted that most of the sequence evolution in CD4 has been concentrated to its D1 domain, the region that contacts HIV and SIV (Meyerson et al., 2014; Zhang et al., 2008). Even though natural selection operates at the level of alleles and SNPs circulating within primate populations (Ohainle and Malik, 2021; Russell et al., 2021), the ultimate outcome is fixed CD4 sequence divergence between species. As would be expected because of the selection at play, we have demonstrated that different primate orthologs vary dramatically in the viruses that they will engage (Warren et al., 2019a). Thus, there has been intense selection on CD4 in host species that are endemically infected with SIV.

Here, we focus on a key event in the emergence of HIV into humans – the transition of SIVs from monkeys to apes. First, we reconstructed and tested ancestral forms of CD4 at two important nodes in ape speciation, prior to the infection of chimpanzees and gorillas with these viruses. These ancestral CD4s fully supported entry of diverse SIV isolates representing the virus(es) that made this initial jump to apes. In stark contrast, modern chimpanzee and gorilla CD4s are less supportive of infection by these viruses, consistent with natural selection having shaped CD4 to resist infection. Second, we investigated the subsequent spillover of SIV from chimpanzees to gorillas. We gathered CD4 sequences from 32 gorilla individuals of 2 species, and identified CD4 alleles that encode 8 unique CD4 proteins. We then identified allele-specific CD4 differences in susceptibility to SIVcpz entry (the virus that spilled over to gorillas). By engineering single point mutations from gorilla CD4 alleles into a permissive human CD4 receptor, we demonstrate that these SNPs in CD4 are responsible for resistance to virus entry in gorillas, and provide a population genetic analysis to support the theory the selection is acting in favor of more and more resistant CD4 alleles in gorillas. Taken together, our results show that SIV has placed intense selective pressure on ape CD4, acting to drive the generation of SIV-resistant CD4 alleles and orthologs.

RESULTS

Receptor mediated resistance to SIV entry is a trait acquired during ape speciation.

First, we wanted to know what ape CD4 was like before SIVs spilled over to apes and began to exert infection pressure on them. We used an alignment of CD4 from diverse simian primates, and the program PAML (Yang, 2007a), to infer ancestral CD4 sequences at the base of the hominin and hominid clades, at the evolutionary positions shown with red and blue nodes in **Fig. 2A**. These ancestral nodes yielded CD4 sequences that differ from human CD4 by only 2 (hominin) and 5 (hominid) nonsynonymous substitutions. Only one of these changes mapped to the D1 domain of CD4 (N52S; **Fig. 2B**). We then synthesized these ancient and extinct CD4 genes. We constructed stable cells lines where Cf2Th (canine) cells were transduced with retroviral vectors that stably integrated each of these CD4 genes (hominin ancestral CD4, hominoid ancestral CD4, human CD4, gorilla CD4, chimp CD4, or an empty vector). All cell lines were also engineered to express human CCR5, a critical co-receptor for SIV and HIV.

We then tested these extinct and modern CD4 proteins for their ability to support viral entry mediated by SIVcpz Env. Since we don't know the actual genetic sequence of the first SIV(s) to infect chimpanzees, the best alternate strategy is to test a phylogenetically diverse set of extant SIVcpz strains (**Fig. 2C**). We also tested HIV-1 strains that are embedded within the SIVcpz clade, because these represent the virus spillovers from chimpanzees to humans. To generate pseudoviruses bearing SIVcpz and HIV Env, different SIVcpz and HIV-1 Env expression plasmids were co-transfected into 293T cells along with a plasmid encoding HIV-1ΔEnv-eGFP. The cell lines stably expressing various CD4 proteins and human CCR5 were then infected with each of these pseudoviruses. The percent of GFP+ (infected) cells was measured by flow cytometry and viral titers on the different cell lines were calculated and reported as transducing units per milliliter (TDU/mL). All tested pseudoviruses displayed similar levels of infection on cells bearing human or the ancestral CD4 proteins (**Fig. 2D, 2E**). This suggests that the ancestral versions of CD4 in apes were susceptible to primate lentivirus entry, just as human CD4 is known to be today. On the other hand, cells bearing chimpanzee and gorilla CD4s were generally less permissive to virus entry (**Fig. 2D, 2E**). We conclude that CD4 was originally permissive to primate lentiviruses, but that selective pressures exerted by SIVs in the chimpanzee and gorilla lineages have led to the retention of mutations that confer resistance to primate lentivirus infection. This has not happened in humans where selective pressure by HIV-1 is too new.

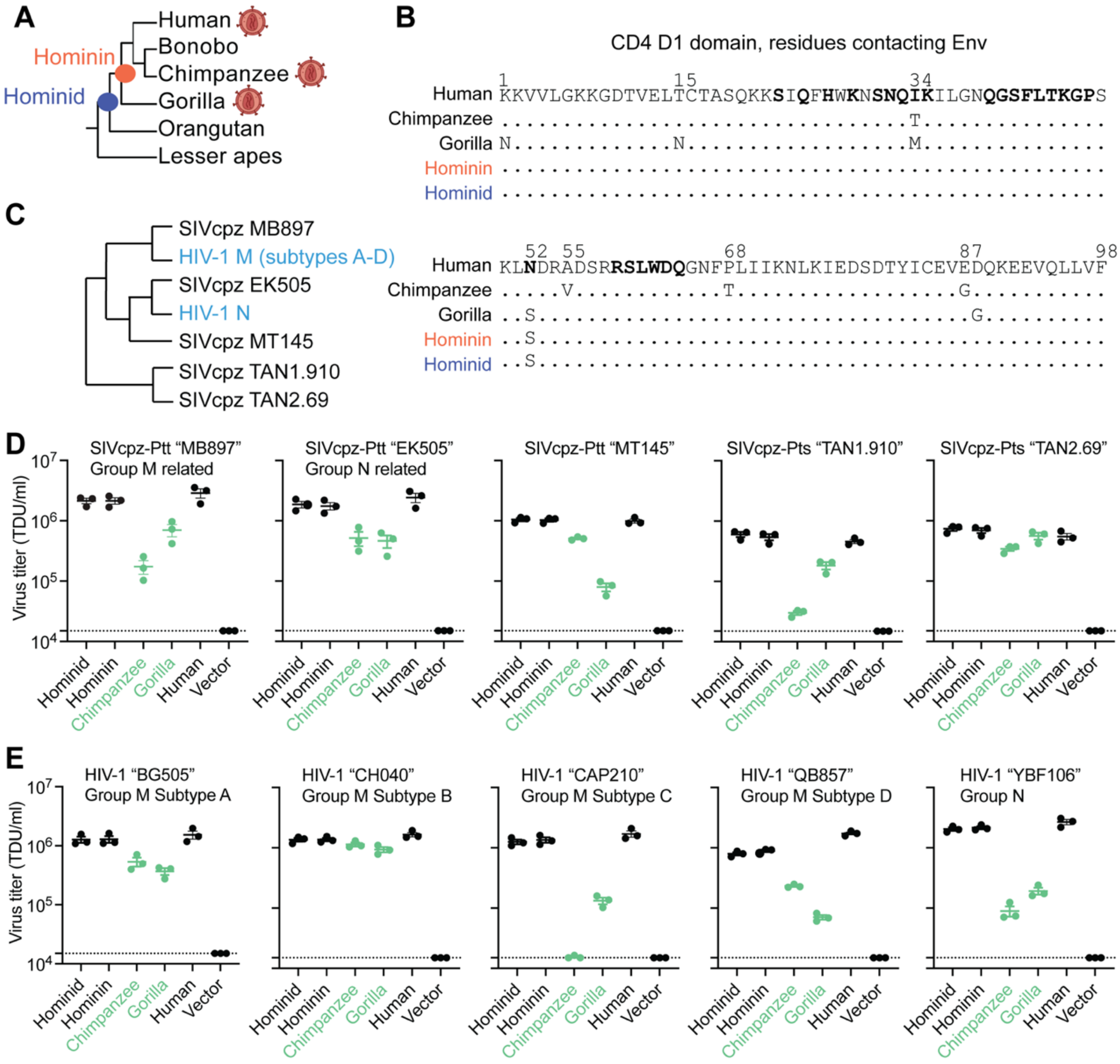


Figure 2. Receptor mediated resistance to SIVcpz entry is a trait acquired during ape speciation. (A) Cladogram of CD4 sequences among apes and the nodes for which ancestral sequences were reconstructed. The virion diagram next to some ape species represent apes that are infected by SIV/HIV. **(B)** An amino acid alignment of the CD4 D1 domain of human, chimpanzee, gorilla, and the inferred ancestral CD4 sequences. Dots represent identical residues compared to human and distinct amino acids and numerical positions are noted. Bolded residues on the human sequence represent sites known to directly interact with HIV-1 Envelope (Liu et al., 2017). **(C)** Cladogram of HIV-1 and SIVcpz was based on previously published work (Takehisa et al., 2007), highlighting genetic relationships of the

envelope (Env) clones used in this study. **(D, E)** HIV-1ΔEnv-GFP viruses were pseudotyped with Env's (top of graphs) from diverse **(D)** SIVcpz or **(E)** HIV-1 strains. Cf2Th cells stably expressing human CCR5 and various CD4s (X-axis) were infected with various volumes of these pseudoviruses and then analyzed by flow cytometry 48 hours post-infection. GFP positive cells were enumerated and virus titers (transducing units per milliliter; TDU/mL) were determined for those samples falling within the linear infection range (n = 2 titration points). The mean virus titers obtained from each of three independent experiments were plotted (dots), with error bars representing the standard error of the mean (SEM). Dotted lines represent the lower limit of detection for this assay. SIVcpz-Ptt and SIVcpz-Pts refer to SIVs derived from the chimpanzee subspecies *Pan troglodytes troglodytes* or *Pan troglodytes schweinfurthii,* respectively.

Gorilla CD4 alleles differentially support entry of SIVcpz.

Natural selection operates on individuals within populations, and only over time can the effects of this selection be seen in the divergence of gene orthologs between species. We and others have already shown that SIVcpz has placed selective pressures on chimpanzees such that multiple CD4 alleles circulate in chimpanzees which resist SIVcpz entry better than human CD4 (Bibollet-Ruche et al., 2019; Warren et al., 2019b). We next wanted to know if the same is true in gorilla populations. To similarly analyze gorilla CD4, we used the SNP data from the Great Ape Genome Project (Prado-Martinez et al., 2013) to identify extant CD4 alleles. We analyzed genetic data from 32 gorilla (*Gorilla gorilla gorilla* [n = 28]; *Gorilla gorilla diehli* [n = 1]; *Gorilla beringei graueri* [n = 3]) individuals and found six non-synonymous and five synonymous SNPs separating the individual alleles encoded. Five out of six of the non-synonymous polymorphisms are located within domain 1 of CD4 (two are in the same codon, codon 27) and one in domain 2 (**Fig. 3A**). A study of over 100 fecal samples from gorillas at field sites in Africa recently identified the same set of SNPs (Russell et al., 2021). Within the gorilla samples, these amino acid differences result in eight distinct allelic CD4 protein haplotypes. The allele frequencies of these protein haplotypes are heterogeneous, where allele 5 is the most common (**Fig. 3B**). (This, allele 5, was also the gorilla CD4 that was tested in **Figure 2D, E** and shown in the alignment in **Figure 2B**.) From looking at the sequences of these different alleles, we noticed a predicted glycosylation site (N-glycosylation tripeptide NXT) at position 15 that is fixed in the gorilla population but absent in the other African apes **(Fig. 3A)**. Interestingly, the gorilla CD4 allele 2 codes for a proline at position 18, immediately after the tripeptide NCT, which strongly reduces the likelihood of glycosylation (Gavel and Heijne, 1990). Since five of the six protein altering polymorphisms are located in domain 1 of gorilla CD4, which directly binds to the lentiviral Env (**Fig. 3C**) we next wanted to test their functional significance.

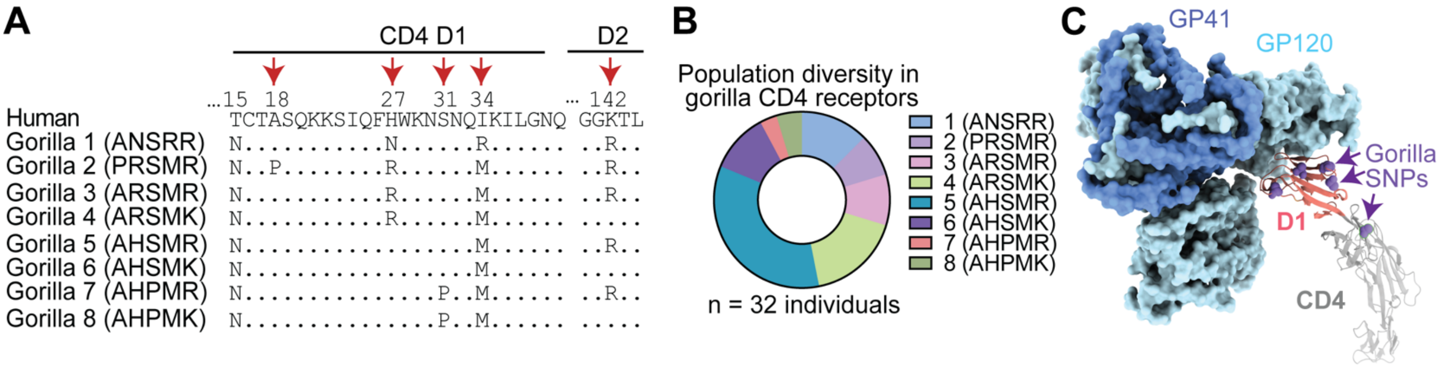

A
CD4 D1
D2
…15 18 27 31 34 … 142
Human TCTASQKKSIQFHWKNSNQIKILGNQ GGKTL
Gorilla 1 (ANSRR) N...........N......R...... ..R..
Gorilla 2 (PRSMR) N..P........R......M...... ..R..
Gorilla 3 (ARSMR) N...........R......M...... ..R..
Gorilla 4 (ARSMK) N...........R......M...... .....
Gorilla 5 (AHSMR) N..................M...... ..R..
Gorilla 6 (AHSMK) N..................M...... .....
Gorilla 7 (AHPMR) N...............P..M...... ..R..
Gorilla 8 (AHPMK) N...............P..M...... .....
B
Population diversity in gorilla CD4 receptors
1 (ANSRR)
2 (PRSMR)
3 (ARSMR)
4 (ARSMK)
5 (AHSMR)
6 (AHSMK)
7 (AHPMR)
8 (AHPMK)
n = 32 individuals
C
GP41
GP120
Gorilla SNPs
D1
CD4

Figure 3. Identification of diverse gorilla CD4 alleles. (A) Eight unique protein variants of gorilla CD4 were identified. The polymorphic sites (red arrows) are shown in the alignment, where dots indicate amino acid residues that are identical to human. **(B)** The frequencies of the eight CD4 protein haplotypes are shown for three gorilla subspecies, *Gorilla gorilla gorilla* (n = 28), *Gorilla beringei graueri* (n = 3), and *Gorilla gorilla diehli* (n = 1). **(C)** Cryo-EM structure of an HIV-1 Env trimer in complex with human CD4 (PDB 5U1F) was visualized in ChimeraX (Goddard et al., 2017). Individual gp120 and gp41 subunits are colored in light and dark blue, respectively. The CD4 D1 domain (red) and D2-D4 domains (gray) are shown, with gorilla SNPs shown on the human sequence as purple spheres.

We made stable cell lines expressing each gorilla CD4 allele, along with human CCR5 (**Fig. S1**). We then infected each of these with GFP pseudoviruses displaying envelopes from different strains of SIVcpz, as described above. We quantified the number of GFP+ cells to measure for the differential usage of each CD4 allele. Again, we don't know the exact strain of SIVcpz that initially infected gorillas, so instead we have tested a phylogenetic diversity of SIVcpz strains. We found substantial differences in susceptibility to pseudovirus entry between the alleles, varying by up to 2 orders of magnitude in some cases (**Figure 4**). All gorilla alleles were equal to, or more resistant to infection than, the human CD4. We also tested pseudoviruses displaying a diverse set of Envs from HIV-1 groups M and N, and found similar patterns (**Figure S2**). These data are consistent with SIV putting selective pressure on gorillas in favor of resistant alleles of CD4. However, as would be expected in a host-virus arms race, the viruses are evolving too. As such, we found considerable differences on the entry phenotype for each SIVcpz strain evaluated, where a single host CD4 allele can be highly restrictive to one strain, while being fully functional for entry of another. As an outlier, SIVcpz TAN2.69 showed a uniformly strong ability to use any of the gorilla CD4 alleles.

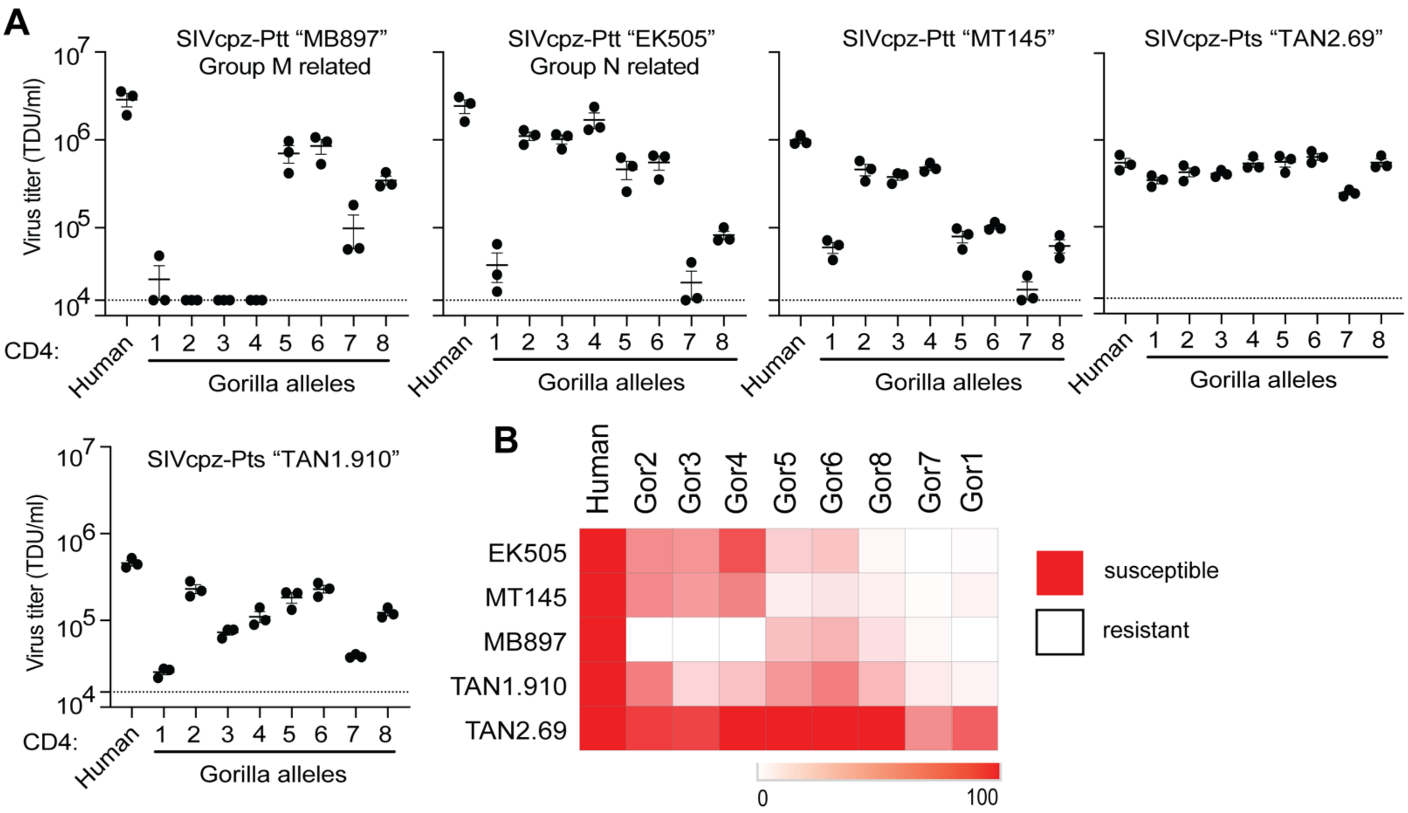


Figure 4. Gorilla CD4 alleles differentially support entry of SIVcpz. (A) HIV-1ΔEnv-GFP viruses were pseudotyped with Envs (top of graphs) from diverse SIVcpz strains. Cf2Th cells stably expressing human CCR5 and various CD4s (X-axis) were infected with various volumes of these pseudoviruses and then analyzed by flow cytometry 48 hours post infection. GFP positive cells were enumerated and virus titers (transducing units per milliliter; TDU/mL) were determined for those samples falling within the linear infection range (n = 2 titration points). The mean virus titers obtained from each of three independent experiments were plotted (dots), with error bars representing the standard error of the mean (SEM). **(B)** Data from each pseudotyped Env in A were used to calculate virus titer means normalized to human CD4 expressing cells and were plotted as a heat map, where red and white represent high susceptibility or resistance to viral entry, respectively. CD4 alleles and Envs were hierarchically clustered to depict similarities in phenotype.

Human CD4 engineered to encode gorilla SNPs supports less SIVcpz entry

We next sought to evaluate if CD4 polymorphisms found in gorilla individuals are protective when engineered in the human version of CD4, a widely susceptible receptor for primate lentiviruses. First, we investigated gorilla CD4 allele 2, which encodes a proline at position 18 that is predicted to prevent an otherwise fixed N-glycosylation at position 15 (**Fig. 3A**). We noticed that gorilla allele 2 CD4 is highly susceptible to most of the SIVcpz strains tested in this study (**Fig. 4**). Allele 3, which differs from allele 2 only by this proline, supported less entry by SIVcpz TAN1.910 presumably due to this change in glycosylation status (**Fig. 5A**). To explore the effects of this gorilla specific glycan at residue 15, we generated cell lines stably expressing a mutated version of human CD4 that encodes for the gorilla specific glycosylation motif. We then challenged these cells with pseudoviruses displaying the envelope of different SIVcpz strains and consistently found a decrease in susceptibility to entry compared to wild type human CD4 (**Fig. 5B**). To confirm the glycosylation status of CD4, we performed CD4 western blotting on lysates from cells stably expressing each of the different versions of CD4. As expected, human T15N CD4, as well as gorilla allele 3, migrated at a higher molecular weight compared to human wild type CD4 and gorilla allele 2, corresponding to the predicted number of glycans on CD4 domain 1 (**Fig. 5C**). We then treated the lysates with PNGase F, an N-linked glycosidase, and found that all CD4 versions migrated at the same rate, confirming that the mobility shift is due to the glycosylation status of CD4. Thus, most gorilla CD4 alleles have gained a glycan at position 15 that reduces entry of SIV as compared to human CD4. It seems that allele 2, which doesn't have this glycan, would be at a fitness disadvantage. In support of this, allele 2 is one of the least frequent alleles in the gorilla population that we surveyed **(Figure 3B)**.

We proceeded to test the amino acid residues at the other 3 polymorphic positions in domain 1 of gorilla CD4. We infected cells individually expressing mutant forms of human CD4 that coded for the gorilla specific residues at positions 27, 31, and 34. We found that in all cases, the mutant form of human CD4 encoding the gorilla specific amino acid was significantly more restrictive to at least two of the four SIVcpz strains when compared to human wild type CD4 (**Fig. 5D**). In several cases, such as H27R CD4 expressing cells infected with SIVcpz MB897, we found drastic changes rendering a fully supportive receptor now highly refractory to infection by a single amino acid substitution. We found that the protective role of these gorilla specific substitutions was SIVcpz strain specific, demonstrating that collectively, the diversity found in gorilla individuals can confer relative protection to all the SIV strains we tested, but that SIVs are counter-evolving as well. These results suggest that single amino acid changes in domain 1 can drastically modify the interaction between CD4 and the lentivirus envelope, directly influencing virus entry.

To evaluate the reverse - if gorilla CD4 mutated to recapitulate the amino acids encoded in human CD4 may render the CD4 a better receptor for SIVcpz - we made and constructed cells expressing those CD4s and quantified the level of infection. We did not observe a full gain-of-function phenotype here and instead found only minimal increases in entry of SIVcpz through these receptors (**Fig. 5E**). These results imply that the resistance to SIVcpz found in gorilla individuals is not dependent on single amino acids, but rather the cumulative effect of multiple SNPs. Overall, our data suggest that population-level diversity of CD4 in SIV-endemic gorillas confers some level of protection against multiple SIVcpz strains.

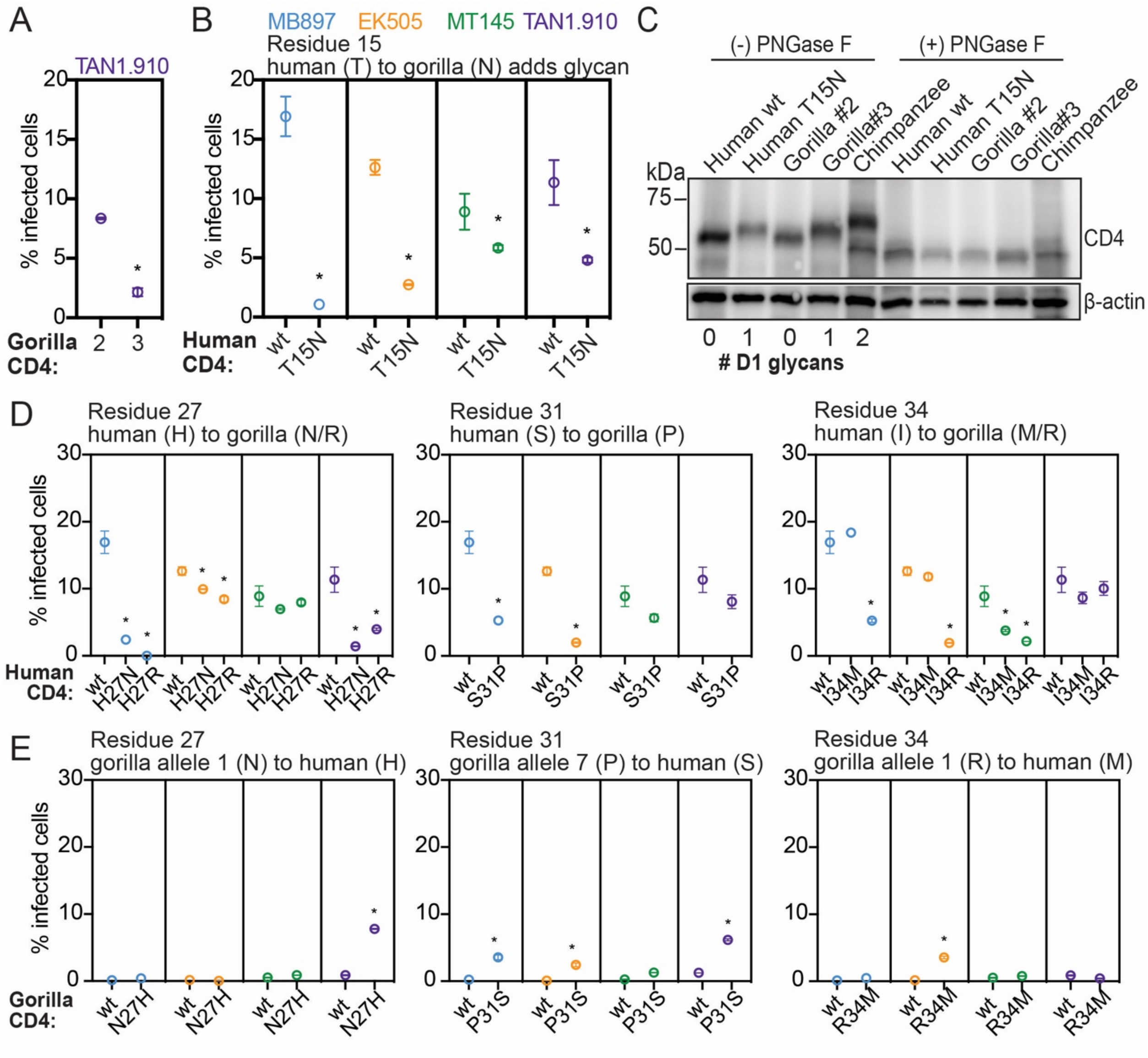


Figure 5. The CD4 SNPs found in gorilla populations are functionally significant. (A-E) HIV-1ΔEnv-GFP viruses were pseudotyped with Envs from diverse SIVcpz isolates (MB897, blue; EK505, orange; MT145, green; TAN1.910, purple). Cf2Th cells stably expressing human CCR5 and wild-type (wt) or human or gorilla CD4s with point mutations (X-axis) were infected with these pseudoviruses and then the percent cells infected (GFP-positive cells) were enumerated by flow cytometry 48 hours post infection. Data represent the mean +/- SEM from two independent experiments, each with two technical replicates. Stars above data sets signify that both independent experiments showed significant statistical differences ($p < 0.05$) when compared to wild-type by one-way ANOVA. **(C)** Lysates of Cf2Th cells stably expressing the indicated CD4 receptors in "A" and "B" were treated with PNGase F (to remove N-specific glycans) or left untreated and then probed for CD4 expression by western blotting. The number of N-specific glycosylation sites

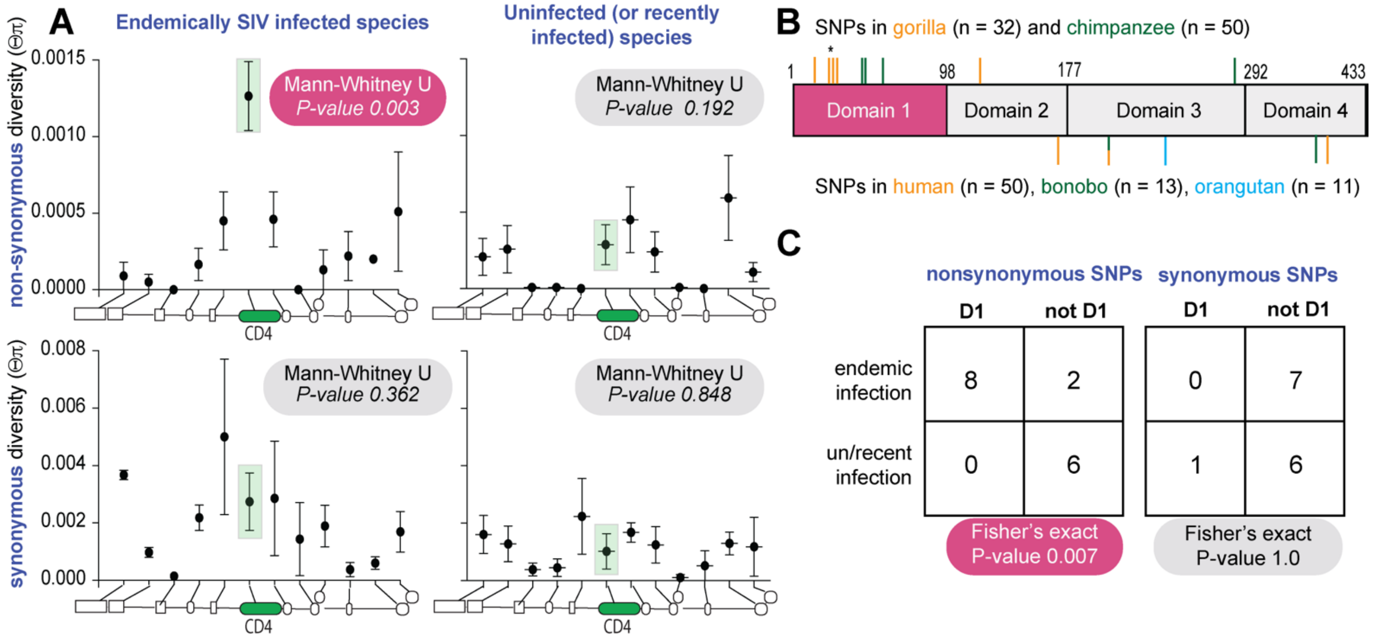
within the D1-domain of CD4 was determined computationally (Gupta and Brunak, 2002) and is shown under the blot. β-Actin served as a loading control.

Positive selection has shaped the evolutionary trajectory of CD4 SNPs in SIV endemic host species.

Natural selection influences the frequency of SNPs within populations. SNPs with deleterious effects will be kept at low frequency by purifying selection. On the other hand, mutations that confer a selective advantage will reach higher frequencies and/or be maintained in a population longer than expected due to different forms of positive selection (i.e., selective sweeps or frequency dependent selection). We next tested how polymorphism in CD4 has been shaped in ape species, with comparison made between apes that have been endemically infected with SIV (chimpanzees and gorillas) and those that have not (bonobos and orangutans).

Formal methods to detect the influence of selection on population-level nucleotide variation exist (Fay and Wu, 2000; Fu and Li, 1993; Tajima, 1989), but their statistical power is decreased in non-human ape species due to their small sample sizes and lower levels of variation (Prado-Martinez et al., 2013). Thus, we use a comparative approach to detect the signature of natural selection. To do this, we compare patterns of population level diversity between CD4 verus its neighboring genes. We compared SIV endemic apes (chimpanzee and gorillas) to apes uninfected (bonobos and orangutans) or recently infected on an evolutionary timescale (humans). This was performed for both nonsynonymous (protein altering) and synonymous (not protein altering) polymorphisms. Polymorphism data for chimpanzee, gorilla, bonobo, and two orangutan species were obtained from the Great Ape Genome Project (Prado-Martinez et al., 2013) for CD4 and 11 neighboring genes spanning 250 kb of the X-chromosome. For these same genes, human variation was obtained from the 1000 Genomes project. We calculated nucleotide diversity either based on the number of single nucleotide polymorphisms (SNPs; Watterson's $\boldsymbol{\theta}_{\omega}$(Watterson, 1975)) or mean pairwise difference between individuals ($\boldsymbol{\theta}_{\pi}$ (Tajima, 1983)). The mutation rate at CD4 does not appear to be elevated given similar levels of variation between CD4 and its neighboring loci when based on the number of SNPs ($\boldsymbol{\theta}_{\omega}$) (**Table S1** and **Fig. S2**).

However, within the endemically infected species, nonsynonymous SNPs in CD4 are at a significantly higher frequency compared to neighboring loci, represented by $\boldsymbol{\theta}\pi$, (**Fig. 6A**) (Tajima, 1983). This difference is not observed for synonymous variation. This discordance between nonsynonymous and synonymous variation suggests that the higher frequency of nonsynonymous variants at CD4 in the endemically infected species is not explained by neutral or demographic evolutionary forces. In addition, the higher frequency of segregating nonsynonymous variation is restricted to the endemically infected species. Taken together, these patterns are consistent with positive selection increasing the frequency of and/or maintaining nonsynonymous polymorphisms at CD4 within the endemically infected species only. Also, in support of this, we find that gorilla and chimpanzee nonsynonymous polymorphic sites are significantly concentrated on the domain 1 of CD4 when compared to the un/recently infected species (**Fig. 6B**). This difference is statistically significant **(Fig. 6C)**. This data suggests that long-term endemic infection of SIV in ape populations may be driving nonsynonymous SNPs to higher frequency in CD4, particularly on the domain 1, the region that directly interacts with the primate lentivirus Env glycoprotein.



	nonsynonymous SNPs		synonymous SNPs	
	D1	not D1	D1	not D1
endemic infection	8	2	0	7
un/recent infection	0	6	1	6

Figure 6. Positive selection has shaped CD4 polymorphism in host species where SIV has long been endemic. (**A**) Mean and standard error of mean of synonymous and nonsynonymous nucleotide heterozygosity (pi) at CD4 and neighboring loci across species endemically infected with SIV (chimpanzee and gorilla) or recently/uninfected (human, bonobo and orangutans). Schematic along bottom of each graph depicts the relative location of each locus as follows 5' to 3': ZNF384, PIANP, COPS7A, MLF2, PTMS, CD4, GPR162, GNB3, CDCA3, TPI1, LRRC23 and ENO2. Mann-Whitney test indicates whether heterozygosity at CD4 is significantly different than neighboring loci. (**B**) Schematic of CD4 domain regions. Ticks above and below the CD4 box indicate the location of polymorphic sites for the infected and un/recent infected species groups, respectively. One of the polymorphic residues in gorilla contains two non-synonymous changes in a single codon, marked by a star above the tick. (**C**) 2x2 contingency table and test results comparing synonymous and nonsynonymous polymorphism location relative to domain 1 between infected and recently/uninfected species groups.

DISCUSSION

Pathogens are strong selective drivers of host gene evolution. We and others have previously shown that the CD4 gene has evolved under strong positive selection throughout the evolution and speciation of simian primates (Meyerson et al., 2014; Zhang et al., 2008). Selection on CD4 is thought to be driven by its direct interaction with the lentiviral envelope (Env), which mediates viral entry. Indeed, most of the sequence evolution in CD4 has occurred in the D1 domain that contacts the lentivirus Env (Meyerson et al., 2014; Zhang et al., 2008). We performed an updated analysis of positive selection in CD4 including new CD4 orthologs that have become available (**Figure S4**). We found that removing the D1 sequence from the analysis renders the gene no longer under positive selection. This sets the stage for the current study, which focuses on selection on CD4 within ape populations.

Within populations of animals, when alleles of CD4 arise that can resist SIV, they would be predicted to rise in frequency. We and others have demonstrated that many alleles circulating in chimpanzees convey an increased ability to restrict viral entry by SIVs (Bibollet-Ruche et al., 2019; Warren et al., 2019b). This is also observed for many other African primate species, where amino acid polymorphisms in CD4 resist viral entry (Russell et al., 2021). It is important to note that many African primate species still harbor lentiviruses endemically, meaning that CD4 remains functional for viral entry of at least some viruses despite the selective pressure to resist it. Therefore, we understand CD4 to be evolving to convey natural tolerance to primates. "Natural tolerance" refers to a species' ability to resist or tolerate a virus to an acceptable level for peaceful co-existence of the virus and host. It is obtained by evolutionary adaptations that occur over time, allowing the species to develop mechanisms to reduce the negative effects of the virus (Pagán and García-Arenal, 2018). For example, some species may evolve barriers (like resistant forms of CD4) that reduce the titers that a virus can achieve in their body. Natural tolerance is often required before a virus can establish itself long-term in a host reservoir, and thus understanding it is key to understanding virus reservoirs in nature.

Herein, we strengthen the insight into lentiviral tolerance via CD4 evolution in three ways. 1) We reconstruct extinct ancestral forms of ape CD4 that pre-date SIV, and find that they were highly vulnerable to SIV entry. We then show that CD4 became less permissive to SIV in species that experienced long-term endemic infection. This resistant phenotype is associated with the accumulation of specific amino acid substitutions in the D1 domain of CD4. 2) We show that gorillas harbor a diversity of CD4 alleles, all of which are more resistant to SIV entry than is human CD4. Again, we show that these alleles are gaining resistance by accumulating amino acid substitutions in the D1 domain, one of which creates a new motif for post-translational addition of a glycan to the CD4 protein. Protective (to the host) glycosylation of CD4 has recently also been observed by us and others in chimpanzees (Bibollet-Ruche et al., 2019; Warren et al., 2019b), and in another population sample of gorillas (Russell et al., 2021). Indeed, the evolutionary acquisition of a glycan shield on CD4 may be a recurring theme in the evolution of primate species that are plagued with SIVs (Russell et al., 2021). 3) Using population genetics analysis, we show that nonsynonymous SNPs are enriched within ape species that are endemically infected with SIV (chimpanzees and gorillas) relative to those that have not (bonobos and orangutans) or which have been infected for less than 100 years (humans). This increased population level diversity is observed only for CD4, and not shared by other genes neighboring the CD4 loci.

Collectively, it is now clear that the sequence diversity (within species) and divergence (between species) of primate CD4 is strongly driven by infection pressure from lentiviruses. There is a surprising outcome of virus-driven host evolution in that the divergence and diversity of these host genes ultimately comes at a detriment to the very viruses that drove this evolution. When host genes like CD4 become highly diverse within species, a given virus strain may only be able to infect a small number of individuals within the population. For instance, gorilla CD4 allele 1 is highly resistant to most of the SIVs we tested **(Fig. 4)**. In fact, we only found one SIV isolate, SIVcpz “TAN2.69,” that could enter cells through the receptor encoded by allele 1. That suggests that gorillas that are homozygous for allele 1 would largely be protected from most circulating SIV strains. Taking this example further, if allele 1 were to become high frequency within gorilla populations, many strains of SIV in gorillas could go extinct.

In the long-term, the virus-driven evolution of genes like CD4 also mean that virus spillover between species – including the zoonotic spillovers that yield new human viruses - are less likely to happen. Indeed, a prevailing theme that has emerged in recent years is that receptor sequence divergence serves as a potent barrier to the movement of viruses between species. Likewise, this study suggests that SIV entry is blocked by the CD4 receptor of some primate species, and some individuals, that it might encounter. Therefore, spillover of lentiviruses between species will only happen when virus is transmitted between *key individuals* of two different species. The donor individual would need to have CD4 alleles that yield high titers of SIV in its body, and the recipient individual would need to have CD4 alleles that make it receptive to infection by this new virus.

Materials and Methods

Ancestral reconstruction of the CD4 sequence at the base of the hominin and hominid clades

The ancestral state of CD4 was determined using the PAML software package as previously described (Yang, 2007a, 2007b; Yang et al., 1995). As input, we used an alignment of CD4 sequences from the following species: human (*Homo* sapiens; NM_000616.4), common chimpanzee (*Pan troglodytes*; NM_001009043.1), western lowland gorilla (*Gorilla gorilla gorilla*; XM_004052582.2), bonobo (*Pan paniscus*; XM_008973678.1), northern white-cheeked gibbon (*Nomascus leucogenys*; XM_004092147.1), Sumatran orangutan (*Pongo abelii*; XM_024256502.1), rhesus monkey (*Macaca mulatta*; NM_001042662.1), green monkey (*Chlorocebus sabaeus*; XM_007967413.1), sooty mangabey (*Cercocebus atys*; NM_001319342.1), pig-tailed macaque (NM_001305921.1), crab-eating macaque (*Macaca nemestrena*; XM_005569956.2), gelada (*Theropithecus gelada*; XM_025401282.1), black snub-nosed monkey (*Rhinopithecus bieti*; XM_017891844.1), drill (*Mandrillus leucophaeus*; XM_011982990.1), Angolan colobus (*Colobus angolensis palliatus*; XM_011952091.1), golden snub-nosed monkey (*Rhinopithecus roxellana*; XM_010385914.1), and olive baboon (*Papio anubis*; XM_003905871.3).

Genotype and allele determination of CD4 from gorillas

Short-read data available through the National Center for Biotechnology Information's (NCBI) Short Read Archive (BioProject PRJNA189439) were mapped onto the *G. gorilla* genome using BWA-MEM (Li, 2013). We applied GATK base quality score recalibration, indel realignment, duplicate removal, and SNP discovery and genotyping in each individual separately (McKenna et al., 2010). Joint genotyping and variant recalibration was performed in a species-specific manner and in accordance to the GATK best practices recommendations (Auwera et al., 2013; DePristo et al., 2011). Variant recalibration was performed using SNPs called by the neighbor quality score method of ssahaSNP on capillary sequencing runs from NCBI's Trace Read Archive (Ning et al., 2001), dbSNP (if available), and high-quality SNPs called on the hg18 genome lifted over to the assembly used for mapping (Prado-Martinez et al., 2013). Processing was performed using custom scrips written in Python. Nucleotide sequence data reported are available in the

Third Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: BK063765-BK063795.

Receptor expression constructs and site directed mutagenesis.

Human (Genbank ID# MK170450) and chimpanzee (Genbank ID# NM_001009043.1) CD4 expression plasmids were constructed in a previous study (Warren et al., 2019a). The chimpanzee CD4 allele tested here is “allele 6” as defined by us previously (Warren et al., 2019b), and has 2 glycans that impede virus binding to the receptor. Gorilla CD4 alleles and ancestral CD4s were commercially synthesized (IDT GeneBlocks) and gateway cloned into the pLPCX retroviral packaging vector (Clontech). Mutant versions of human and gorilla CD4 were constructed by standard site-directed mutagenesis methods using overlapping PCR primers encoding the modification. Both wild-type and mutant CD4 constructs were analyzed by Sanger sequencing prior to use.

Generation of stable cell lines expressing CD4

HEK293T (ATCC CRL-11268) were cultured in DMEM (Invitrogen) with 10% FBS, 2 mM L-glutamine, and 1X penicillin-streptomycin (complete medium) at 37 °C and 5% CO2. Cf2Th (ATCC CRL-1430) cells stably expressing human CCR5 (from (Warren et al., 2019a)) were cultured in complete medium supplemented with 250 μg/mL hygromycin. To produce retroviruses for transduction, HEK293T cells plated in antibiotic free media (1x10^6 cells per well in a six well plate) were transfected with 2 μg of pLPCX transfer vector containing the CD4 gene of interest (or empty), 1 μg of pCS2-mGP (MLV gag/pol), and 0.2 μg of pC-VSV-G (VSV-G envelope) using a 3:1 ratio of TransIT-293 (Mirus) transfection reagent to DNA according to the manufacturer’s instructions. Forty-eight hours post transfection, supernatant was collected, filtered through 0.22 μm cellulose acetate filters, and retrovirus stored at -80 °C in single-use aliquots. Cf2Th cells stably expressing human CCR5 were plated at 2x10^4 cells per well of a 12-well dish (15% confluent) and 24-h later, transduced with 500 μL of retroviral supernatant by spinoculation at 1,200 xg for 75 min in the presence of 5 μg/mL polybrene. Forty-eight hours post transduction, the cells were placed in complete medium containing selection antibiotics (250 μg/mL hygromycin and 3 μg/mL puromycin) and cultured until stable outgrowth was noted (>1 week). Stable cell lines were maintained indefinitely in selection media. To confirm expression of CD4, cells were analyzed by flow cytometry (**Fig. S3**). Briefly, cells were harvested from culture plates, washed two times with PBS, fixed in 2% paraformaldehyde, and washed 2 times in flow buffer (1X PBS, 2% FBS, 1mM EDTA). Fixed cells were stained for 30 min at 4° C with PerCP-Cy5.5 mouse anti-human CD195 (CCR5, BD Biosciences 560635) and AlexaFluor647 mouse anti-human CD4 (BD Biosciences, 566681), and analyzed using a BD Accuri C6 Plus flow cytometer (BD Biosciences).

HIV/SIV Envelope clones used in this study.

Envelope clones for HIV-1 and SIVcpz EK505 and MB897 were constructed in a previous study (Warren et al., 2019b). SIVcpz MT145, TAN1.910, and TAN2.69 molecular clones were a gift from Brandon Keele, Frederick National Laboratory for Cancer Research, Frederick, MD and used as template for PCR amplification. The RevEnv cassettes of SIVcpz were amplified by PCR using the following primer pairs, where the lowercase sequence corresponds to an added Kozak sequence for enhanced translation: MT145 (JN835462) forward 5'-tcgccaccATGGCAGGAAGAAGCGAGGGAGACG-3', reverse 5'-TTAAAGCAAAGCTCTTTCTAAGCCTTGT-3'; TAN1.910 (AF447763.1) forward 5'-tcgccaccATGGCAGGAAGAGAAGAGGACGC-3', reverse 5'- TTAATTTAAGGCTAGTTCCAGACCC-3'; TAN2.69 (DQ374657.1) forward 5'-tcgccaccATGGCAGGAAGAGAAGAGGACGC-3', reverse 5'-TTAATTTAAGGCTATTTCTAGACCCTGT-3'. PCR products were cloned into the pCR8/GW/TOPO TA plasmid (Thermo Fisher) and then shuttled into a Gateway-converted pCDNA3.1 mammalian expression vector (Invitrogen).

Single-cycle HIV and SIV pseudovirus infections

To produce HIV-1ΔEnv-eGFP reporter viruses, 13x10^6 HEK293T cells were seeded into a 15-cm dishes in antibiotic free media and 24 h later transfected with 13.25 μg of Q23ΔEnv-GFP (group M backbone; (Humes and Overbaugh, 2011)) and 6.75 μg of envelope plasmid. Forty-eight hours post transfection, the cell supernatant was harvested, concentrated (~100-fold) using Amicon Ultracel 100K filters (Millipore), and stored at -80 °C in single use aliquots. Cf2Th cells stably expressing CD4 and CCR5 were plated at 3x10^4

cells/well of a 48-well plate 24 h before infection. The cells (~80% confluent) were then infected with HIV-1 pseudoviruses in three different volumes (Fig 2 and 4), or a volume corresponding to 10-20% infection of cells expressing human CD4 (Fig 5). Infections were carried by spinoculation at 1,200 xg for 75 min in the presence of 5 μg/mL of polybrene. Forty-eight hours post infection, the cells were harvested from the plate and fixed in 2% paraformaldehyde. Fixed cells were washed three times with PBS and resuspended in 50 μL flow buffer (1X PBX buffer containing 2% FBS and 1 mM EDTA) and stained for 30 min at 4 °C with the following antibody mixture: PerCP-Cy5.5 mouse anti-human CD195 (CCR5, BD Biosciences 560635) and AlexaFluor647 mouse anti-human CD4 (BD Biosciences, 566681), and analyzed using a BD Accuri C6 Plus flow cytometer (BD Biosciences). Following singlet cell discrimination, gates were drawn to capture double-positive cells expressing CD4 and CCR5, and then the percent GFP+ cells was enumerated within that population. The data from ~2x10^4 cells per technical replicate were analyzed using FlowJo v10. To calculate virus titers (Fig 2 and 4), the linear range of the infectivity curve was determined, and two points within the linear range were selected to calculate the mean virus titer in TDU/mL. The limit of detection for the titer calculation corresponds to a value of 0.2 % GFP positive cells. TDU/mL mean values were normalized to the titer of infection in cells expressing human CD4, and data used to construct a heat map using the Morpheus server (https://software.broadinstitute.org/morpheus); rows and columns were hierarchically clustered by Euclidian distance.

Statistical comparisons were performed between percentages of infected cells in some cases. Values of technical replicates of each biological replicate were compared between mutant and wild type CD4 versions by one-way ANOVA. If a statistically significant difference was found ($p < 0.05$) in both independent biological replicates, an asterisk was added to the mutant column in the dot plot.

Glycosylation state of CD4 by western blotting.

Cf2Th cells stably expressing CD4 cells were lysed in Nonidet P-40 buffer [150 mM NaCl, 50 mM Tris·HCl pH 7.4, 1% Nonidet P-40 substitute, 1 mM DTT, 1 μL/mL Benzonase (Sigma-Aldrich #E1014), and protease inhibitor mixture (Sigma- Aldrich, #11873580001)] by resuspending the cell pellet and rocking at 4° C for 30 min. Cell lysate was cleared by centrifugation at maximum speed for 15 min. Whole-cell extracts were quantified using the BCA assay and 10 μg was subjected to PNGase F treatment according to the manufacturer's protocol, including a paired sample with no glycosidase as control (New England Biolabs, #P0705S). Treated whole cell extracts (5 μg per lane) were resolved on a 12% TGX Stain-free polyacrylamide gel (Bio-Rad, #1610185) by applying 180V until loading dye ran off the gel. Protein was transferred to a PVDF membrane (Millipore Sigma, #IPVH07850) using a wet transfer apparatus set at 100V for 60 min. The membrane was incubated with blocking buffer (tris-buffered saline 1X, Tween-20 0.1%, 5% milk) for 60 min at room temperature. Primary antibodies were diluted in blocking buffer and incubated with the membrane overnight at 4° C (1:1,000 anti-CD4, Abcam #ab133616). After primary Ab incubation, membrane was washed 4 x 5 min in TBST (0.1% Tween-20). Secondary antibodies were diluted in blocking buffer and incubated with the membrane for 60 min at room temperature (1:10,000 anti-rabbit-HRP, Promega #W401B). After secondary Ab incubation, membrane was washed 4 x 5 min in TBST (0.1% Tween-20) and developed using ECL reagent (Sigma-Aldrich, #GERPN2232), and imaged on a Bio-Rad ChemiDoc Imaging System. As loading control, membranes were reblotted to detect β-acting using a primary (Cell Signal #3700) and secondary (Promega #W402B) antibodies and developed as described.

Analysis of population-level selection acting on CD4

To compare the pattern of molecular evolution at CD4 relative to neighboring loci we pulled population level re-sequencing data for loci located within 100 kb downstream and upstream of CD4. Primate sequences were obtained from the Great Ape Genome project (Prado-Martinez et al., 2013) and size-matched human sequences were selected to represent diverse ethnic groups from Human 1000 Genomes project.

To identify the individual-specific SNPs within the selected loci, genotype data in variant call format (VCF) was directly downloaded from International Genome Sample Resources (internationalgenome.org/) and the Great Ape Genome Project (biologiaevolutiva.org/greatape/). For human variants, the variant calls were made based on human reference genome annotation hg38, and individual-specific haplotypes were extracted by altering the reference sequence with the alternative SNPs annotated in the VCF files via custom Perl script. For non-human primate variants, the primate genome short read sequences were mapped to human reference genome hg19 to generate the VCF files, as previously described(Prado-Martinez et al., 2013). The SNPs in the VCF files were further filtered by the variant call quality (GQ ≥ 15). Like the human sequences, the individual-specific haplotype sequences are re-constructed by correcting the reference sequence with VCF annotations.

In total we obtained population level variation for CD4 plus 15 other loci (six upstream and nine downstream). Four loci were removed from analysis because they have previously been shown to directly interact with a viral protein (USP5 and SPSB2; (Jia et al., 2020; Rathore et al., 2020; Wang et al., 2019; Zhang et al., 2021)) or non-human primate sequencing reads did not map well with the human reference due to repetitive sequence (LAG3 and P3H3). Coding loci included in this study (in order 5' to 3') are: ZNF384, PIANP, COPS7A, MLF2, PTMS, CD4, GPR162, GNB3, CDCA3, TPI1, LRRC23 and ENO2. This was done for great ape species endemically infected with immunodeficiency viruses (chimpanzee and gorilla) and those newly or not infected (human, bonobo, Sumatran and Bornean orangutans).

Sequences were aligned for each species individually using the Muscle alignment program (Edgar, 2004). DnaSP (Rozas et al., 2017) was used to haplotype-phase the downloaded sequences and to calculate levels of nucleotide diversity for each locus. Rarely we would observe an internal stop codon within a locus' reading frame. In these cases, both haplotypes for that individual were removed from analysis. We analyzed the subspecies of gorilla and chimpanzee together. While there is evidence of genetic differentiation between these subspecies (Prado-Martinez et al., 2013) this should not affect our comparisons as the differentiation is expected to be similar across all loci.

Analysis of positive selection of CD4 in primates

Sequence alignments. *CD4* sequences were aligned to the longest human isoform in MEGA X for macOS (Stecher et al., 2020) using the ClustalW alignment tool. Multiple sequence alignments were visually inspected, duplicate gene sequences were removed, and the gene isoform from each species that best aligned to the human reference was retained for further analysis. The terminal stop codon was removed and aligned DNA and protein sequences were exported as fasta files. Codon alignments were generated using PAL2NAL (Suyama et al., 2006). Species cladograms for use in PAML were constructed following the species-level phylogenetic relatedness of primates (Perelman et al., 2011). Cladograms were generated using Newick formatted files and viewed with Njplot version 2.3.

Evolutionary analysis. Codon alignments and unrooted species cladograms were used as input files for analysis of positive selection using the PAML4.8 software package (Yang, 2007a). To detect selection, multiple sequence alignments were fit to the NSites models M7 (neutral model, codon values of dN/dS fit to a beta distribution bounded between 0 and 1), M8a (neutral model, similar to M7 but with an extra codon class fixed at dN/dS = 1) and M8 (positive selection model, similar to M8a but with the extra codon class allowed to have a dN/dS > 1). A likelihood ratio test was performed to assess whether the model of positive selection (M8) yielded a significantly better fit to the data compared to null models (model comparisons M7 vs. M8 and M8a vs M8). Posterior probabilities (Bayes Empirical Bayes analysis) were assigned to individual codons with dN/dS values > 1. To calculate the posterior mean of ω over a sliding window, the per-site ω value was extracted from the M8 model, and the average ω value within the designated window size (80 amino acids) was calculated across the open reading frame in a sliding manner. With the window slide 1 amino acid each time to calculate the smoothed mean ω values.

Acknowledgements

This work was funded by the NIH (DP1-DO-33547, DP1-DA-046108, R01-AI-137011, R01-OD-034046 to SLS; T32 A1007447-25, F32 GM125442, and K99 AI151256 to CJW). The flow cytometry work was performed at the BioFrontiers Institute Flow Cytometry Core supported by NIH Grant S10ODO21601.

Supplemental Figures and Table

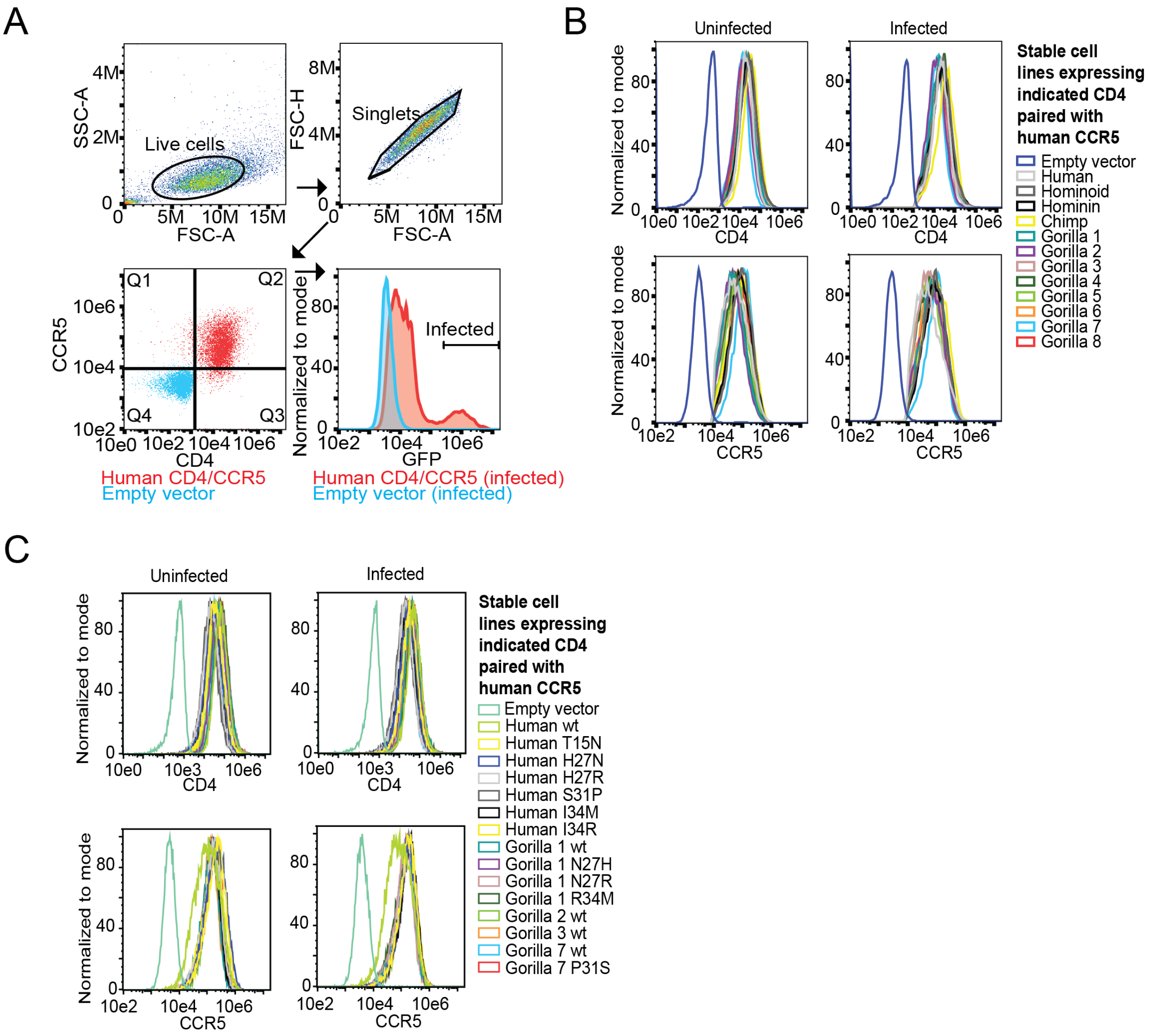


Figure S1. Flow cytometry gating strategy. A) Collected events were selected for live cells and singlets based on forward and side scatter values. Singlets were gated for CD4 and CCR5 fluorescent signal and then the double-positive poulation (Q2) was further analyzed for viral infection based on a shift in GFP fluorescence compared to virus exposed cells lacking CD4/CCR5 receptors (empty vector transduced cells). B) Expression levels for CD4 and CCR5 were compared amongst all stable cell lines under uninfected and infected conditions, demonstrating that viral infection does not impact receptor expression levels. For empty vector control cells, singlets were used for comparison. Data shown are representative of multiple independent experiments.

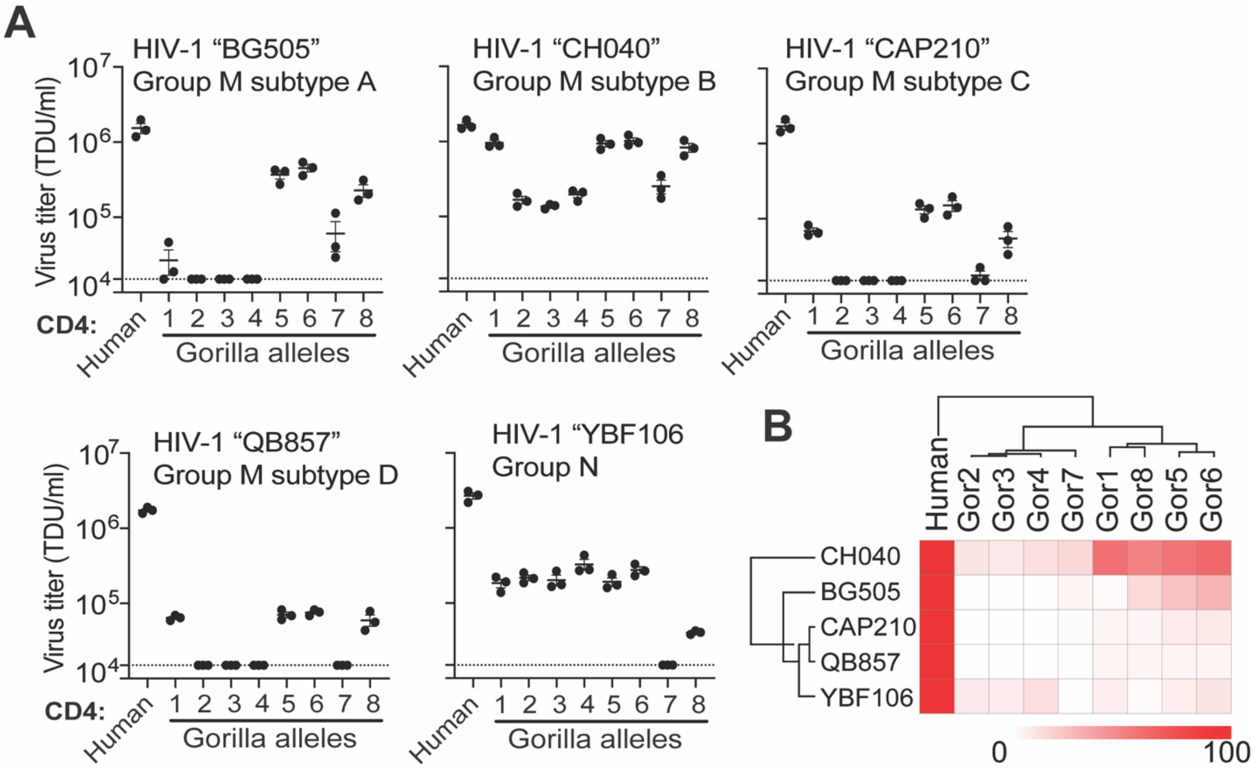


Figure S2. Gorilla CD4 alleles differentially support entry of HIV-1. (A) HIV-1ΔEnv-GFP viruses were pseudotyped with Envs (top of graphs) from globally diverse HIV strains. Cf2Th cells stably expressing human CCR5 and various CD4s (X-axis) were infected with various volumes of these pseudoviruses and then analyzed by flow cytometry 48 hours post infection. GFP positive cells were enumerated and virus titers (transducing units per milliliter; TDU/mL) were determined for those samples falling within the linear infection range (n = 2 titration points). The mean virus titers obtained from each of three independent experiments were plotted (dots), with error bars representing the standard error of the mean (SEM). **(B)** Data from each pseudotyped Env in D were used to calculate virus titer means normalized to human CD4 expressing cells and were plotted as a heat map. CD4 alleles and Envs were hierarchically clustered to depict similarities in phenotype.

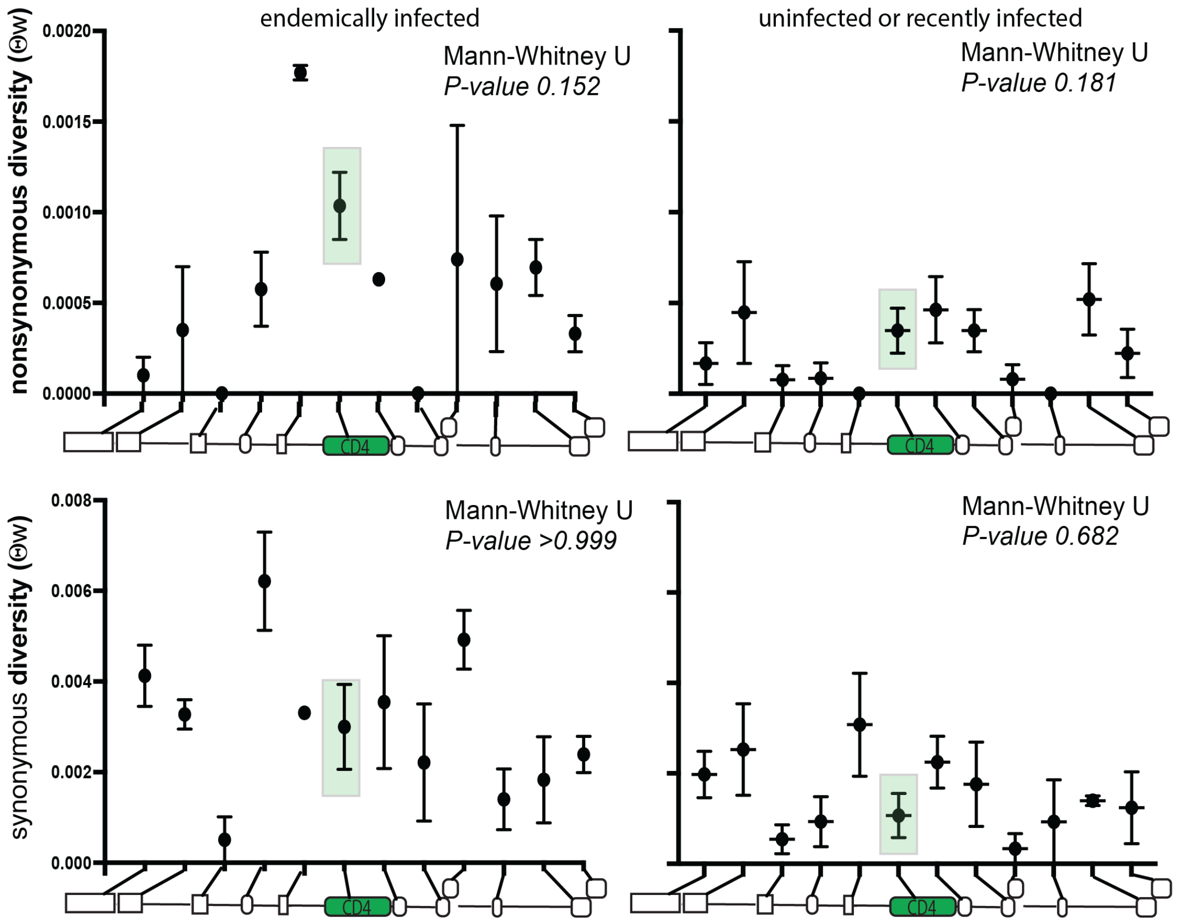
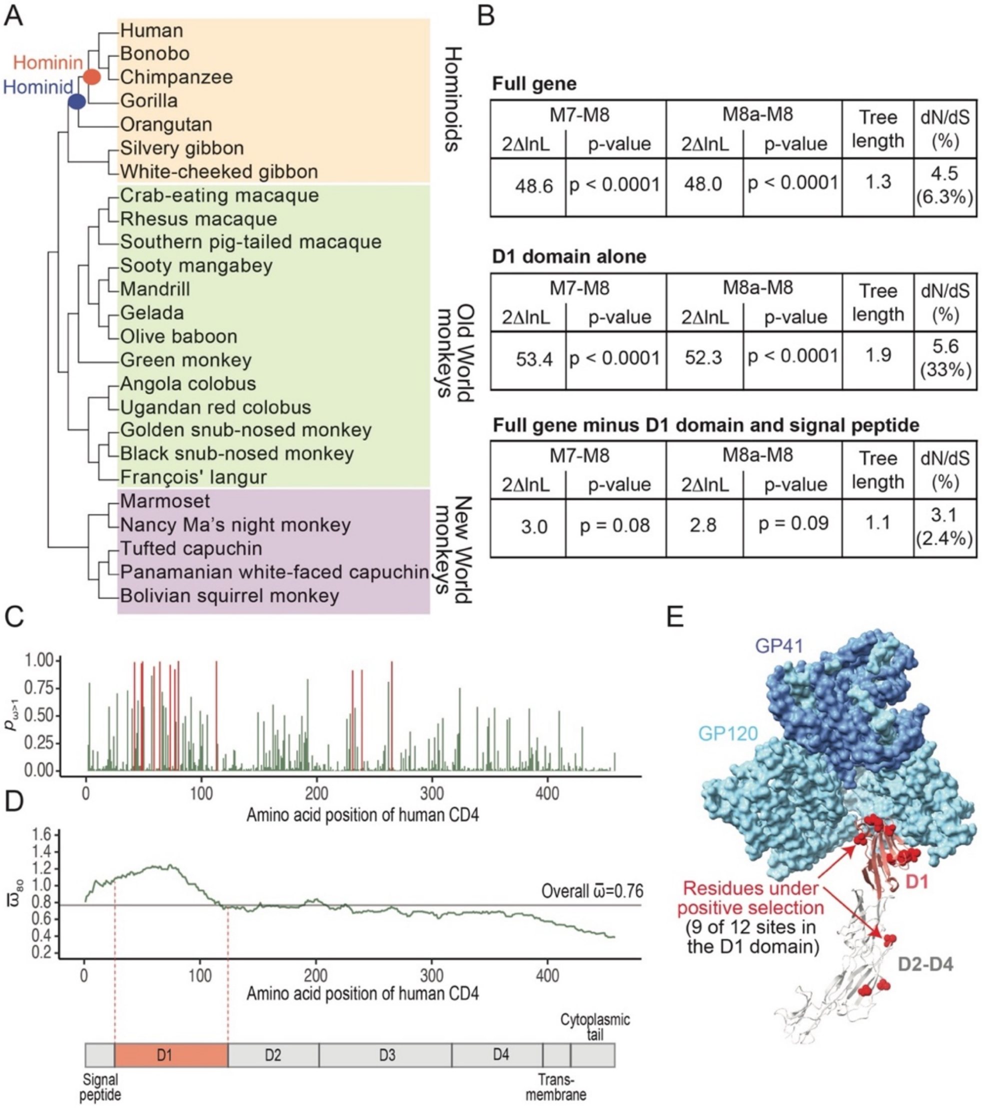


Figure S3. Single nucleotide polymorphisms in ape species. Population nucleotide diversity at a locus is estimated either based on the number of single nucleotide polymorphisms (SNPs; Watterson's $\boldsymbol{\theta}_{\omega}$(Watterson, 1975)) or mean pairwise difference between individuals ($\boldsymbol{\theta}_{\pi}$ (Tajima, 1983)). Level of variability based on the number of single nucleotide polymorphisms at a locus ($\boldsymbol{\theta}_{\pi}$) is not significantly different between CD4 and neighboring loci. Y-axis shows the mean and standard error of $\boldsymbol{\theta}_{\pi}$ for synonymous and nonsynonymous nucleotide variants at CD4 and neighboring loci across species endemically infected with SIV (chimpanzee and gorilla) or recently/uninfected (human, bonobo and orangutans). Schematic along bottom of each graph depicts the relative location of each locus and are as follows 5' to 3': ZNF384, PIANP, COPS7A, MLF2, PTMS, CD4, GPR162, GNB3, CDCA3, TPI1, LRRC23 and ENO2. Mann-Whitney test indicates whether heterozygosity at CD4 is significantly different than neighboring loci. We observe no difference in the total number of nonsynonymous and synonymous SNPs, represented by $\boldsymbol{\theta}_{\omega}$ between CD4 and its genomic neighbors in the endemic and un/recently infected species (see also **Table S1**).



Full gene

M7-M8		M8a-M8		Tree length	dN/dS (%)
2ΔlnL	p-value	2ΔlnL	p-value		
48.6	p < 0.0001	48.0	p < 0.0001	1.3	4.5 (6.3%)

D1 domain alone

M7-M8		M8a-M8		Tree length	dN/dS (%)
2ΔlnL	p-value	2ΔlnL	p-value		
53.4	p < 0.0001	52.3	p < 0.0001	1.9	5.6 (33%)

Full gene minus D1 domain and signal peptide

M7-M8		M8a-M8		Tree length	dN/dS (%)
2ΔlnL	p-value	2ΔlnL	p-value		
3.0	p = 0.08	2.8	p = 0.09	1.1	3.1 (2.4%)

Figure S4. CD4 is under positive selection in primates. Previous studies have identified CD4 as an HIV-1 cofactor that is evolving under positive (diversifying) selection (Meyerson et al., 2014; Zhang et al., 2008). However, these studies were limited in that they analyzed *CD4* sequences from a narrow set of primate species (Zhang et al., 2008), or included a larger species panel but lacked the complete *CD4* coding sequence (Meyerson et al., 2014). To extend on these studies, we collected full-length *CD4* sequences from 25 primate species (**Fig. 1A**) and tested for evidence of site-specific selective pressures using the *codeml* program on the Phylogenetic Analysis by Maximum Likelihood (PAML) package. **(A)** Cladogram of the primate species (n = 25) analyzed in this study. **(B)** The most amino-terminal extracellular domain of CD4 (domain 1, D1) is bound by the primate lentivirus (HIV/SIV) envelope glycoprotein (Env) during entry (Bour et al., 1995). We next sought to assess whether D1 alone is evolving under positive selection (presumably due to selective pressures exerted by SIVs), or if other regions of CD4 are also experiencing selective pressures for diversification. Site-specific selective pressures in primate CD4 (full gene; top), the CD4 D1 domain alone (amino acids 26-123; middle), and CD4 minus the signal peptide and the D1 domain (amino acids 123-458; bottom) were detected using the phylogenetic analysis maximum-likelihood (PAML) program (Yang, 2007a). Positive selection among amino acid sites was tested using two model comparisons, M7 vs. M8 and M8a vs. M8. In each of these comparisons, the null models (M7, M8a) do not allow for sites under positive selection, while the alternative model (M8) does. Tables summarize the likelihood ratio test between the M7-M8 and M8a-M8 models. The 2ΔlnL value (twice the difference in the natural log of the likelihoods) is shown, along with the p-value with which the neutral models (M7 or M8a) are rejected in favor of the model of positive selection (M8). **(C)** To further identify codon sites in CD4 under positive selection, we calculated the posterior probability of ω > 1 (where ω is the dN [nonsynonymous]/dS [synonymous] rate ratio, and values > 1 in the model M8 indicate sites under selection) using the Bayes empirical Bayes approach. Plot of posterior probabilities (ω>1 under maximum likelihood random-sites model M8) for all CD4 sites. Sites under positive selection ($p_{\omega} > 0.9$) are shown in red. (**D**) The posterior mean of ω over a sliding window of 80 amino acids is shown (green line), along with the overall mean of ω across the entire gene (grey line). In both panels C and D, the amino acid positions are shown in relationship to human CD4, and the D1 domain of CD4 is highlighted in orange. **(E)** Cryo-EM structure of an HIV-1 Env trimer in complex with human CD4 (PDB 5U1F) was visualized in ChimeraX (Goddard et al., 2017). Individual gp120 and gp41 subunits are colored in light and dark blue, respectively. The CD4 D1 domain (red) and D2-D4 domains (gray) are shown, with sites under positive selection ($P_{\omega} > 0.9$) shown on the human sequence as red spheres. 9 of the 12 sites passing this stringent cutoff map to the Env-CD4 D1 domain interface.

Table S1. Table S1 Levels for nonsynonymous and synonymous variation at CD4 and 11 genomically neighboring loci. For each locus we list the sample size and number of segregating nonsynonymous and synonymous single nucleotide polymorphisms (SNPs) for each species included in this study. We also include two measurements of nucleotide variability. One based on the number of SNPs ($\boldsymbol{\theta}$w) and the other based on the frequency of each SNP ($\boldsymbol{\theta}\pi$).

(this will be an extra excel file, it is too big for a word table)

REFERENCES

Auwera GAV der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics* **43**:11.10.1-33. doi:10.1002/0471250953.bi1110s43

Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, Hahn BH, Sharp PM. 2003. Hybrid Origin of SIV in Chimpanzees. *Science* **300**:1713–1713. doi:10.1126/science.1080657

Bibollet-Ruche F, Russell RM, Liu W, Stewart-Jones GBE, Sherrill-Mix S, Li Y, Learn GH, Smith AG, Gondim MVP, Plenderleith LJ, Decker JM, Easlick JL, Wetzel KS, Collman RG, Ding S, Finzi A, Ayouba A, Peeters M, Leendertz FH, Schijndel J van, Goedmakers A, Ton E, Boesch C, Kuehl H, Arandjelovic M, Dieguez P, Murai M, Colin C, Koops K, Speede S, Gonder MK, Muller MN, Sanz CM, Morgan DB, Atencia R, Cox D, Piel AK, Stewart FA, Ndjango J-BN, Mjungu D, Lonsdorf EV, Pusey AE, Kwong PD, Sharp PM, Shaw GM, Hahn BH. 2019. CD4 receptor diversity in chimpanzees protects against SIV infection. *P Natl Acad Sci Usa* **116**:3229–3238. doi:10.1073/pnas.1821197116

Bour S, Geleziunas R, Wainberg MA. 1995. The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. *Microbiol Rev* **59**:63–93. doi:10.1128/mr.59.1.63-93.1995

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Angel G del, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**:491–498. doi:10.1038/ng.806

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

Fay JC, Wu C-I. 2000. Hitchhiking Under Positive Darwinian Selection. *Genetics* **155**:1405–1413. doi:10.1093/genetics/155.3.1405

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**:693–709. doi:10.1093/genetics/133.3.693

Gavel Y, Heijne G von. 1990. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng* **3**:433–442. doi:10.1093/protein/3.5.433

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. 2017. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci Publ Protein Soc* **27**:14–25. doi:10.1002/pro.3235

Gupta R, Brunak S. 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symposium Biocomput Pac Symposium Biocomput* 310–22.

Heuverswyn FV, Li Y, Neel C, Bailes E, Keele BF, Liu W, Loul S, Butel C, Liegeois F, Bienvenue Y, Ngolle EM, Sharp PM, Shaw GM, Delaporte E, Hahn BH, Peeters M. 2006. SIV infection in wild gorillas. *Nature* **444**:164–164. doi:10.1038/444164a

Humes D, Overbaugh J. 2011. Adaptation of Subtype A Human Immunodeficiency Virus Type 1 Envelope to Pig-Tailed Macaque Cells. *J Virol* **85**:4409–4420. doi:10.1128/jvi.02244-10

Jia P, Zhang W, Xiang Y, Lu X, Liu W, Jia K, Yi M. 2020. Ubiquitin-specific protease 5 was involved in the interferon response to RGNNV in sea perch (Lateolabrax japonicus). *Fish Shellfish Immun* **103**:239–247. doi:10.1016/j.fsi.2020.04.065

Keele BF, Heuverswyn FV, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JFY, Sharp PM, Shaw GM, Peeters M, Hahn BH. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**:523–526. doi:10.1126/science.1126531

Klatt NR, Silvestri G, Hirsch V. 2012. Nonpathogenic Simian Immunodeficiency Virus Infections. *Csh Perspect Med* **2**:a007153–a007153. doi:10.1101/cshperspect.a007153

Li H. 2013. Aligning sequences reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303.3997v2**:1–3.

Liu Q, Acharya P, Dolan MA, Zhang P, Guzzo C, Lu J, Kwon A, Gururani D, Miao H, Bylund T, Chuang G-Y, Druz A, Zhou T, Rice WJ, Wigge C, Carragher B, Potter CS, Kwong PD, Lusso P. 2017. Quaternary contact in the initial interaction of CD4 with the HIV-1 envelope trimer. *Nat Struct Mol Biol* **24**:370–378. doi:10.1038/nsmb.3382

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**:1297–1303. doi:10.1101/gr.107524.110

Meyerson NR, Rowley PA, Swan CH, Le DT, Wilkerson GK, Sawyer SL. 2014. Positive selection of primate genes that promote HIV-1 replication. *Virology* **454–455**:291–8. doi:10.1016/j.virol.2014.02.029

Meyerson NR, Sawyer SL. 2011. Two-stepping through time: mammals and viruses. *Trends Microbiol* **19**:286–294. doi:10.1016/j.tim.2011.03.006

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res* **11**:1725–1729. doi:10.1101/gr.194201

Ohainle M, Malik HS. 2021. A balancing act between primate lentiviruses and their receptor. *Proc National Acad Sci* **118**:e2104741118. doi:10.1073/pnas.2104741118

Pagán I, García-Arenal F. 2018. Tolerance to Plant Pathogens: Theory and Experimental Evidence. *Int J Mol Sci* **19**:810. doi:10.3390/ijms19030810

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J. 2011. A molecular phylogeny of living primates. *PLoS Genet* **7**:e1001342. doi:10.1371/journal.pgen.1001342

Plantier J-C, Leoz M, Dickerson JE, Oliveira FD, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F. 2009. A new human immunodeficiency virus derived from gorillas. *Nat Med* **15**:871–872. doi:10.1038/nm.2016

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann

M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. 2013. Great ape genetic diversity and population history. *Nature* **499**:471–475. doi:10.1038/nature12228

Rathore A, Iketani S, Wang P, Jia M, Sahi V, Ho DD. 2020. CRISPR-based gene knockout screens reveal deubiquitinases involved in HIV-1 latency in two Jurkat cell models. *Sci Rep-uk* **10**:5350. doi:10.1038/s41598-020-62375-3

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol Biol Evol* **34**:3299–3302. doi:10.1093/molbev/msx248

Russell RM, Bibollet-Ruche F, Liu W, Sherrill-Mix S, Li Y, Connell J, Loy DE, Trimboli S, Smith AG, Avitto AN, Gondim MVP, Plenderleith LJ, Wetzel KS, Collman RG, Ayouba A, Esteban A, Peeters M, Kohler WJ, Miller RA, François-Souquiere S, Switzer WM, Hirsch VM, Marx PA, Piel AK, Stewart FA, Georgiev AV, Sommer V, Bertolani P, Hart JA, Hart TB, Shaw GM, Sharp PM, Hahn BH. 2021. CD4 receptor diversity represents an ancient protection mechanism against primate lentiviruses. *Proc National Acad Sci* **118**. doi:10.1073/pnas.2025914118

Sharp PM, Hahn BH. 2011. Origins of HIV and the AIDS pandemic. *CSH Perspect Med* **1**:a006841. doi:10.1101/cshperspect.a006841

Sharp PM, Shaw GM, Hahn BH. 2005. Simian Immunodeficiency Virus Infection of Chimpanzees. *J Virol* **79**:3891–3902. doi:10.1128/jvi.79.7.3891-3902.2005

Stecher G, Tamura K, Kumar S. 2020. Molecular evolutionary genetics analysis (MEGA) for macOS. *Mol Biol Evol* **37**:1237–1239. doi:10.1093/molbev/msz312

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**:W609–W612. doi:10.1093/nar/gkl315

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595. doi:10.1093/genetics/123.3.585

Tajima F. 1983. EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS. *Genetics* **105**:437–460. doi:10.1093/genetics/105.2.437

Takehisa J, Kraus MH, Ayouba A, Bailes E, Heuverswyn FV, Decker JM, Li Y, Rudicell RS, Learn GH, Neel C, Ngole EM, Shaw GM, Peeters M, Sharp PM, Hahn BH. 2009. Origin and Biology of Simian Immunodeficiency Virus in Wild-Living Western Gorillas. *J Virol* **83**:1635–1648. doi:10.1128/jvi.02311-08

Takehisa J, Kraus MH, Decker JM, Li Y, Keele BF, Bibollet-Ruche F, Zammit KP, Weng Z, Santiago ML, Kamenya S, Wilson ML, Pusey AE, Bailes E, Sharp PM, Shaw GM, Hahn BH. 2007. Generation of Infectious Molecular Clones of Simian Immunodeficiency Virus from Fecal Consensus Sequences of Wild Chimpanzees. *J Virol* **81**:7463–7475. doi:10.1128/jvi.00551-07

Wang M, Wang Y, Liu Y, Wang H, Xin X, Li J, Hao Y, Han L, Yu F, Zheng C, Shen C. 2019. SPSB2 inhibits hepatitis C virus replication by targeting NS5A for ubiquitination and degradation. *Plos One* **14**:e0219989. doi:10.1371/journal.pone.0219989

Warren CJ, Meyerson NR, Dirasantha O, Feldman ER, Wilkerson GK, Sawyer SL. 2019a. Selective use of primate CD4 receptors by HIV-1. *PLoS Biol* **17**:e3000304. doi:10.1371/journal.pbio.3000304

Warren CJ, Meyerson NR, Stabell AC, Fattor WT, Wilkerson GK, Sawyer SL. 2019b. A glycan shield on chimpanzee CD4 protects against infection by primate lentiviruses (HIV/SIV). *Proc National Acad Sci* **116**:11460–11469. doi:10.1073/pnas.1813909116

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**:256–276. doi:10.1016/0040-5809(75)90020-9

Yang Z. 2007a. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Yang Z. 2007b. PAML User Manual.

Yang Z, Kumar S, Nei M. 1995. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequence. *Genetics* **141**:1641–1650.

Zhang H, Zheng H, Zhu J, Dong Q, Wang J, Fan H, Chen Y, Zhang X, Han X, Li Q, Lu J, Tong Y, Chen Z. 2021. Ubiquitin-Modified Proteome of SARS-CoV-2-Infected Host Cells Reveals Insights into Virus–Host Interaction and Pathogenesis. *J Proteome Res* **20**:2224–2239. doi:10.1021/acs.jproteome.0c00758

Zhang ZD, Weinstock G, Gerstein M. 2008. Rapid Evolution by Positive Darwinian Selection in T-Cell Antigen CD4 in Primates. *J Mol Evol* **66**:446–456. doi:10.1007/s00239-008-9097-1