

1 **Modern-Day SIV Viral Diversity Generated by Extensive Recombination and**
2 **Cross-Species Transmission.**

3

4 Sidney M. Bell^{1,2,*}, Trevor Bedford¹

5 ¹ Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Wash-
6 ington, United States of America

7 ² Molecular and Cellular Biology Program, University of Washington, Seattle, Washington, United
8 States of America

9 * Corresponding author

10 E-mail: sbell23@fredhutch.org (SB)

11 **Abstract**

12 Cross-species transmission (CST) has led to many devastating epidemics, but is still a poorly under-
13 stood phenomenon. HIV-1 and HIV-2 (human immunodeficiency virus 1 and 2), which have collectively
14 caused over 35 million deaths, are the result of multiple CSTs from chimpanzees, gorillas, and sooty
15 mangabeys. While the immediate history of HIV is known, there are over 45 lentiviruses that infect spe-
16 cific species of primates, and patterns of host switching are not well characterized. We thus took a phy-
17 logenetic approach to better understand the natural history of SIV recombination and CST. We mod-
18 eled host species as a discrete character trait on the viral phylogeny and inferred historical host switch-
19 es and the pairwise transmission rates between each pair of 24 primate hosts. We identify 14 novel,
20 well-supported, ancient cross-species transmission events. We also find that lentiviral lineages vary
21 widely in their ability to infect new host species: SIV*col* (from colobus monkeys) is evolutionarily isolat-
22 ed, while SIV*agms* (from African green monkeys) frequently move between host subspecies. We also
23 examine the origins of SIV*cpz* (the predecessor of HIV-1) in greater detail than previous studies, and
24 find that there are still large portions of the genome with unknown origins. Observed patterns of CST
25 are likely driven by a combination of ecological circumstance and innate immune factors.

26

27

28 **Introduction**

29 As demonstrated by the recent epidemics of EBOV and MERS, and by the global HIV pandem-
30 ic, viral cross-species transmissions (CST) can be devastating [1,2]. As such, understanding the pro-
31 pensity and ability of viral pathogens to cross the species barrier is of vital public health importance [3].
32 Of particular interest are transmissions that not only “spillover” into a single individual of a new host
33 species, but that result in a virus actually establishing a sustained chain of transmission and becoming
34 endemic in the new host population (“host switching”).

35 HIV is the product of not just one successful host switch, but a long chain of host switch events
36 [4,5]. There are two human immunodeficiency viruses, HIV-2 and HIV-1. HIV-2 arose from multiple
37 cross-species transmissions of SIV_{smm} (simian immunodeficiency virus, *sooty mangabey*) from sooty
38 mangabeys to humans [6–8]. HIV-1 is the result of four independent cross-species transmissions from
39 chimpanzees and gorillas. Specifically, SIV_{cpz} was transmitted directly from chimpanzees to humans
40 twice; one of these transmissions generated HIV-1 group M, which is the primary cause of the human
41 pandemic [9]. SIV_{cpz} was also transmitted once to gorillas [10], generating SIV_{gor}, which was in turn
42 transmitted twice to humans [11].

43 Looking further back, SIV_{cpz} itself was also generated by lentiviral host switching and recombi-
44 nation. Based on the SIV sequences available at the time, early studies identified SIV_{mon/-mus/-gsn}
45 (which infect mona, mustached, and greater spot-nosed monkeys, respectively) and SIV_{rcm} (which in-
46 fects red-capped mangabeys) as probable donors [12]. Functional analysis of accessory genes from
47 these putative parental lineages indicate that the specific donors and genomic locations of these re-
48 combination event(s) were crucial for enabling what became SIV_{cpz} to cross the high species barrier
49 and establish an endemic lineage in hominids [13].

50 The complex evolutionary history resulting in HIV illustrates the importance of natural history to
51 modern day viral diversity, and although the history leading to HIV is well detailed, broader questions
52 regarding cross-species transmission of primate lentiviruses remain [14]. With over 45 known extant
53 primate lentiviruses, each of which is endemic to a specific host species [4,5,15], we sought to charac-

54 terize the history of viral transmission between these species to the degree possible given the precision
55 allowed from a limited sample of modern-day viruses.

56 Here, we utilized phylogenetic inference to reconstruct the evolutionary history of primate lentivi-
57 ral recombination and cross-species transmission. We assembled datasets from publicly available len-
58 tiviral genome sequences and conducted discrete trait analyses to infer rates of transmission between
59 primate hosts. We find evidence for extensive interlineage recombination and identify many novel host
60 switches that occurred during the evolutionary history of lentiviruses. We also find that specific lentiviral
61 lineages exhibit a broad range of abilities to cross the species barrier. Finally, we also examined the
62 origins of each region of the SIVcpz genome in greater detail than previous studies to yield a more nu-
63 ancanced understanding of its origins.

64

65 Results

66 **There have been at least 13 interlineage recombination events that confuse the lentiviral phy-**
67 **logeny.**

68 In order to reconstruct the lentiviral phylogeny, we first had to address the issue of recombi-
69 nation, which is frequent among lentiviruses [16]. In the context of studying cross-species transmission,
70 this is both a challenge and a valuable tool. Evidence of recombination between viral lineages endemic
71 to different hosts is also evidence that at one point in time, viruses from those two lineages were in the
72 same animal (i.e., a cross-species transmission event must have occurred in order to generate the ob-
73 served recombinant virus). However, this process also results in portions of the viral genome having
74 independent evolutionary—and phylogenetic—histories.

75 To address the reticulate evolutionary history of SIVs, we set out to identify the extent and na-
76 ture of recombination between lentiviral lineages. Extensive sequence divergence between lineages
77 masks site-based methods for linkage estimation (Fig. S1). However, topology-based measures of re-
78 combination allow for “borrowing” of information across nearby sites, and are effective for this dataset.
79 We thus utilized a phylogenetic model to group segments of shared ancestry separated by recombi-

80 tion breakpoints instantiated in the HyPhy package GARD [17]. For this analysis, we used a version of
81 the SIV compendium alignment from the Los Alamos National Lab (LANL), modified slightly to reduce
82 the overrepresentation of HIV sequences (N=64, see Methods) [18]. Importantly, because each virus
83 lineage has only a few sequences present in this alignment, these inferences refer to *inter*-lineage re-
84 combination, and not the rampant *intra*-lineage recombination common among lentiviruses.

85 GARD identified 13 locations along the genome that had strong evidence of inter-lineage re-
86 combination (Fig. 1). Here, evidence for a particular model is assessed via Akaike Information Criterion
87 (AIC) [19] and differences in AIC between models indicate log probabilities, so that a delta-AIC of 10
88 between two models would indicate that one model is $e^{10/2} = \sim 148$ as likely as the other. In our case,
89 delta-AIC values ranged from 154 to 436 for each included breakpoint, indicating that these breakpoints
90 are strongly supported by the underlying tree likelihoods. The 14 resulting segments ranged in length
91 from 351 to 2316 bases; in order to build reliable phylogenies, we omitted two of the less supported
92 breakpoints from downstream analyses, yielding 12 segments ranging in length from 606 – 2316 bases.
93 We found no evidence to suggest linkage between non-neighboring segments (Fig. S2). While it has
94 been previously appreciated that several lineages of SIV are recombinant products [12,20,21], the 13
95 breakpoints identified here provide evidence that there have been at least 13 inter-lineage recombina-
96 tion events during the evolution of SIVs. Identifying these recombination breakpoints allowed us to con-
97 struct a putatively valid phylogeny for each segment of the genome that shares an internally cohesive
98 evolutionary history.

99

100 **Most primate lentiviruses were acquired by cross-species transmissions.**

101 We then looked for phylogenetic evidence of cross-species transmission in the tree topologies
102 of each of the 12 genomic segments. For this and all further analyses, we constructed a dataset from
103 all publicly available primate lentivirus sequences, which we curated and subsampled by host and virus
104 lineage to ensure an equitable distribution of data (see Methods). This primary dataset consists of virus

105 sequences from the 24 primate hosts with sufficient data available (5 – 25 sequences per viral lineage,
106 N=423, Fig. S3). Alignments used the fixed compendium alignment as a template (see Methods).

107 In phylogenetic trees of viral sequences, cross-species transmission appears as a mismatch
108 between the host of a virus and the host of that virus's ancestor. To identify this pattern and estimate
109 how frequently each pair of hosts has exchanged lentiviruses, we used the established methods for
110 modeling evolution of discrete traits, as implemented in BEAST [22,23]. In our case, the host of each
111 viral sample was modeled as a discrete trait. This is analogous to treating the "host state" of each viral
112 sample as an extra column in an alignment, and inferring the rate of transition between all pairs of host
113 states along with inferring ancestral host states across the phylogeny. This approach is similar to com-
114 mon phylogeographic approaches that model movement of viruses across discrete spatial regions [24]
115 and has previously been applied to modeling discrete host state in the case of rabies virus [22]. Here,
116 we took a fully Bayesian approach and sought the posterior distribution across phylogenetic trees, host
117 transition rates and ancestral host states. We integrate across model parameters using Markov chain
118 Monte Carlo (MCMC). The resulting model provides phylogenetic trees for each segment annotated
119 with ancestral host states alongside inferred transition rates.

120 Figure 2 shows reconstructed phylogenies for 3 segments along with inferred ancestral host
121 states. Trees are color coded by known host state at the tips, and inferred host state at internal
122 nodes/branches; color saturation indicates the level of certainty for each ancestral host assignment. A
123 visual example of how the model identifies cross-species transmissions can be seen in the
124 SIV_{mon}/SIV_{tal} clade, which infect mona- and talapoin monkeys, respectively (starred in Fig. 2A-C).
125 Due to the phylogenetic placement of the SIV_{mon} tips, the internal node at the base of this clade is red,
126 indicating that the host of the ancestral virus was most likely a mona monkey. This contrasts with the
127 host state of the samples isolated from talapoin monkeys (tips in green). These changes in the host
128 state across the tree are what inform our estimates of the rate of transmission between host pairs. In
129 total, the support for each possible transmission is derived from both A) whether the transmission is

130 supported across the posterior distribution of phylogenies for a particular segment, and B) whether this
131 is true for multiple genomic segments.

132 Notably, the tree topologies are substantially different between segments, which emphasizes
133 both the extent of recombination and the different evolutionary forces that have shaped the phylogenies
134 of individual portions of the genome. In all segments' trees, we also see frequent changes in the host
135 state between internal nodes (illustrated as changes in color going up the tree), suggestive of frequent
136 ancient cross-species transmissions. On average, primate lentiviruses switch hosts once every 6.25
137 substitutions per site per lineage across the SIV phylogeny.

138 The cross-species transmission events inferred by the model are illustrated in Fig. 3, with raw
139 rates and Bayes factors (BF) in Fig. S4. As shown, the model correctly infers the pairs of hosts with
140 previously identified CST events [6,9,11,12,20,21,25,26]. Importantly, we also identify 14 novel cross-
141 species transmission events with strong statistical support (cutoff of BF ≥ 10.0). Each of these trans-
142 missions is clearly and robustly supported by the tree topologies (all 12 trees are illustrated in Fig. S5).

143 To control for sampling effects, we repeated the analysis with a supplemental dataset built with
144 fewer hosts, and more sequences per host (15 host species, subsampled to 16-40 sequences per viral
145 lineage, N=510), and see consistent results. As illustrated in Figs. S6-8, we find qualitatively similar re-
146 sults. When directly comparing the average indicator values, host state transition rates, and BF values
147 between analyses, the results from the main and supplemental datasets are strongly correlated, indicat-
148 ing robust quantitative agreement between the two analyses (Fig. S9). Taken together, these results
149 represent a far more extensive pattern of CST among primate lentiviruses than previously described,
150 with all but a few lineages showing robust evidence of host switching [4,5,14].

151 However, the distribution of these host switches is not uniform; when we assess the network
152 centrality of each virus we find a broad range (Figs. 3, S6, as node size), indicating some hosts act as
153 sources in the SIV transmission network and other hosts act as sinks. From this, we infer that some
154 viruses have either had greater opportunity or have a greater ability to cross the species barrier than
155 others. In particular, the SIVs from four of the subspecies of African green monkeys (SIVsab, SIVtan,

156 SIV*ver*, SIV*pyg*; collectively, SIV*agm*) appear to exchange viruses with other host species frequently
157 (Fig. 3, 12' o'clock). An example of SIV*agm* CST events can be seen in the tree topologies from *gag*,
158 *prot*, *reverse transcriptase (RT)*, and *vif* (Figs. 2 and S5, segments 2, 3, 4, & 6). Here, SIV*tan* isolates
159 reported by Jin *et al.* (denoted with † in Fig. 2) clearly cluster with SIV*sab*, in a distant part of the phy-
160 logeny from the rest of the SIV*agm* viruses (including the majority of SIV*tan* isolates) [20]. For all other
161 segments, however, the SIV*agms* cluster together. We thus concur with the conclusion of Jin *et al.* that
162 these samples represent a recent spillover of SIV*sab* from sabaeus monkeys to tantalus monkeys, and
163 the model appropriately identifies transmission.

164 Although SIV*agms* appear to be particularly prone to CST among African green monkeys, near-
165 ly every primate clade has at least one inbound, robustly supported viral transmission from another
166 clade. We thus conclude that the majority of lentiviruses have arisen from a process of host switching,
167 followed by a combination of intraclade host switches and host-virus coevolution.
168

169 **SIV*col* appears to be evolutionarily isolated.**

170 Previous studies of the lentiviral phylogeny have noted that SIV*col* is typically the outgroup to
171 other viral lineages, and have hypothesized that this may implicate SIV*col* as the “original” primate len-
172 tivirus [27]. We find this hypothesis plausible, but the evidence remains inconclusive. For the majority of
173 genomic segments, we also observe SIV*col* as the clear outgroup (Figs. 2 and S5). In contrast, for por-
174 tions of *gag/pol* (segments 3 and 4) and some of the accessory genes (segment 7), we find that there is
175 not a clear outgroup. For these segments, many other lineages of SIV are just as closely related to
176 SIV*col* as they are to each other. However, with the occasional exception of single heterologous taxa
177 with poorly supported placement, SIV*col* remains a monophyletic clade (N=16), and does not interca-
178 late within the genetic diversity of any other lineage in our dataset.

179 Based on these collective tree topologies, our model does not identify strong evidence for any
180 specific transmissions out of colobus monkeys, and identifies only a single, weakly supported inbound

181 transmission (likely noise in the model caused by the fact that red-capped mangabeys are the marginal-
182 ly supported root host state; see below). This is consistent with previous findings that the colobinae
183 have a unique variant of the *APOBEC3G* gene, which is known to restrict lentiviral infection and specu-
184 lated to be a barrier to cross-species transmission [27]. These observations generally support the idea
185 of SIV*col* as having maintained a specific relationship with its host over evolutionary time. However, we
186 find inconclusive evidence for the hypothesis that SIV*col* is the most ancient or “source” lineage of pri-
187 mate lentivirus; if this hypothesis were true, then these lentiviral transmissions out of the colobinae
188 must have been followed by extensive recombination that obscures the relationships between SIV*col*/
189 and its descendants.

190

191 **SIV*cpz*, the precursor to HIV-1, has a mosaic origin with unknown segments.**

192 Unlike SIV*col*, SIV*cpz* appears to be the product of multiple CSTs and recombination events.
193 SIV*cpz* actually encompasses two viral lineages: SIV*cpzPtt* infects chimpanzees of the subspecies *Pan*
194 *troglodytes troglodytes*, and SIV*cpzPts* infects chimpanzees of the subspecies *Pan troglodytes*
195 *schweinfurthii* [28]. There are two additional subspecies of chimpanzees that have not been found har-
196 bor an SIV despite extensive surveys, suggesting that SIV*cpzPtt* was acquired after chimpanzee sub-
197 speciation [26]. Both previous work [26,28] and our own results support the hypothesis that SIV*cpz* was
198 later transmitted from one chimpanzee subspecies to the other, and SIV*cpzPtt* is the only SIV*cpz* line-
199 age that has crossed into humans. Given the shared ancestry of the two lineages of SIV*cpz*, we use
200 “SIV*cpz*” to refer specifically to SIV*cpzPtt*.

201 Based on the lentiviral sequences available in 2003, Bailes et al [12] conclude that the SIV*cpz*
202 genome is a recombinant of just two parental lineages. SIV*rcm* (which infects red-capped mangabeys)
203 was identified as the 5' donor, and an SIV from the SIV*mon-/mus-/gsn* clade (which infect primates in
204 the *Cercopithecus* genus) was identified as the 3' donor. Since the time of this previous investigation
205 many new lentiviruses have been discovered and sequenced. In incorporating these new data, we find
206 clear evidence that the previous two-donor hypothesis may be incomplete.

207 The tree topologies from *env* in the 3' end of the genome (segments 8-11) support the previous
208 hypothesis [12] that this region came from a virus in the SIV*mon*-*mus*-*gsn* clade. These viruses form a
209 clear sister clade to SIV*cpz* with high posterior support (Fig. 2C,D). We find strong evidence for trans-
210 missions from mona monkeys (SIV*mon*) to mustached monkeys (SIV*mus*), and from mustached mon-
211 keys to greater spot-nosed monkeys (SIV*gsn*) (see discussion of potential coevolution below). We also
212 find more evidence in support of a transmission from mona monkeys to chimpanzees than from the
213 other two potential donors, but more sampling is required to firmly resolve which of these viruses was
214 the original donor of the 3' end of SIV*cpz*.

215 We find phylogenetic evidence to support the previous hypothesis [12,13] that the *integrase* and
216 *vif* genes of SIV*cpz* (segments 4-6) originated from SIV*rcm*; however, we find *equally strong evidence*
217 to support the competing hypothesis that *pol* came from SIV*mnd*-2, which infects mandrils (Fig. 2B,D).
218 In these portions of the genome, SIV*mnd*-2 and SIV*rcm* together form a clear sister clade to SIV*cpz*.
219 The *vpr* gene, in segment 7, is also closely related to both SIV*rcm* and SIV*mnd*-2, but this sister clade
220 also contains SIV*smm* from sooty mangabeys.

221 Interestingly, we do *not* find evidence to support either SIV*rcm/mnd*-2 or SIV*mon/mus/gsn* as
222 the donor for the 5' most end of the genome (segments 1-5), including the 5' *LTR*, *gag*, and *RT* genes.
223 This is also true for the 3' *LTR* (segment 12). SIV*cpz* lacks a clear sister clade or ancestor in this re-
224 gion, and SIV*rcm* groups in a distant clade; we therefore find no evidence to suggest that an ancestor
225 of an extant SIV*rcm* was the parental lineage of SIV*cpz* in the 5' most end of the viral genome as previ-
226 ously believed (Fig. 2A,D). This may support the possibility of a third parental lineage, or a number of
227 other plausible scenarios (discussed below).

228

229 **Discussion**

230 **Limitations and strengths of the model**

231 *Additional sampling is required to fully resolve the history of CST among lentiviruses.*

232 In addition to the 14 strongly supported novel transmissions ($BF \geq 10$) described above, we
233 also find substantial evidence for an additional 8 possible novel transitions, but with lower support (BF
234 ≥ 3) (Fig. 3, S4). These transmissions are more difficult to assess, because many of them are inferred
235 on the basis of just a few “outlier” tips of the tree that group apart from the majority of viral samples
236 from the same lineage. In each case, the tips’ phylogenetic position is strongly supported, and the pri-
237 mary literature associated with the collection of each of these “outlier” samples clearly specifies the
238 host metadata. However, due to the limited number of lentiviral sequences available for some hosts, we
239 are unable to control for sampling effects for some of these lower-certainty transmissions. We report
240 them here because it is unclear whether these outliers are the result of unidentified separate endemic
241 lineages, one-time spillovers from other hosts, or species misidentification during sample collection. It is
242 also important to note that while some of these less-supported transmissions are potentially sampling
243 artifacts, many of them may be real, and may be less supported simply because they lack the requisite
244 available data for some genome segments. Ultimately, far more extensive sampling of primate lentivi-
245 ruses is required in order to resolve these instances.

246

247 *Most lentiviruses were originally acquired by CST and have since coevolved with their hosts.*

248 Some of these “noisier” transmission inferences, particularly within the same primate clade, may
249 be the result of coevolution, ie. lineage tracking of viral lineages alongside host speciation. Within the
250 model, viral jumps into the *common ancestor* of two extant primate species appear as a jump into *one*
251 of the extant species, with a secondary jump between the two descendants. For example, the model
252 infers a jump from mona monkeys into mustached monkeys, with a secondary jump from mustached
253 monkeys into their sister species, red-tailed guenons (Fig. 3). Comparing the virus and host phyloge-
254 nies, we observe that this host tree bifurcation between mustached monkeys and red-tailed guenons is
255 mirrored in the virus tree bifurcation between SIV*mus* and SIV*rtg* for most segments of the viral ge-
256 nome (Figs. 2, 3). This heuristically suggests that the true natural history may be an ancient viral

257 transmission from mona monkeys into the common ancestor of mustached monkeys and red-tailed
258 guenons, followed by host/virus coevolution during primate speciation to yield SIV*mus* and SIV*rtg*.

259 The possibility of virus/host coevolution means that while we also observe extensive host
260 switching *between* primate clades, many of the observed jumps *within* a primate clade may be the re-
261 sult of host-virus coevolution. However, we also note that the species barrier is likely lower between
262 closely related primates, making it challenging to rigorously disentangle coevolution vs. true host
263 switches within a primate clade [29].

264

265 *The model propagates and accounts for residual uncertainty from the recombination analyses.*

266 The deep phylogenies and extensive sequence divergence between SIV lineages makes any
267 assessment of recombination imperfect. Our analysis aims to 1) place a lower bound on the number of
268 interlineage recombination events that must have occurred to explain the observed extant viruses, and
269 2) use this understanding to construct a model of CST among these viruses. As the most well devel-
270 oped package currently available for topology-based recombination analysis, GARD was an appropri-
271 ate choice to identify in broad strokes the extent and nature of recombination among SIVs. Some
272 (though not all) of the remaining uncertainty as to the exact location of breakpoints is represented within
273 the posterior distribution of trees for each segment, and is thus propagated and accounted for in the
274 discrete traits model (inferences of CST). Most other studies of SIV evolutionary history simply split the
275 phylogeny along gene boundaries or ignore recombination (e.g., [5,12,29]). Thus, while there is still
276 some remaining uncertainty in our recombination analyses, these results still represent a major step
277 forward in attempting to systematically assess recombination among all extant SIV lineages and to in-
278 corporate it into the phylogenetic reconstruction.

279

280 **Cross-species transmission is driven by exposure and constrained by host genetic distance.**

281 Paleovirology and, more recently, statistical models of lentiviral evolution estimate that primate
282 lentiviruses share a common ancestor approximately 5-10 million years ago [27,30,31]. This, along with

283 the putative viral coevolution during primate speciation, suggests that many of these transmissions
284 were ancient, and have been acted on by selection for millions of years. Thus, given that the observed
285 transmissions almost exclusively represent evolutionarily successful host switches, it is remarkable that
286 lentiviruses have been able to repeatedly adapt to so many new host species. In the context of this vast
287 evolutionary timescale, however, we conclude that while lentiviruses have a far more extensive history
288 of host switching than previously understood, these events remain relatively rare overall.

289 Looking more closely at the distribution of host switching among lineages, it is clear that some
290 viruses have crossed the species barrier with far greater frequency than others. This is likely governed
291 by both ecological and biological factors. Ecologically, frequency and form of exposure are likely key
292 determinants of transmission, but these relationships can be difficult to describe statistically [3]. For ex-
293 ample, many primates are chronically exposed to many exogenous lentiviruses through predation [32].
294 Using log body mass ratios [33] as a proxy for predation, we do not see a statistically significant asso-
295 ciation between body mass ratio and non-zero transmission rate (Fig. 4A, blue; $p=0.678$, coef. 95% CI
296 (-0.311, 0.202)). We believe the lack of signal is likely due to the imperfect proxy, although it is also
297 possible that predator-prey relationships do not strongly structure the CST network.

298 Biologically, increasing host genetic distance has a clear negative association with the probabili-
299 ty of cross-species transmission (Fig. 4B, blue, $p<0.001$, coef. 95% CI (-7.633, -2.213)). Importantly, as
300 already discussed, the strength of this association may be inflated by instances of lineage tracking (vi-
301 rus/host cospeciation). However, it is well established that increasing host genetic distance is associat-
302 ed with a higher species barrier; as previously documented in the literature (reviewed in [34]), we ex-
303 pect this is largely due to the divergence of host restriction factor genes, which are key components of
304 the innate immune system. Functional assays of these host restriction factors against panels of SIVs,
305 while outside the scope of this study, will be important for further identifying the molecular bases of the
306 species barriers that have led to the transmission patterns identified here.

307

308 **Origins of HIV-1 and HIV-2**

309

310 *Epidemiological factors were key to the early spread of HIV.*

311 Understanding the underlying dynamics of lentiviral CST provides important biological context to
312 the transmissions that generated the HIV pandemic. As discussed above, our results support a view of
313 lentiviral cross-species transmission as a rare event. Notably, only two lentiviruses have crossed the
314 high species barrier from Old World monkeys into hominids: SIV_{smm} and the recombinant SIV_{cpz}.
315 Both HIV-1 and HIV-2 have arisen in human populations in the last century [5,35]. While it is possible
316 that this has occurred by chance, even without increased primate exposure or other risk factors, we
317 nevertheless find it striking that humans would acquire two exogenous viruses within such a short evo-
318 lutionary timespan.

319 Examining this phenomenon more closely, the history of HIV-2 is enlightening. HIV-2 has been
320 acquired through at least 8 *independent* spillover events from sooty mangabeys [5]. Notably, 6 of these
321 transmissions have resulted in only a single observed infection (spillovers) [7,8,36]; only 2 of these
322 events have established sustained transmission chains and successfully switched hosts to become en-
323 demic human pathogens [37–40]. Given this pattern, it is conceivable that there have been many iso-
324 lated introductions of lentiviruses into humans over the past 200,000 years [14,41]. However, these
325 other viral exposures did not result in new endemic human pathogens either because of species-
326 specific immune barriers, non-conducive epidemiological conditions, or a combination thereof. The rap-
327 id and repeated emergence of HIV-1 and HIV-2 is on a timescale more congruent with changes in epi-
328 demiological conditions than mammalian evolution, perhaps emphasizing the importance of the concur-
329 rent changes in human population structure and urbanization in facilitating the early spread of the epi-
330 demic [35]. Significantly, though, this also highlights the importance of careful public health surveillance
331 and interventions to prevent future epidemics of zoonotic viruses.

332

333 *The exact origins of SIV_{cpz} may not be identifiable.*

334 The ancestry of SIV_{cpz} appears to be tripartite: the unknown ancestry of the 5' end; the puta-
335 tively SIV_{rcm} or SIV_{mnd-2} derived *int* and *vif* genes; and the putatively SIV_{gsn/mon/mus} derived 3' end
336 of the genome. For any of these three portions of the genome, there are multiple evolutionary histories
337 supported by available sequence data. However, it is possible that further sampling of lentivirus line-
338 ages (both known and currently undiscovered) will be able to definitively resolve the ancestry of
339 SIV_{cpz}. Alternatively, it may be that the ancestral virus(es) that gave rise to any of these three portions
340 is extinct. In the case of the last two portions of the SIV_{cpz} genome, it may be that the common ances-
341 tor of these putative genetic donors (SIV_{rcm/mnd-2} and SIV_{mon/mus/gsn}, respectively) was the true
342 source. However, it is also a distinct possibility that SIV_{cpz} has sufficiently diverged since its acquisition
343 by chimpanzees such that its history is obscured.

344

345 *Evolutionary time obscures the identity of the “original” primate lentivirus.*

346 Among primates, our results clearly illustrate that the vast majority of lentiviral lineages were
347 originally acquired by cross-species transmission. It is intriguing to speculate as to which virus was the
348 “original” source of all of these lineages. Because of its consistent position as the outgroup of primate
349 lentiviral trees, there has long been a popular hypothesis in the field that SIV_{co} was this original lentivi-
350 rus among primates [27]. While SIV_{co} is certainly the most evolutionary isolated extant lentivirus that
351 has been sampled to date, this does not definitively place it as the ancestral lentivirus. Alternative sce-
352 narios (also noted by Compton *et al.*) include an extinct original lentiviral lineage (and/or primate host
353 species) or an unsampled ancestral lentivirus. It is also plausible that another known extant lentivirus
354 was the “original” lineage, but has diverged and/or recombined to such an extent that its origins are ob-
355 scured.

356

357 *Conclusions*

358 Here, we have shown that lentiviruses have a far more extensive history of host switching than
359 previously described. Our findings also demonstrate that the propensity of each lentiviral lineage to

360 switch between distant hosts, or to spillover among related hosts, is highly variable. In examining spe-
361 cific lineages, our findings are consistent with the prevalent hypothesis that SIV*co*l has evolved in isola-
362 tion from other SIVs. Contrastingly, we have also demonstrated that the mosaic origins of SIV*cpz* are
363 far more complex than previously recognized; the currently available sequence data is unable to re-
364 solve the ancestry of nearly half of the SIV*cpz* genome. Together, our analyses move closer to a full
365 understanding of the pattern of cross-species transmission among primate lentiviruses, but additional
366 efforts to obtain high quality viral genome sequences from under sampled lineages will be necessary to
367 fully resolve the natural history of these viruses.

368 **Methods**

369 All datasets, config files, documentation, and scripts used in this analysis or to generate figures are
370 available publicly at <<https://github.com/blab/siv-cst>>.

371

372 **Datasets & Alignments**

373 Lentiviral genomes are translated in multiple reading frames; we therefore utilized nucleotide sequence
374 data for all analyses.

375

376 *Recombination*

377 For all recombination analyses (GARD, R², rSPR), we used the 2015 Los Alamos National Lab
378 (LANL) HIV/SIV Compendium [18]. The compendium is a carefully curated alignment of high-quality,
379 representative sequences from each known SIV lineage and each group of HIV-1 and HIV-2. We re-
380 duced the overrepresentation of HIV in this dataset, but maintained at least one high quality sample
381 from each group of HIV. In total, this dataset contains 64 sequences from 24 hosts (1-10 sequences
382 per host). Maximum likelihood trees for each segment—used for the rSPR analysis and displayed in
383 Figure 1—were built with RAxML v.8.2.9 [42] with the rapid bootstrapping algorithm under a GTR model
384 with 25 discrete bins of site to site rate variation and were rooted to SIVcol.

385

386 *Main & Supplemental Datasets*

387 Primate lentiviral sequences were downloaded from the comprehensive LANL HIV sequence
388 database [18]. Sequences from lineages known to be the result of artificial cross-species transmissions
389 (SIVmac, -stm, -mne, and –wcm) were excluded. We also excluded any sequences shorter than 500
390 bases or that were flagged as problematic by standard LANL criteria. We grouped host subspecies to-
391 gether except for cases where there is a known specific relationship between host subspecies and virus
392 lineage (chimpanzees and African green monkeys). To construct datasets with a more equitable distri-
393 bution of sequences per host, we preferentially subsampled sequences from the LANL Compendium,

394 followed by samples isolated from Africa (more likely to be primary sequences), and finally supplement-
395 ing with samples isolated elsewhere (excluding experimentally generated sequences). For humans,
396 mandrils and mustached monkeys, which are host to 2, 2 and 3 distinct viral lineages, respectively, we
397 allowed a few additional sequences (if available) to represent the full breadth of documented lentiviral
398 diversity. The “main” dataset consists of 24 host species, with 5-31 sequences per host (total N=422).
399 As an alternative dataset to control for sampling bias and data availability, we also constructed a “sup-
400 plemental” dataset with just 15 hosts but with more viral sequences per host (16 – 40 sequences per
401 lineage, N=510).

402

403 *Alignments*

404 Alignments were done with the L-INSI algorithm implemented in mafft v7.294b [43]; the Com-
405 pendium alignment was held fixed, with other sequences aligned to this template. Insertions relative to
406 the fixed compendium were discarded. This alignment was then split along the breakpoints identified by
407 GARD to yield the segment-specific alignments.

408

409 **Recombination**

410 *Topology-Based Analysis*

411 Each portion of the genome was analyzed with the Genetic Algorithm for Recombination Detec-
412 tion (GARD), implemented in HyPhy v.2.2.0 [17], with a nucleotide model selected by HyPhy’s Nu-
413 cModelCompare package (#012234) and general discrete distribution (3 bins) of site variation.

414 Computational intensity was eased by analyzing the recombination dataset in 3kb long portions,
415 with 1kb overlaps on either end (e.g., bases 1:3000, 2000:5000, 4000:7000, etc.). To control for sites’
416 proximity to the ends of the genomic portion being analyzed, this was repeated with the windows offset
417 (e.g., bases 1:2500, 1500:4500, 3500:6500, etc.). In total, this resulted in every site of the alignment
418 being analyzed at least two-fold, with at least one of these replicates providing 500-1500 bases of ge-
419 nomic context on either side (other than at the very ends of the total alignment). Disagreement be-

420 tween window-offset replicates for a given breakpoint were minimal and almost always agreed within a
421 few bases, with two exceptions: offset replicates for the breakpoint between segments 7 and 8 disa-
422 greed by 529 bases, and offset replicates for the breakpoint between 8 and 9 disagreed by 263 bases.
423 In these instances, we used the average site placement.

424

425 *Sitewise Linkage Analysis*

426 Biallelic sites were identified across the genome (ignoring gap characters and polymorphisms
427 present at <5%). These biallelic sites were compared pairwise to generate an observed and expected
428 distribution of haplotypes (combinations of alleles between the two sites), and assessed with the statis-
429 tic $R^2 = \chi^2/(d*N)$, where χ^2 is chi-square, d is the degrees of freedom, and N is the number of haplotypes
430 for this pair of sites [44]. This statistic follows the canonical interpretation of R^2 , i.e., if the allele at site 1
431 is known, how well does it predict the allele at site 2 (0 indicating no linkage between sites, and 1 indi-
432 cating perfect linkage).

433

434 *Segment Linkage Analysis*

435 Segment linkage was assessed by comparing the similarity of tree topologies between seg-
436 ments. This was done with the pairwise method in the Rooted Subtree-Prune-and-Regraft (rSPR)
437 package [45], which measures the number of steps required to transform one topology into another.
438 Segment pairs with similar topologies have lower scores than segments with less similar topologies.

439

440 **Phylogenies & Discrete Trait Analysis**

441 *Empirical tree distributions*

442 For each segment alignment, a posterior distribution of trees was generated with BEAST v.
443 1.8.2 under a general time reversible substitution model with gamma distributed rate variation and a
444 strict molecular clock [46]. We used a Yule birth-death speciation model tree prior [47]; all other priors
445 were defaults. Trees were estimated using a Markov chain Monte Carlo (MCMC). The main dataset

446 was run for 25 million steps, sampled every 15,000 states after discarding the first 10% as burnin, re-
447 sulting in ~1600 trees per segment in the posterior distribution. Convergence was determined by visual-
448 ly inspecting the trace in Tracer v. 1.6.0 [48]. All but two effective sample size (ESS) values (segment 2:
449 AG substitution rate [ESS=164] and frequencies 4 [ESS=189]) were well over 200. The supplemental
450 dataset was run for 65 million steps, sampled every 15,000 states after discarding the first 10% as
451 burnin (~3900 trees per segment in the posterior distribution). Convergence was again determined in
452 Tracer. All ESS values except for the AG substitution rate in segment 2 (ESS=190) were well above
453 200.

454

455 *Discrete Trait Analysis*

456 These tree distributions were used to estimate the transmission network. As in Faria *et al.* [22],
457 we model hosts as states of a discrete trait, and model cross-species transmission as a stochastic dif-
458 fusion process among n host states. We use a non-reversible state transition matrix with $n \times (n-1)$ indi-
459 vidual transition parameters [49]. We also utilize standard methodology in using a Bayesian stochastic
460 search variable selection process to encourage a sparse network, implemented as in [23]. An exponen-
461 tial distribution with mean=1.0 was used as a prior on each of the pairwise rate parameters (552 in total
462 for the main dataset with 24 hosts; 210 parameters in total for the supplemental dataset with 15 hosts).
463 The prior placed on the *sum* of the binary indicator variables corresponding to each pairwise rate pa-
464 rameter (i.e., the number of transitions that are non-zero) was an exponential distribution with
465 mean=20. Convergence was again assessed by visually inspecting the log files in Tracer. The main
466 dataset MCMC chain was run for 25 million steps and sampled every 5,000 steps after discarding the
467 first 5 million steps as burnin. The supplemental dataset MCMC chain was run for 45 million steps and
468 sampled every 5,000 steps after discarding the first 20 million steps as burnin. For both datasets, more
469 than 97% of the rate, indicator, and actualRate (stepwise rate*indicator values) parameters had an ESS
470 well over 200 (most > 1000).

471 Statistical support for each transmission is summarized as a Bayes factor (BF), calculated by
472 comparing the posterior and prior odds that a given rate is non-zero (i.e., that there has been *any*
473 transmission between a given pair of hosts) [23]. The ancestral tree likelihoods of each of the 12 tree
474 distributions contribute equally to the inference of a shared transmission rate matrix. However, not eve-
475 ry lineage has recombined along every breakpoint, which means that the tree likelihoods from each
476 segment are not fully statistically independent. Thus, we report conservative estimates of BF by dividing
477 all BF values by 12.

478

479 **Acknowledgements**

480 We thank Michael Emerman, Janet Young, Gytis Dudas, Chris Whidden, and the Emerman and Bed-
481 ford labs for their help. We also acknowledge the Los Alamos National Labs HIV database, which pro-
482 vided the curated data for this study. T.B. is a Pew Biomedical Scholar and his work is supported by
483 NIH award R35 GM119774-01. S.B. is a NSF Graduate Research Fellow and her work is supported by
484 NSF grant [DGE-1256082](#) and NIH training grant 5-T32-CA-80416-17.

485

486

487

488

489

- 490 1. Morens D, Folkers G, Fauci A. Emerging infections: a perpetual challenge. In: LANCET INFECTIOUS DISEASES [Internet]. 2008 [cited 11 Aug 2015] pp. 710–719. doi:10.1016/S1473-3099(08)70256-1
- 493
- 494 2. Parrish CR, Holmes EC, Morens DM, Park E-C, Burke DS, Calisher CH, et al. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev.* 2008;72: 457–70. doi:10.1128/MMBR.00004-08
- 497
- 498 3. Locatelli S, Peeters M. Cross-species transmission of simian retroviruses: how and why they could lead to the emergence of new diseases in the human population. *AIDS.* 2012;26: 659–73. doi:10.1097/QAD.0b013e328350fb68
- 501
- 502 4. Apetrei C, Robertson DL, Marx PA. The history of SIVs and AIDS: epidemiology, phylogeny and biology of isolates from naturally SIV infected non-human primates (NHP) in Africa. *Front Biosci.* 2004;9: 225–54. Available: <http://www.ncbi.nlm.nih.gov/pubmed/14766362>
- 505
- 506 5. Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med.* 2011;1: a006841. doi:10.1101/cshperspect.a006841
- 508
- 509 6. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature.* 1989;339: 389–92. doi:10.1038/339389a0
- 511
- 512 7. Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, et al. Human infection by genet-

- 513 ically diverse SIVSM-related HIV-2 in west Africa. *Nature*. 1992;358: 495–9.
514 doi:10.1038/358495a0
515
- 516 8. Chen Z, Telfier P, Gettie A, Reed P, Zhang L, Ho DD, et al. Genetic characterization of new
517 West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived
518 SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses
519 from a single feral sooty mangabey t. *J Virol*. 1996;70: 3617–27. Available:
520 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=190237&tool=pmcentrez&rendertype=abstract>
521
- 522
- 523 9. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, et al. Origin of HIV-1 in the
524 chimpanzee *Pan troglodytes troglodytes*. *Nature*. 1999;397: 436–41. doi:10.1038/17130
525
- 526 10. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, et al. Human immunodeficiency vi-
527 ruses: SIV infection in wild gorillas. *Nature*. 2006;444: 164. doi:10.1038/444164a
528
- 529 11. D'arc M, Ayouba A, Esteban A, Learn GH, Boué V, Liegeois F, et al. Origin of the HIV-1 group O
530 epidemic in western lowland gorillas. *Proc Natl Acad Sci*. 2015;112: 201502022.
531 doi:10.1073/pnas.1502022112
532
- 533 12. Etienne L, Hahn BH, Sharp PM, Matsen FA, Emerman M. Gene loss and adaptation to hominids
534 underlie the ancient origin of HIV-1. *Cell Host Microbe*. 2013;14: 85–92.
535 doi:10.1016/j.chom.2013.06.002
536

- 537 13. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, et al. Hybrid Origin of SIV
538 in Chimpanzees. *Science* (80-). 2003;300.
- 539
- 540 14. Aghokeng AF, Ayouba A, Mpoudi-Ngole E, Loul S, Liegeois F, Delaporte E, et al. Extensive sur-
541vey on the prevalence and genetic diversity of SIVs in primate bushmeat provides insights into
542risks for potential new cross-species transmissions. *Infect Genet Evol*. 2010;10: 386–96.
543 doi:10.1016/j.meegid.2009.04.014
- 544
- 545 15. Gifford RJ. Viral evolution in deep time: Lentiviruses and mammals. *Trends in Genetics*. 2012.
546 pp. 89–100. doi:10.1016/j.tig.2011.11.003
- 547
- 548 16. Chen J, Powell D, Hu W-S. High frequency of genetic recombination is a common feature of pri-
549mate lentivirus replication. *J Virol*. 2006;80: 9651–8. doi:10.1128/JVI.00936-06
550
- 551 17. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. Automated phylogenetic
552 detection of recombination using a genetic algorithm. *Mol Biol Evol*. 2006;23: 1891–901.
553 doi:10.1093/molbev/msl051
- 554
- 555 18. Los Alamos SIV/HIV Sequence Database [Internet]. 2015 [cited 11 Aug 2015]. Available:
556 <http://www.hiv.lanl.gov/content/index>
- 557
- 558 19. Jin MJ, Hui H, Robertson DL, Müller MC, Barré-Sinoussi F, Hirsch VM, et al. Mosaic genome
559 structure of simian immunodeficiency virus from west African green monkeys. *EMBO J*. 1994;13:
560 2935–47. Available: <http://www.ncbi.nlm.nih.gov/pubmed/8026477>

561

562 20. Souquière S, Bibollet-Ruche F, Robertson DL, Makuwa M, Apetrei C, Onanga R, et al. Wild
563 Mandrillus sphinx are carriers of two types of lentivirus. *J Virol. American Society for Microbiolo-*
564 *gy (ASM); 2001;75: 7086–96. doi:10.1128/JVI.75.15.7086-7096.2001*

565

566 21. Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously reconstructing viral
567 cross-species transmission history and identifying the underlying constraints. *Philos Trans R Soc
568 Lond B Biol Sci. 2013;368: Supplementary Text 1. doi:10.1098/rstb.2012.0196*

569

570 22. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots.
571 Fraser C, editor. *PLoS Comput Biol. Public Library of Science; 2009;5: e1000520.*
572 doi:10.1371/journal.pcbi.1000520

573

574 23. Jin MJ, Rogers J, Phillips-Conroy JE, Allan JS, Desrosiers RC, Shaw GM, et al. Infection of a
575 yellow baboon with simian immunodeficiency virus from African green monkeys: evidence for
576 cross-species transmission in the wild. *J Virol. American Society for Microbiology (ASM);
577 1994;68: 8454–60. Available: <http://www.ncbi.nlm.nih.gov/pubmed/7966642>*

578

579 24. Leitner T, Dazza M-C, Ekwalanga M, Apetrei C, Saragosti S. Sequence diversity among chim-
580 panzee simian immunodeficiency viruses (SIVcpz) suggests that SIVcpzPts was derived from
581 SIVcpzPtt through additional recombination events. *AIDS Res Hum Retroviruses. Mary Ann
582 Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA; 2007;23: 1114–8.*
583 doi:10.1089/aid.2007.0071

584

- 585 25. Compton AA, Emerman M. Convergence and divergence in the evolution of the APOBEC3G-Vif
586 interaction reveal ancient origins of simian immunodeficiency viruses. PLoS Pathog. 2013;9:
587 e1003135. doi:10.1371/journal.ppat.1003135
- 588
- 589 26. Charleston MA, Robertson DL. Preferential host switching by primate lentiviruses can account
590 for phylogenetic similarity with the primate phylogeny. Syst Biol. 2002;51: 528–35.
591 doi:10.1080/10635150290069940
- 592
- 593 27. Aiewsakun P, Katzourakis A. Time-Dependent Rate Phenomenon in Viruses. J Virol. American
594 Society for Microbiology; 2016;90: 7184–95. doi:10.1128/JVI.00593-16
- 595
- 596 28. Nerrienet E, Amouretti X, Müller-Trutwin MC, Poaty-Mavoungou V, Bedjebaga I, Nguyen HT, et
597 al. Phylogenetic analysis of SIV and STLV type I in mandrills (*Mandrillus sphinx*): indications that
598 intracolony transmissions are predominantly the result of male-to-male aggressive contacts.
599 AIDS Res Hum Retroviruses. 1998;14: 785–96. doi:10.1089/aid.1998.14.785
- 600
- 601 29. Harris RS, Hultquist JF, Evans DT. The restriction factors of human immunodeficiency virus. J
602 Biol Chem. 2012;287: 40875–83. doi:10.1074/jbc.R112.416925
- 603
- 604 30. Chen Z, Luckay A, Sodora DL, Telfer P, Reed P, Gettie A, et al. Human immunodeficiency virus
605 type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the
606 natural range of simian immunodeficiency virus-infected sooty mangabeys. J Virol. 1997;71:
607 3953–60. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9094672>
- 608

- 609 31. Pieniazek D, Ellenberger D, Janini LM, Ramos AC, Nkengasong J, Sasan-Morokro M, et al.
610 Predominance of human immunodeficiency virus type 2 subtype B in Abidjan, Ivory Coast. AIDS
611 Res Hum Retroviruses. 1999;15: 603–8. doi:10.1089/088922299311132
- 612
- 613 32. Ishikawa K, Janssens W, Banor JS, Shinno T, Piedade J, Sata T, et al. Genetic analysis of HIV
614 type 2 from Ghana and Guinea-Bissau, West Africa. AIDS Res Hum Retroviruses. 2001;17:
615 1661–3. doi:10.1089/088922201753342077
- 616
- 617 33. Peeters M, Toure-Kane C, Nkengasong JN. Genetic diversity of HIV in Africa: impact on diagno-
618 sis, treatment, vaccine development and trials. AIDS. 2003;17: 2547–60.
619 doi:10.1097/01.aids.0000096895.73209.89
- 620
- 621 34. Damond F, Descamps D, Farfara I, Telles JN, Puyeo S, Campa P, et al. Quantification of proviral
622 load of human immunodeficiency virus type 2 subtypes A and B using real-time PCR. J Clin Mi-
623 crobiol. 2001;39: 4264–8. doi:10.1128/JCM.39.12.4264-4268.2001
- 624
- 625 35. Apetrei C, Gaufin T, Gautam R, Vinton C, Hirsch V, Lewis M, et al. Pattern of SIVagm Infection in
626 Patas Monkeys Suggests that Host Adaptation to Simian Immunodeficiency Virus Infection May
627 Result in Resistance to Infection and Virus Extinction. J Infect Dis. Oxford University Press;
628 2010;202: S371–S376. doi:10.1086/655970
- 629
- 630 36. Gifford RJ. Viral evolution in deep time: lentiviruses and mammals. Trends Genet. 2012;28: 89–
631 100. doi:10.1016/j.tig.2011.11.003
- 632

- 633 37. Compton A a., Emerman M. Convergence and Divergence in the Evolution of the APOBEC3G-
634 Vif Interaction Reveal Ancient Origins of Simian Immunodeficiency Viruses. PLoS Pathog.
635 2013;9. doi:10.1371/journal.ppat.1003135
636
- 637 38. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements
638 in performance and usability. Mol Biol Evol. Oxford University Press; 2013;30: 772–80.
639 doi:10.1093/molbev/mst010
640
- 641 39. Whidden C, Matsen FA. Ricci-Ollivier Curvature of the Rooted Phylogenetic Subtree-Prune-
642 Refgraft Graph. 2015; Available: <http://arxiv.org/abs/1504.00304>
643
- 644 40. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the
645 BEAST 1.7. Mol Biol Evol. 2012;29: 1969–73. doi:10.1093/molbev/mss075

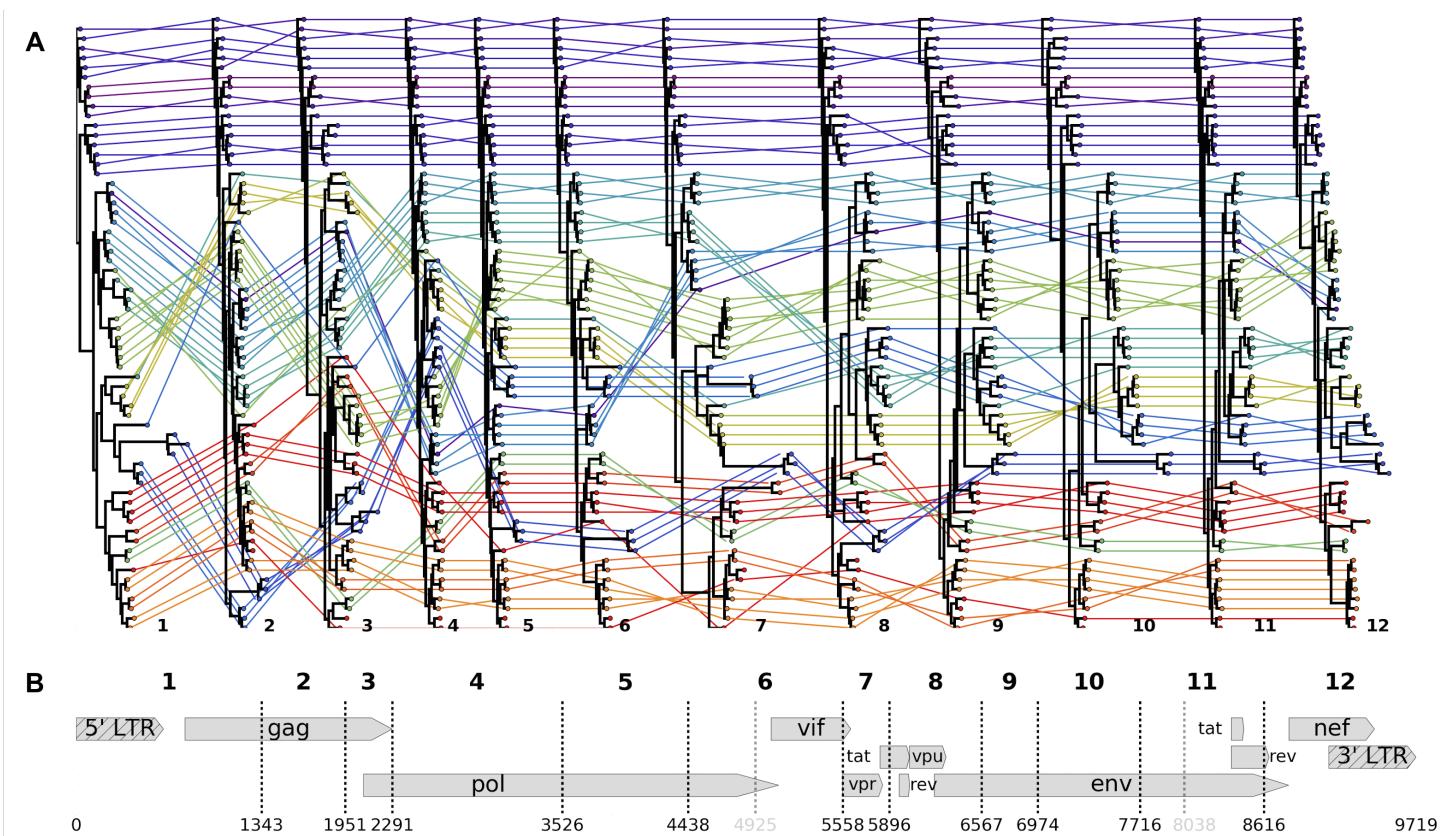


Figure 1: There have been at least 13 interlineage recombination events among SIVs.

The SIV LANL compendium, slightly modified to reduce overrepresentation of HIV, was analyzed with GARD to identify the 13 recombination breakpoints across the genome (dashed lines in **B**; numbering according to the accepted HXB2 reference genome--accession K03455, illustrated). Two of these breakpoints were omitted from further analyses because they created extremely short fragments (< 500 bases; gray dashes in **B**). For each of the 11 remaining breakpoints used in further analyses, we split the compendium alignment along these breakpoints and built a maximum likelihood tree, displayed in **A**. Each viral sequence is color-coded by host species, and its phylogenetic position is traced between trees. Heuristically, straight, horizontal colored lines indicate congruent topological positions between trees (likely not a recombinant sequence); criss-crossing colored lines indicate incongruent topological positions between trees (likely a recombinant sequence).

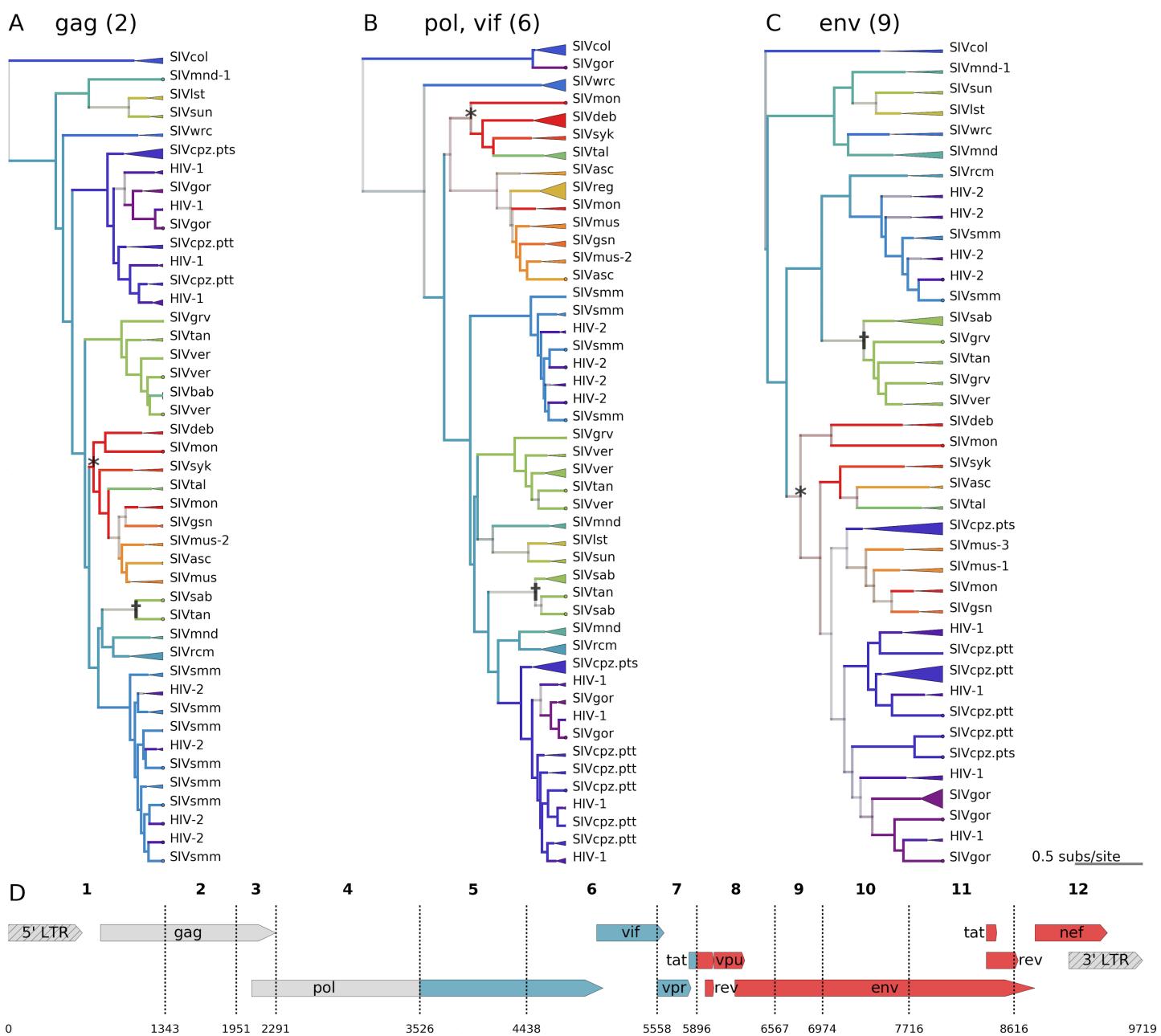


Figure 2: Cross-species transmissions are inferred from tree topologies; SIVcpz has mosaic origins.

A,B,C - Bayesian maximum clade credibility (mcc) trees are displayed for segments 2 (*gag* - A), 6 (*int* and *vif* - B), and 9 (*env* - C) of the main dataset (N=423). Tips are color coded by known host species; internal nodes and branches are colored by inferred host species, with saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences. An example of likely cross-species transmission is starred in each tree, where the host state at the internal node (red / mona monkeys) is incongruent with the descendent tips' known host state (green / talapoin monkeys), providing evidence for a transmission from mona monkeys to talapoin monkeys. Another example of cross-species transmission of a recombinant virus among African green monkeys is marked with a dagger.

D - The genome map of SIVcpz, with breakpoints used for the discrete trait analysis, is color coded and labeled by the most likely ancestral host for each segment of the genome.

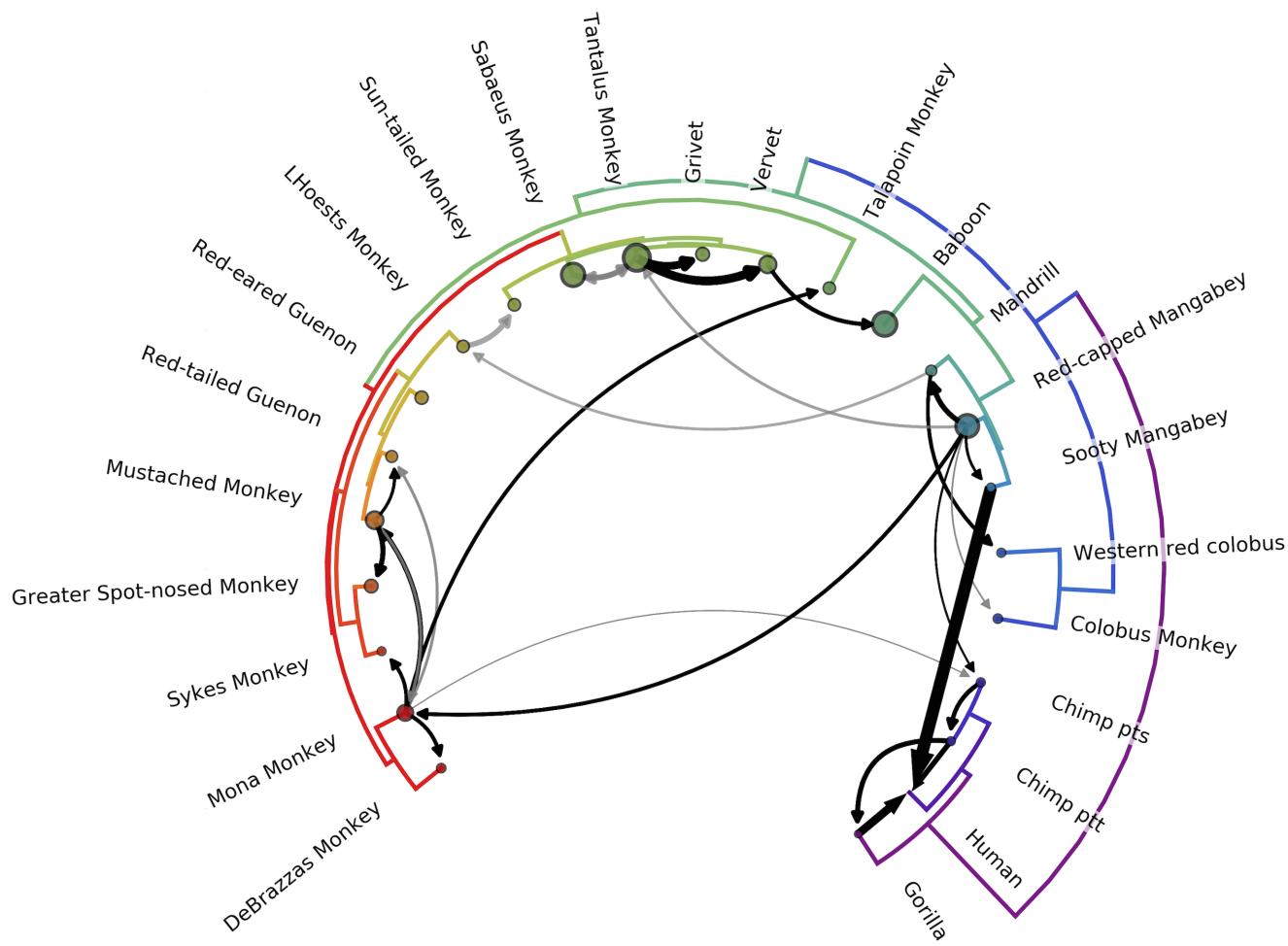


Figure 3: Most lentiviruses are the product of ancient cross-species transmissions.

The phylogeny of the host species' mitochondrial genomes forms the outer circle. Arrows represent transmission events inferred by the model with Bayes' factor ($BF \geq 3.0$); black arrows have $BF \geq 10$, with opacity of gray arrows scaled for BF between 3.0 and 10.0. Width of the arrow indicates the rate of transmission (actual rates = rates * indicators). Circle sizes represent network centrality scores for each host. Transmissions from chimps to humans; chimps to gorillas; gorillas to humans; sooty mangabeys to humans; sabaeus to tantalus; and vervets to baboons have been previously documented. To our knowledge, all other transmissions illustrated are novel identifications.

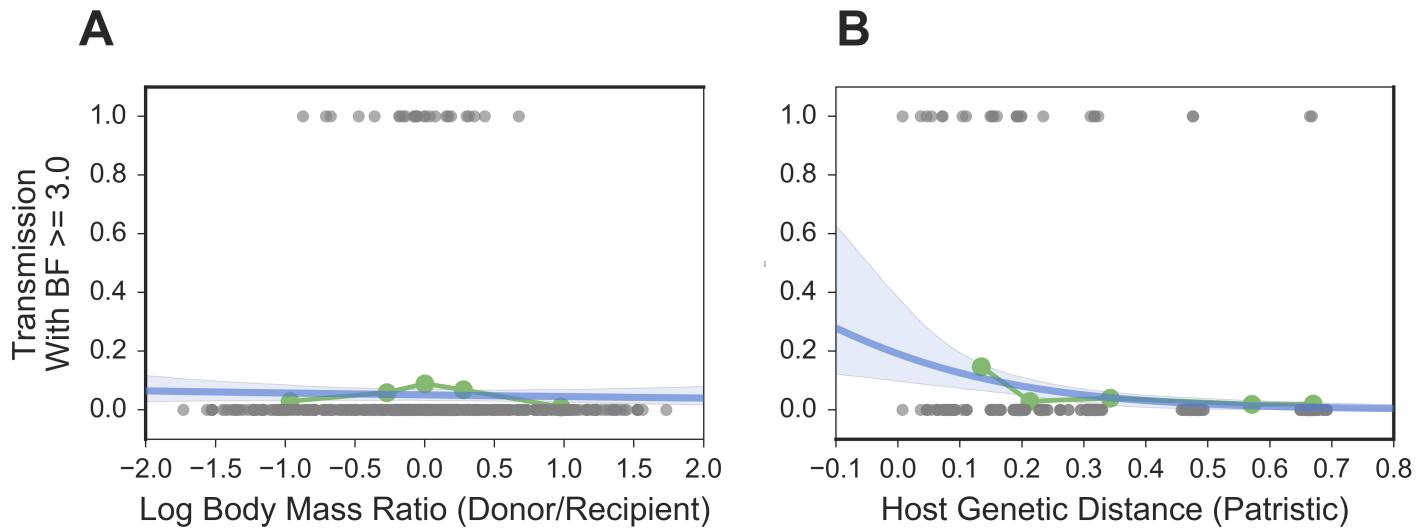


Figure 4: Cross-species transmission is driven by exposure and constrained by host genetic distance.

For each pair of host species, we (A) calculated the log ratio of their average body masses and (B) found the patristic genetic distance between them (from a maximum-likelihood tree of mtDNA). To investigate the association of these predictors with cross-species transmission, we treated transmission as a binary variable: 0 if the Bayes factor for the transmission (as inferred by the discrete traits model) was < 3.0 , and 1 for a Bayes factor ≥ 3.0 . Each plot shows raw predictor data in gray; the quintiles of the predictor data in green; and the logistic regression and 95% CI in blue.

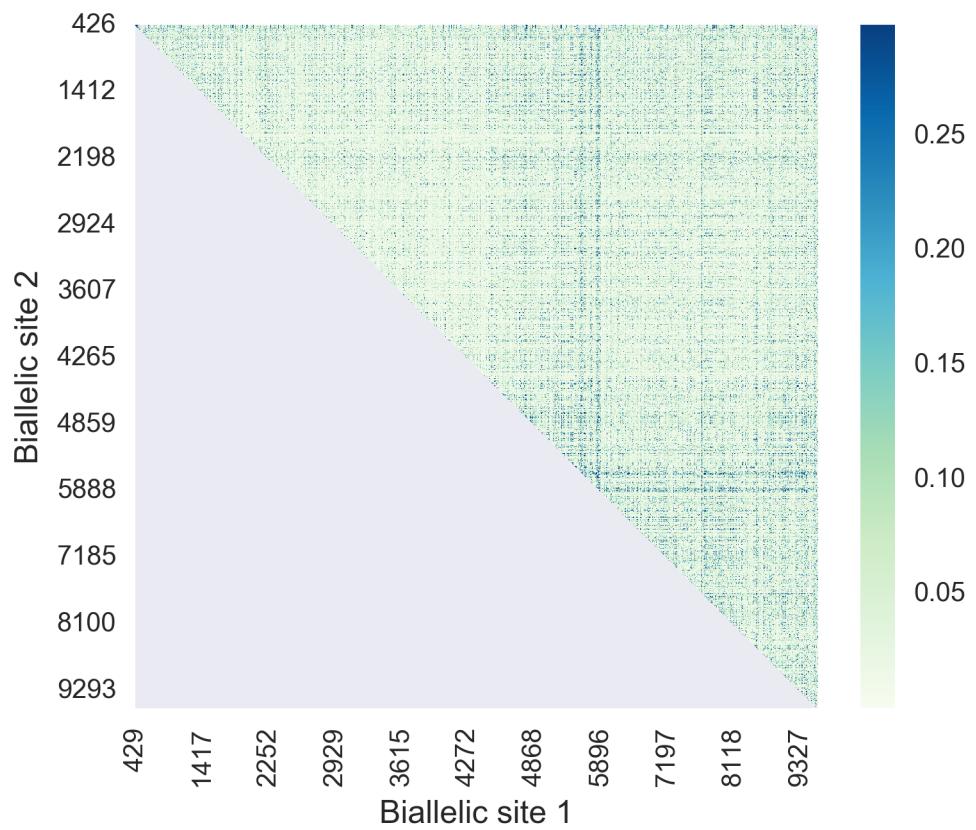


Figure S1: Extensive divergence makes sitewise measures of genetic linkage ineffective

For pairs of biallelic sites (ignoring rare variants), R^2 was used to estimate how strongly the allele in one site predicts the allele in the second site, with values of 0 indicating no linkage and 1 indicating perfect linkage. The mean value of R^2 was 0.044, indicating very low levels of linkage overall.

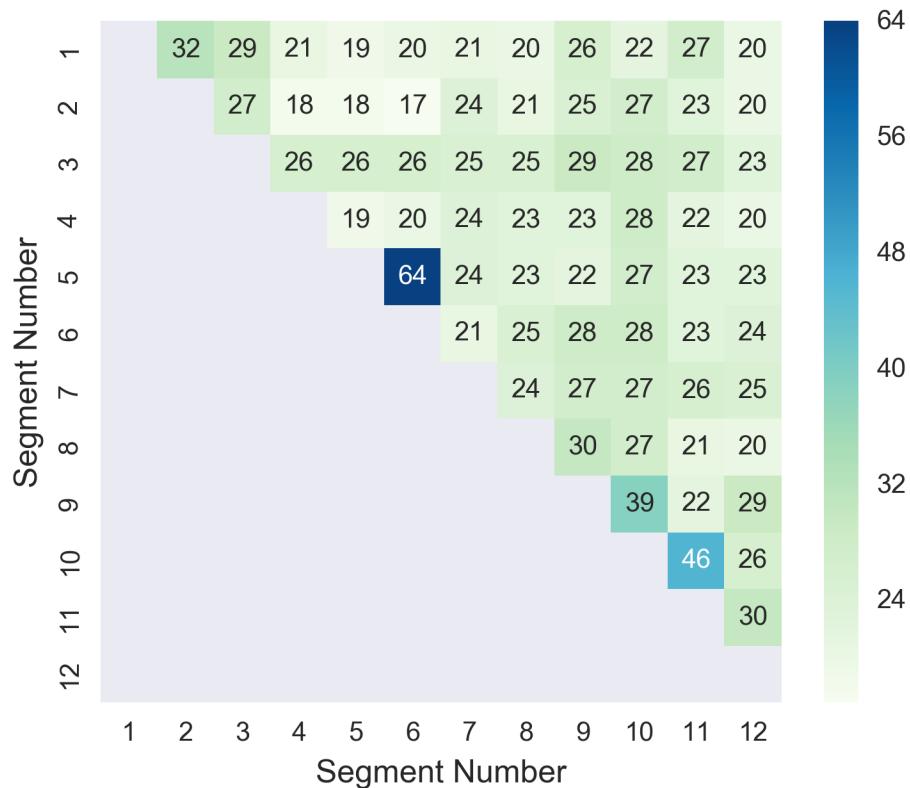


Figure S2: No evidence of linkage between nonadjacent segments of the SIV genome.

The alignment used for GARD analyses (LANL compendium with HIV overrepresentation reduced) was split along the breakpoints identified by GARD to yield the 12 genomic segments, and a maximum likelihood tree was constructed for each. The number of steps required to turn one tree topology into another was assessed for each pair of trees with the Rooted Subtree-Prune-and-Regraft (rSPR) package. Segment pairs with similar topologies have lower scores than segments with less similar topologies.

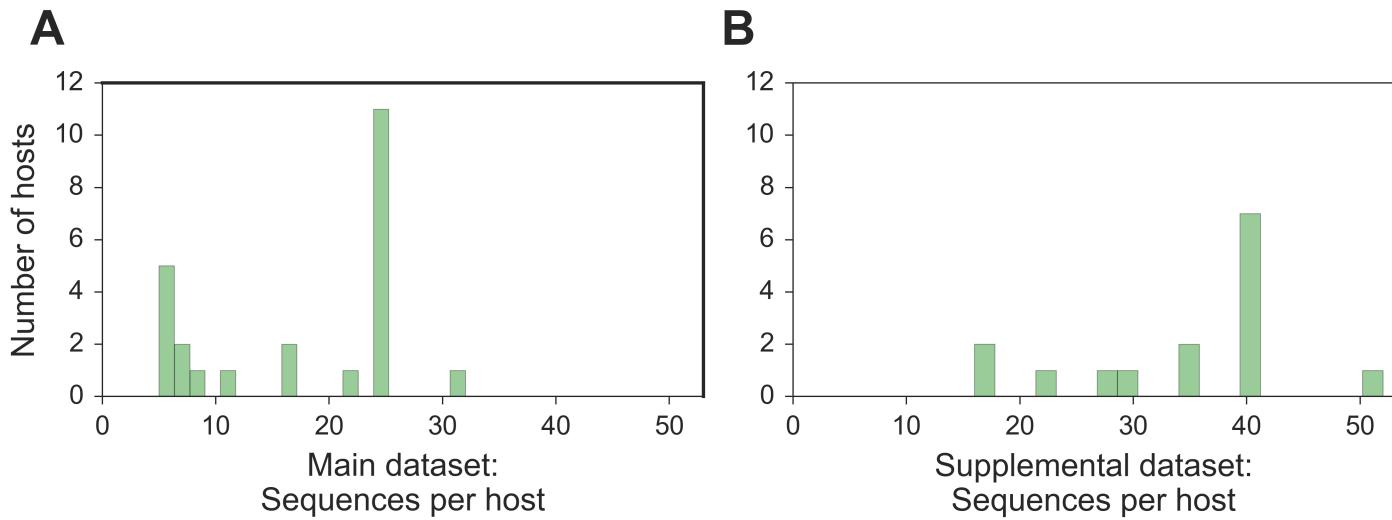


Figure S3 Distribution of the number of sequences per host included in analyses

A: All available high-quality lentivirus sequences were randomly subsampled up to 25 sequences per host for the main dataset. We included the 24 hosts with at least 5 sequences available in this dataset. B: For the supplemental dataset, we randomly subsampled up to 40 sequences per host, and included the 15 hosts with at least 16 sequences available in this dataset. For both datasets, a small number of additional sequences were permitted for the few hosts that are infected by multiple viral lineages in order to represent the full breadth of known genetic diversity of lentiviruses in each host population.

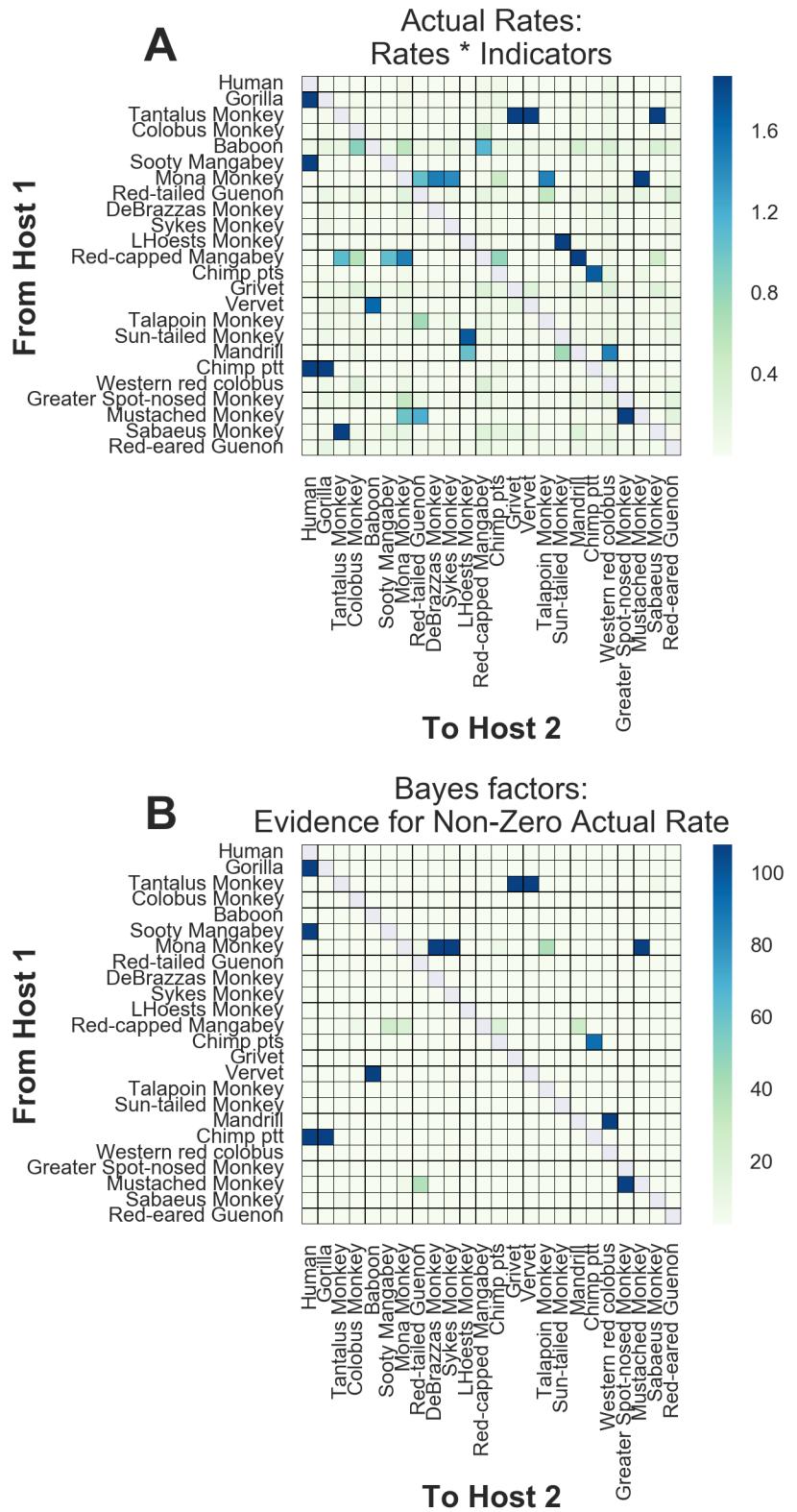


Figure S4: Actual rates and Bayes factors for main dataset discrete trait analyses

Values for the asymmetric transition rates between hosts, as estimated by the CTMC, were calculated as rate * indicator (element-wise for each state logged). We report the average posterior values above. Bayes factors represent a ratio of the posterior odds / prior odds that a given actual rate is non-zero. Because each of the 12 segments contributes to the likelihood, but they have not evolved independently, we divide all Bayes factors by 12 and report the adjusted values above (and throughout the text).

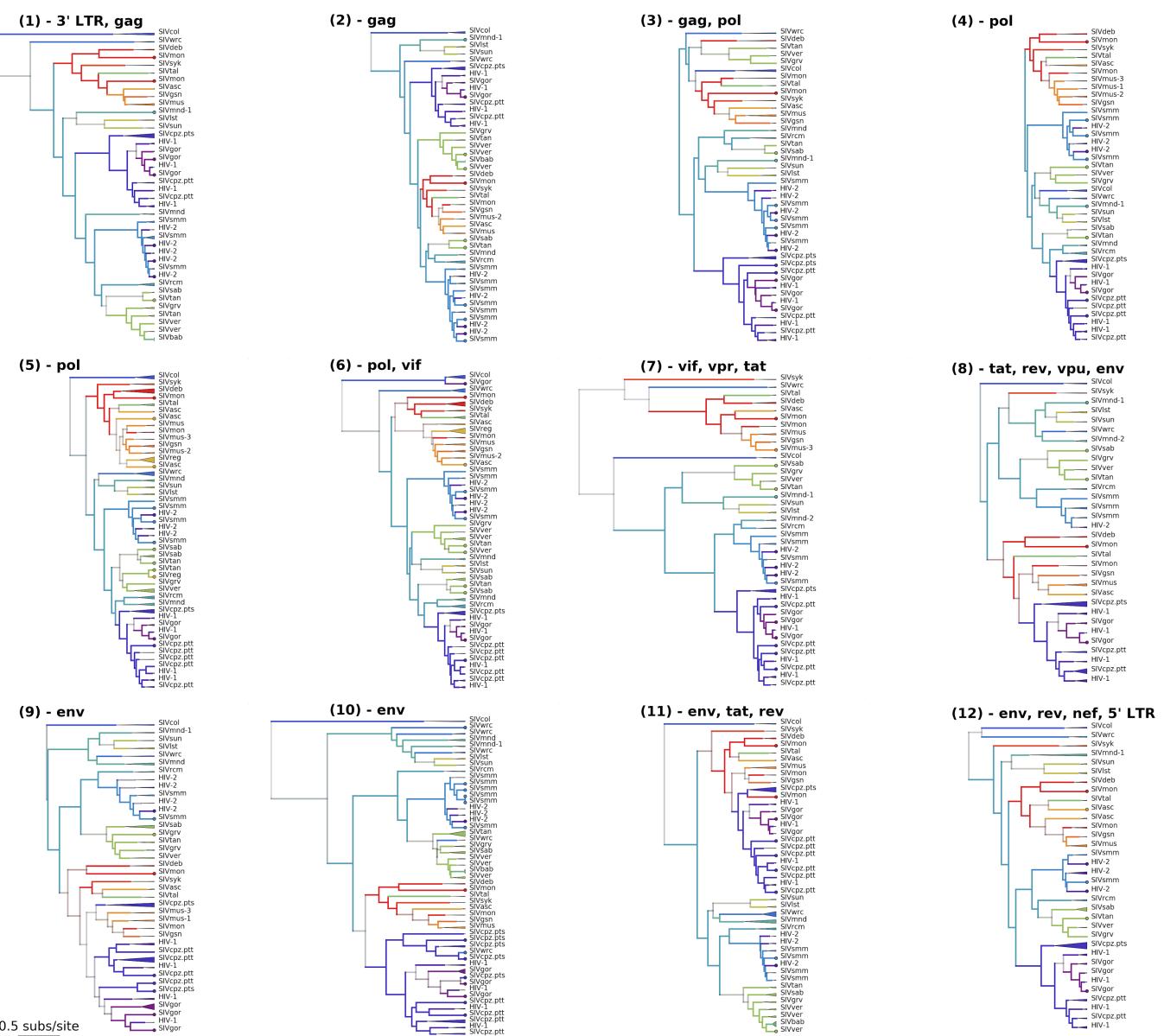


Figure S5: Maximum clade credibility trees for each of the 12 GARD-identified genomic segments of the lentiviral genome

Tips are color coded by known host state; branches and internal nodes are color coded by inferred host state, with color saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences.

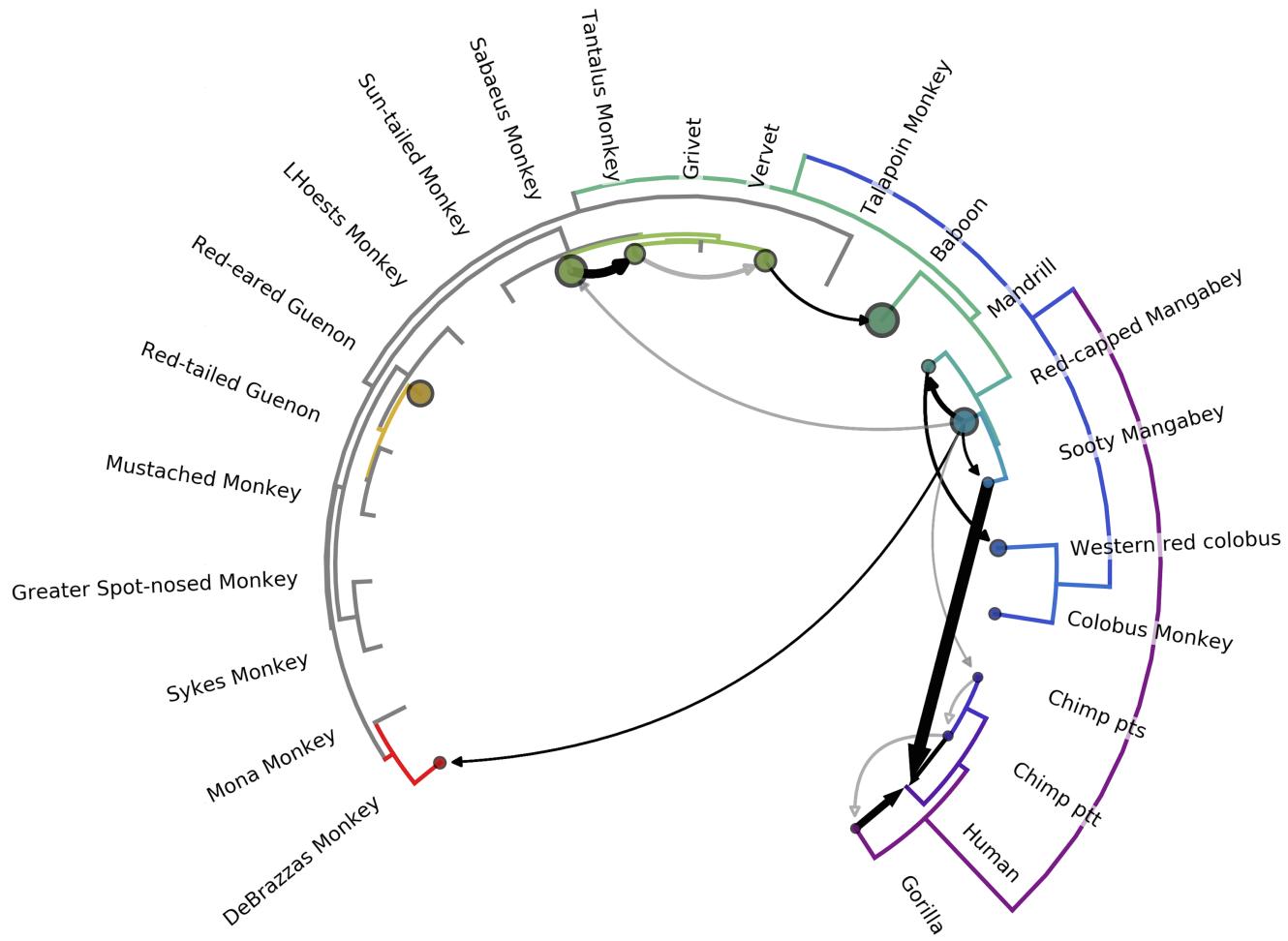
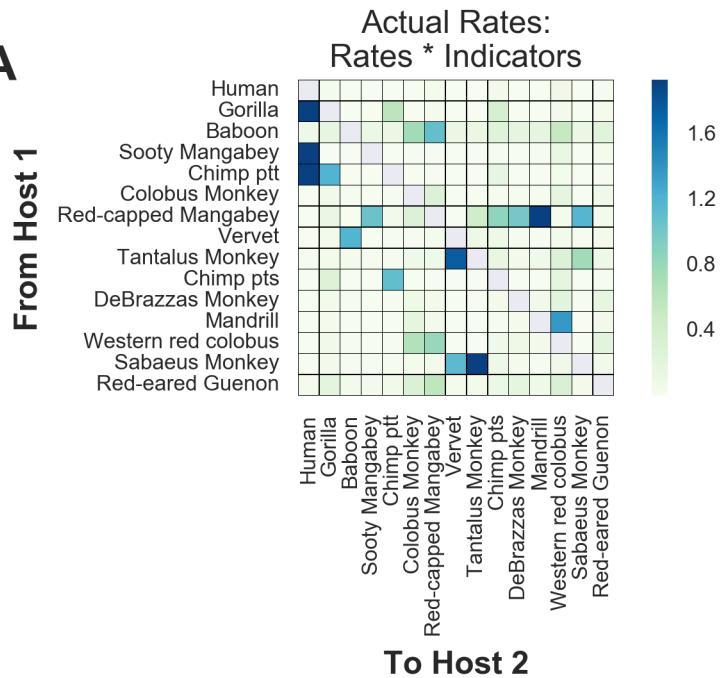


Figure S6: Most lentiviruses are the product of ancient cross-species transmissions (supplemental dataset).

The phylogeny of the host species' mitochondrial genomes forms the outer circle (gray: not included in supplemental dataset). Arrows with filled arrowheads represent transmission events inferred by the model with Bayes' factor ($BF \geq 3.0$; black arrows have $BF \geq 10$, with opacity of gray arrows scaled for BF between 3.0 and 10.0. Transmissions with $2.0 \leq BF < 3.0$ have open arrowheads. Width of the arrow indicates the rate of transmission (actual rates = rates * indicators). Circle sizes represent network centrality scores for each host. Transmissions from chimps to humans; chimps to gorillas; gorillas to humans; sooty mangabeys to humans; sabaeus to tantalus; and vervets to baboons have been previously documented. To our knowledge, all other transmissions illustrated are novel identifications.

A



B

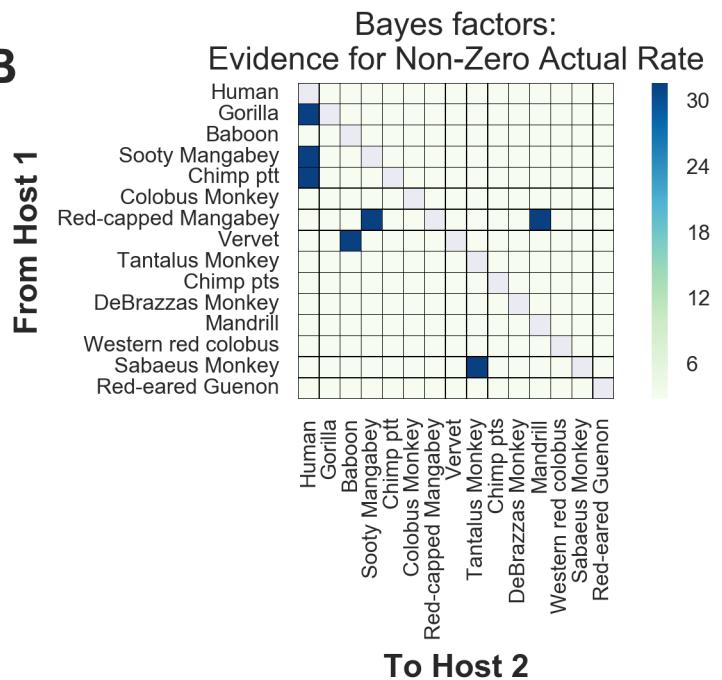


Figure S7: Actual rates and Bayes factors for supplemental dataset discrete trait analyses

Values for the asymmetric transition rates between hosts, as estimated by the CTMC, were calculated as rate * indicator (element-wise for each state logged). We report the average posterior values above. Bayes factors represent a ratio of the posterior odds / prior odds that a given actual rate is non-zero. Because each of the 12 segments contributes to the likelihood, but they have not evolved independently, we divide all Bayes factors by 12 and report the adjusted values above (and throughout the text).

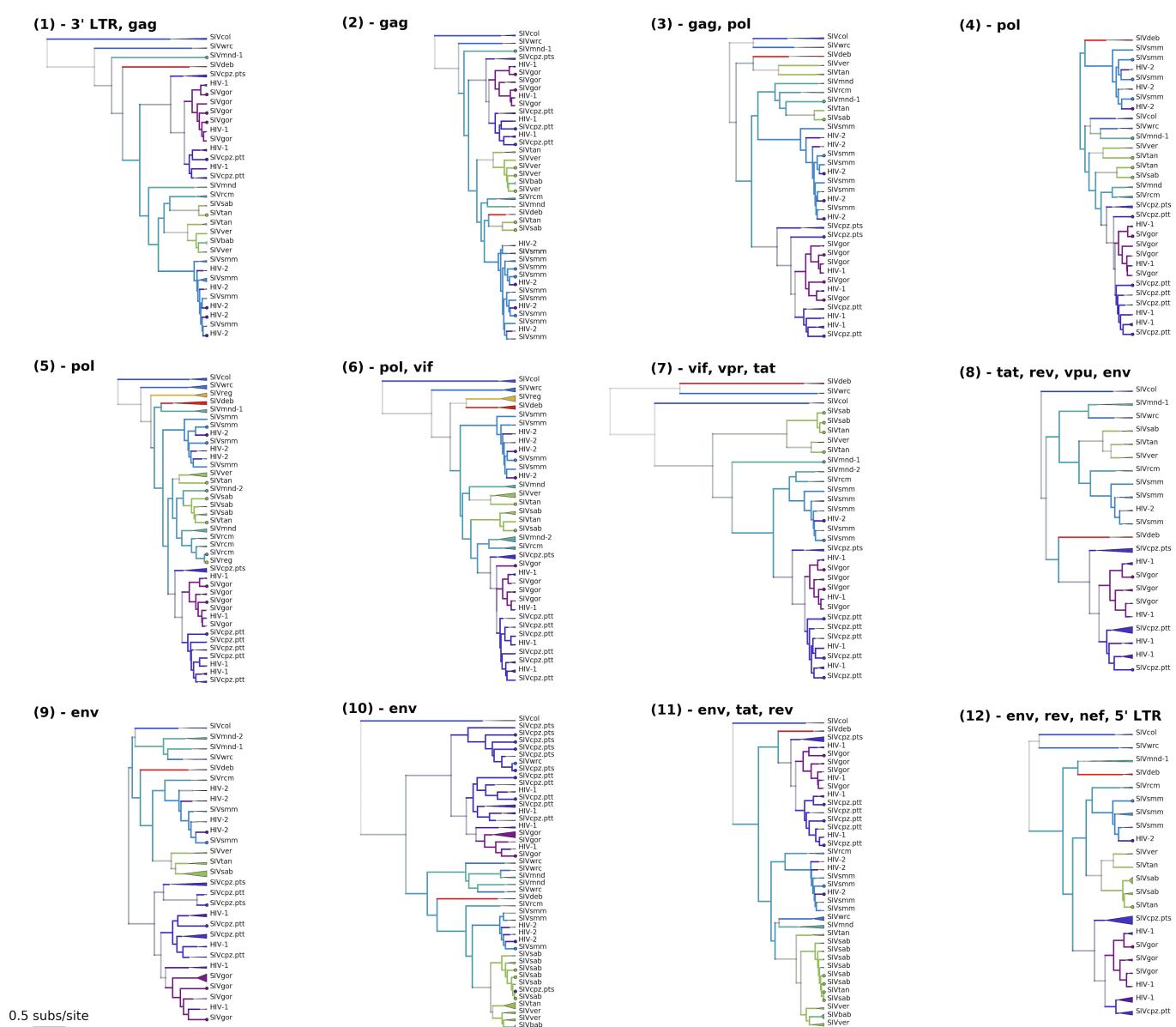


Figure S8: Maximum clade credibility trees for each of the 12 GARD-identified genomic segments of the lentiviral genome (supplemental dataset)

Tips are color coded by known host state; branches and internal nodes are color coded by inferred host state, with color saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences.

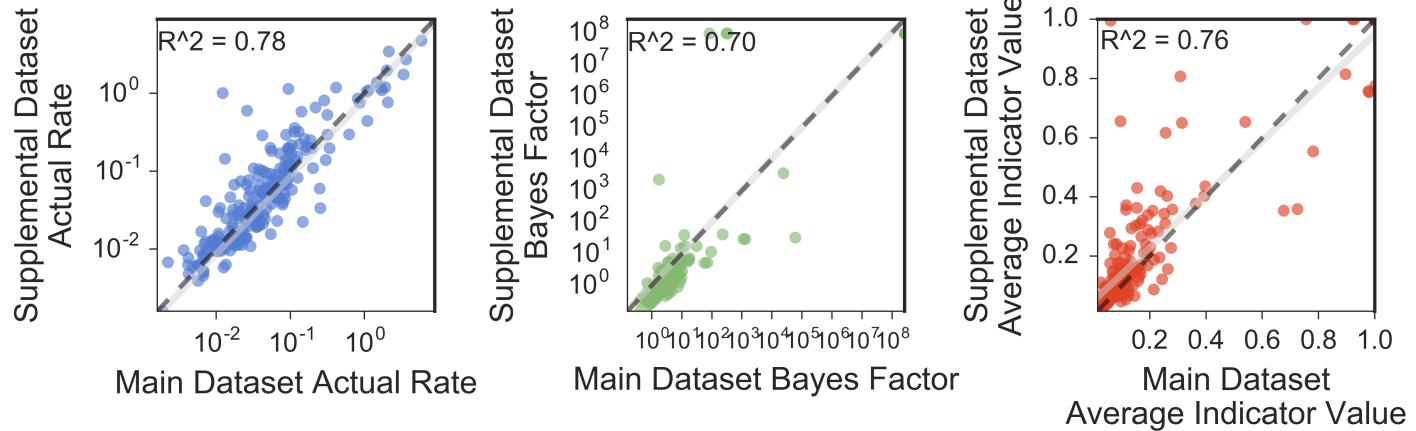


Figure S9: Comparison of Main and Supplemental Dataset Discrete Trait Analysis Results

Each datapoint represents one of the 210 possible transmissions between each pair of the 15 hosts present in both datasets. The black dashed line shows $y=x$; the linear regression and 95% CI are shown in gray.