

1   **Inference of direction, diversity, and frequency of HIV-1 transmission using**  
2   **approximate Bayesian computation**

3

4

5   Ethan O. Romero-Severson<sup>1</sup>, Ingo Bulla<sup>1,2</sup>, Nick Hengartner<sup>1</sup>, Inês Bárto<sup>3</sup>, Ana  
6   Abecasis<sup>4</sup>, José M. Azevedo-Pereira<sup>5</sup>, Nuno Taveira<sup>3,6</sup>, Thomas Leitner<sup>1</sup>

7

8   <sup>1</sup> Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los  
9   Alamos, NM 87545, USA

10   <sup>2</sup> Institut für Mathematik und Informatik, Universität Greifswald, Walther-Rathenau-  
11   Straße 47, 17487 Greifswald, Germany

12   <sup>3</sup> HIV Evolution, Epidemiology and Prevention, Research Institute for Medicines  
13   /Instituto de Investigação do Medicamento (iMed.ULisboa), Faculdade de Farmácia,  
14   Universidade de Lisboa, Portugal

15   <sup>4</sup> Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina  
16   Tropical, IHMT, Universidade Nova de Lisboa, UNL, Rua da Junqueira 100, 1349-008  
17   Lisboa, Portugal

18   <sup>5</sup> Host-Pathogen Interaction Unit, Research Institute for Medicines /Instituto de  
19   Investiga<sup>ç</sup>o<sup>n</sup> do Medicamento (iMed.ULisboa), Faculdade de Farmácia, Universidade  
20   de Lisboa, Portugal

21   <sup>6</sup> Centro de Investigação Interdisciplinar Egas Moniz (CiiEM), Instituto Superior  
22   Ciências da Saúde Egas Moniz, Monte de Caparica, Portugal

23

24

25   GenBank Accession numbers: KT123041-KT123171

26

1    ***Short running title***

2    ABC inference of HIV-1 transmission

3

4    ***Keywords***

5    Coalescent, phylogeny, approximate Bayesian computation, co-infection, super-

6    infection

7

8    ***Corresponding author***

9    Thomas Leitner

10    Theoretical Biology & Biophysics Group T-6

11    Mail Stop K710

12    Los Alamos National Laboratory

13    Los Alamos, NM 87545

14    USA

15

16    Tel: (505) 667-3898

17    Email: [tkl@lanl.gov](mailto:tkl@lanl.gov)

18

19

20

1 **ABSTRACT**

2

3 Diversity of the founding population of Human Immunodeficiency Virus Type 1  
4 (HIV-1) transmissions raises many important biological, clinical, and  
5 epidemiological issues. In up to 40% of sexual infections there is clear evidence for  
6 multiple founding variants, which can influence the efficacy of putative prevention  
7 methods and the reconstruction of epidemiologic histories. To measure the diversity  
8 of the founding population and to compute the probability of alternative  
9 transmission scenarios, while explicitly taking phylogenetic uncertainty into  
10 account, we created an Approximate Bayesian Computation (ABC) method based on  
11 a set of statistics measuring phylogenetic topology, branch lengths, and genetic  
12 diversity. We applied our method to a heterosexual transmission pair showing a  
13 complex paraphyletic-polyphyletic donor-recipient phylogenetic topology. We  
14 found evidence identifying the donor that was consistent with the known facts of  
15 the case (Bayes factor >20). We also found that while the evidence for ongoing  
16 transmission between the pair was as good or better than the singular transmission  
17 event model, it was only viable when the rate of ongoing transmission was  
18 implausibly high (~1/day). We concluded that the singular transmission model,  
19 which was able to estimate the diversity of the founding population (mean 7%  
20 substitutions/site), was more biologically plausible. Our study provides a formal  
21 inference framework to investigate HIV-1 direction, diversity, and frequency of  
22 transmission. The ability to measure the diversity of founding populations in both  
23 simple and complex transmission situations is essential to understanding the  
24 relationship between the phylogeny and epidemiology of HIV-1 as well as in efforts  
25 developing new prevention technologies.

26

1 **INTRODUCTION**

2

3 Most HIV-1 infections are the result of sexual transmission (SHATTOCK AND MOORE  
4 2003), where 20-40% involve transmission of multiple genetic variants (KEELE *et al.*  
5 2008; SALAZAR-GONZALEZ *et al.* 2009; LI *et al.* 2010; RIEDER *et al.* 2011). Transmitting  
6 more than one variant raises many important biological, clinical, and  
7 epidemiological issues. Biologically, successful transmission of >1 variant means  
8 that many viruses in a donor have the capacity to establish infection, and further  
9 that they had similar fitness as they did not outcompete each other in the new host.  
10 Following establishment of infection, the existence of multiple lineages may also  
11 generate virus with higher relative fitness than when single lineages establish  
12 infection (CARRILLO *et al.* 2007), due either to recombination or competition after  
13 transmission (SANBORN *et al.* 2015). Clinically, transmission of several virus variants  
14 may make it harder for the immune system to combat the virus (GROBLER *et al.* 2004;  
15 YANG *et al.* 2005; SMITH *et al.* 2006), easier for the virus to evade antiviral treatment  
16 (SMITH *et al.* 2004), and may accelerate disease progression (GOTTLIEB *et al.* 2004).  
17 Epidemiologically, the establishment of >1 genetic variant can occur simultaneously  
18 at one time or sequentially over a long period of time, which is defined as co-  
19 infection or super-infection, respectively (VAN DER KUYL AND CORNELISSEN 2007). This  
20 has further impact on whether one infection protects against another (ALTFELD *et al.*  
21 2002; RONEN *et al.* 2013), or if later super-infections may induce drug resistance  
22 (SMITH *et al.* 2005), and if a potential vaccine to one form would protect against  
23 another.

24

25 Phylogenetics reconstructs evolutionary history, and for an organism like HIV-1 that  
26 evolves very rapidly, the joint pathogen phylogeny from hosts that have infected  
27 each other reveals details about the host-to-host transmission. Recently, coalescent-  
28 based simulations showed that the resulting phylogeny may reveal both direction  
29 and directness in epidemiologically linked hosts, i.e., who infected whom, and  
30 whether missing host-links were likely (ROMERO-SEVERSON *et al.* 2016). Furthermore,  
31 it has previously been shown that there exists a pretransmission interval that  
32 describes the bias towards the past when using phylogenetic trees to estimate  
33 transmission times (LEITNER AND ALBERT 1999; LEITNER AND FITCH 1999; ROMERO-  
34 SEVERSON *et al.* 2014). Importantly, when multiple phylogenetic lineages have been  
35 transmitted from one host to another the resulting tree opens up alternative  
36 interpretations of whether all lineages were transmitted at one or several occasions.  
37 Thus, while simulations have shown that phylogenies carry detailed information  
38 about who infected whom, and within-host models predict the pretransmission  
39 interval, a single framework to determine the evidence for the various possible  
40 transmission scenarios between two infected hosts is lacking.

41

42 The objective of this study was to create a unified framework to investigate the  
43 nature of an epidemiological link and to apply that to a real HIV-1 transmission case.  
44 Based on previous theoretical work the tree topology should probabilistically  
45 indicate direction and directness, whether >1 lineage were transmitted, as well as  
46 when transmission occurred. Here, we also intended to determine the evidence for

1 whether the infection was established by a single transmission event or an ongoing  
2 process of re-infection. In addition, we wanted to avoid basing our inferences on a  
3 single (best) phylogenetic tree as many trees with different topology and distance  
4 properties may be nearly as likely as the best tree. Basing our method on the entire  
5 posterior distribution of trees allows us to consider the full range of solutions that  
6 the data may support and to propagate uncertainty in phylogenetic reconstruction  
7 onto the parameter estimates. Thus, we extended our previous within-host  
8 coalescent methods to simulate trees corresponding to different transmission  
9 scenarios and parameterizations and analyzed a previously unpublished HIV-1  
10 transmission chain. To test and compare alternative scenarios of the  
11 epidemiological link, i.e., when and how transmission(s) occurred, we developed  
12 and applied an approximate Bayesian computation (ABC) method based on tree  
13 topology, root host-assignment, and patristic tree distance measures. The ABC  
14 method also allowed us to estimate the diversity at the time of transmission rather  
15 than at time of sampling.

16

17

## 18 MATERIALS AND METHODS

19

### 20 Joint linear within-hosts population model

21 We considered two alternative sexual transmission scenarios: i) a singular  
22 transmission event, or ii) multiple transmissions where the donor and recipient are  
23 repeatedly re-infecting each other (Fig 1). In the singular transmission scenario the  
24 within-host effective population size,  $N(t) = \alpha + \beta t$ , is a linear function of time  
25 where  $\alpha$  is the population size at the time of infection at time  $t = 0$  and  $\beta$  is the  
26 linear increase in population size per day. Expanding this model to a transmission  
27 pair, we assume that all times and parameters are defined along a single forward  
28 time axis such that the population size in the donor is simply given by  $N_d(t) = \alpha_d +$   
29  $\beta_d t$ , while the population size in the recipient is given by  $N_r(t) = \alpha_r + \beta_r(t -$   
30  $t_{trans}$ ), where subscript  $d$  indicates the donor and subscript  $r$  indicates the recipient.  
31 The time of transmission is indicated as  $t_{trans}$  when the population size is  $N_d(t_{trans})$   
32 in the donor and  $\alpha_r$  in the recipient.

33

34 In the multiple transmission scenario we assume that single virus lineages are  
35 passed via sexual contact to the female partner at rate  $\rho$  and to the male partner at  
36 rate  $\rho/2$ . The half factor corresponds to the reduced rate of female to male  
37 transmission (BOILY *et al.* 2009). The population sizes are given by the same  
38 equations as in the singular transmission scenario, but where  $\alpha_r = \alpha_d = 1$ . We also  
39 assume that  $\rho$  is small enough that  $N(t)$  is not significantly affected by the migration  
40 of lineages between the donor and recipient. We assume that all extant lineages are  
41 equally probable to migrate.

42

### 43 Simulating trees from the joint coalescent model

44 The derivation of the density of coalescent times for a sample of  $k$  clones follows  
45 Romero-Severson *et al.* (ROMERO-SEVERSON *et al.* 2014) with modifications to account

1 for the increased rate of coalescence in a population with  $k$  extant lineages. As  
2 before, we define the rate of change in coalescent time as a function of calendar time  
3 along the reverse time axis,  $s$ , as  
4

$$5 g(s, t_1) = \int_0^s \frac{du}{\alpha + \beta(t_1 - u)} = \beta^{-1} [\log(\alpha + \beta t_1) - \log(\alpha + \beta(t_1 - s))]$$

6  
7 where  $t_1$  is the current time. The density of the time to the next coalescent event in  
8 Kingman's n-coalescent with normalized population size for  $k$  extant lineages is  
9  $f_A(a) = \binom{k}{2} e^{-a} \binom{k}{2}^{(k)}$  (WAKELEY 2009). Therefore we obtained the density of the time to  
10 coalesce in our linear growth model by the transformations  
11

$$12 f_Z(z) = f_A(g(z, t_1)) g'(z, t_1) = \binom{k}{2} (\alpha + \beta t_1)^{-\binom{k}{2} \frac{1}{\beta}} (\alpha + \beta(t_1 - z))^{\binom{k}{2} \frac{1}{\beta} - 1}$$

13 for  $z \in [0, t_1 + \frac{\alpha}{\beta}]$ .  
14

15 To simulate the time to the next coalescent event we use the inverse cumulative  
16 function

$$17 F_Z^{-1}(u) = \left( 1 - (1 - u)^{\frac{\beta}{\binom{k}{2}}} \right) (\alpha + \beta t_1) \beta^{-1}$$

18 and to simulate the time to the next migration event using the inverse cumulative  
19 function  
20

$$21 F_M^{-1}(u) = \left( 1 - (1 - u)^{\frac{\beta}{k\rho}} \right) (\alpha + \beta t_1) \beta^{-1}$$

22 where  $u$  is a unit uniform random variate. In the singular transmission model a  
23 coalescent process was simulated in each of the derived populations of the donor  
24 and recipient up to the time of transmission. We define a derived population as a  
25 population that only exists in one host after transmission has occurred (in forward  
26 time). The derived populations join into a source population at time of transmission,  
27 when the lineages from both hosts can freely coalesce (Fig 2). In the source  
28 population, a coalescent process was simulated starting with the previous  
29 simulations of the derived populations. In the ongoing transmission model four  
30 possible events can occur: migration from donor to recipient, migration from  
31 recipient to donor, coalescence in donor, and coalescence in the recipient. At a  
32 migration event one random lineage moves from one host to the other. Simulations  
33 stop when the infection time of the donor is reached along the reverse time axis.  
34

### 35 36 **Model priors and constraints**

37 To ease interpretation of  $\alpha$ , we assume that  $\alpha$  only takes integer values  $\geq 1$ . Because  
38 the model is constructed assuming that all parameters and population sizes are  
39 continuous it is theoretically possible that the number of lineages that survive

1 though the transmission bottleneck can exceed  $\alpha$  (e.g. the probability of 5 lineages  
2 surviving a bottleneck of size 4 is extremely small but formally non-zero). This  
3 incongruity virtually never occurs, however to avoid this situation, we forced a  
4 coalescence with branch lengths zero in any case where the number of extant  
5 lineages exceeds  $N(t)$ . We also assume that the donor was infected with a single  
6 lineage,  $\alpha_d = 1$ .  
7

8 To constrain the linear growth rate in the recipient, we assumed that the ratio of the  
9 population sizes in the donor and recipient is equal to the empirically observed ratio  
10 of pairwise diversity between the donor and recipient.  
11

12 To match the data we wanted to analyze here, we assume that at the time of  
13 sampling both donor and recipient were treatment naïve and did not have an AIDS  
14 diagnosis. Based on that, and a lack of other relevant epidemiological information,  
15 we assumed a uniform distribution of infection times from 0 to 12 years. We assume  
16 that the population growth rate in the donor is drawn from  $\beta_d \sim \text{Exponential}(15^{-1})$   
17 units per day. This distribution includes growth rates that correspond to most of the  
18 published estimates of the HIV within-host effective population numbers (LEIGH  
19 BROWN 1997; NIJHUIS *et al.* 1998; PENNINGS *et al.* 2014). In the case of a singular  
20 transmission event we assume that the donor transmits on average 0.7% (s.d. 0.9%)  
21 of the current effective population number (Beta(0.5,70) distributed) to the  
22 recipient. In the ongoing transmission case we assume the transmission rate from  
23 the male to female partner is a uniform random variable between 0 and 2 per day,  
24  $\rho \sim \text{Uniform}(0,2)$ .  
25

## 26 **Phylogenetic measures for approximate Bayesian computation**

27 For a tree with taxa from two hosts, "A" and "B", we used the following statistics to  
28 define the probability that a simulation should be accepted: the root label, the  
29 topological class, the number of monophyletic clades of one of the host labels, the  
30 total number of substitutions in the tree, and the average pairwise distance between  
31 pairs of taxa with mismatched host labels. The root label is defined as the maximum  
32 parsimony host assignment of the root ("A", "B", or ambiguous). The topological  
33 relationship can be one out of three classes: MM (both host sets of taxa are  
34 monophyletic), PM (taxa from one host forms a monophyletic clade that inserts into  
35 the sample of the other host forming a paraphyletic clade), and PP (taxa from one  
36 host are paraphyletic to the other host's taxa that are polyphyletic, or both host's  
37 taxa are polyphyletic). Root label and topological class have been demonstrated to  
38 be associated with the epidemiologic relationship between two sampled hosts  
39 (ROMERO-SEVERSON *et al.* 2016). The number of monophyletic clades of the putative  
40 recipient in the joint tree defines the minimum number of transmitted lineages.  
41 Note that it probabilistically informs the number of transmitted lineages, e.g. a large  
42 number of transmitted lineages is generally—but not always—inconsistent with an  
43 observed single monophyletic clade. With the root label assigned to "A", the number  
44 of "B" clades in the sample is counted by applying Dollo's law (DOLLO 1893), which  
45 logically follows from the irreversible fact that the donor was infected before the  
46 recipient. In principle, this translates on the tree to first assigning the "A" label to

1 each node on a root to “A”-tip path, and then counting the minimum “A” to “B”  
2 transformations needed to observe the tip labels. We call each such “B” lineage a  
3 “monophyletic clade”, including clades with only one “B” taxon.

4  
5 In the single transmission of multiple lineages scenario, rescaling the tree using a  
6 molecular clock identifies the time interval during which transmission could have  
7 occurred. In that case the first coalescence going towards the root between a “A” and  
8 “B” lineage defines the time of when the tree describes the HIV-1 evolution in the  
9 donor, i.e., the “source population” (ROMERO-SEVERSON *et al.* 2016). Thus, the time  
10 during which transmission could have occurred spans from the time of the sampling  
11 of the recipient back until the time that defines the source population (Fig 2). The  
12 total number of substitutions is calculated by assuming a Gamma distributed  
13 uncorrelated relaxed molecular clock with a mean evolutionary rate at 6.7 (s.d. 4.2)  
14  $\times 10^{-3}$  substitutions/site per year that informs both the infection time and the  
15 within-host population growth rate (LEITNER AND ALBERT 1999). Finally, the mean  
16 number of substitutions between donor-recipient pairs informs the donor-recipient  
17 transmission time and is linked to the number of transmitted lineages. Note that the  
18 possible transmission time interval in the multiple transmissions of single lineages  
19 scenario is undefined because it does not have a strict logical boundary going  
20 towards the root. Hence, we do not use that model to estimate the time of the  
21 original transmission.

22

### 23 **Sampling from the posterior density of parameters**

24 To account for variance in the phylogenetic trees that are consistent with the  
25 sequence data, we calculated the statistics described above for each tree and  
26 normalized the results to obtain distributions of the statistics conditional on the  
27 data and phylogenetic model. We considered two separate probabilistic sampling  
28 schemes based on either the topological statistical alone—without having to specify  
29 the evolutionary rate—or the full set of tree statistics. Both schemes consider each  
30 statistic as an independent test that is probabilistically passed with the empirical  
31 probability of the statistic. For example, if the empirical probability of the PP  
32 topology is 1.0 (every tree in the posterior had a PP topology) then any simulation  
33 that does not produce a PP tree is rejected. Likewise, if the probability of the “A”  
34 root label is 0.93, then a simulation with root label “A” will be accepted 93% of the  
35 time. Each statistic is considered independent such that the total probability of  
36 accepting a parameter is proportional to the product of each of the simulated  
37 statistics.

38

39 The first sampling scheme is based only on the topological class and root label  
40 statistics, while the second sampling scheme is based on the topology, root label,  
41 number of monophyletic clades in “A”, the total number of substitutions in the joint  
42 tree, and the average distance (in substitutions) between “A”-“B” pairs.

43

44 We sampled  $10^7$  parameter sets from the prior, for each of the 4 possible models (“A”  
45 donor, singular transmission; “A” donor, multiple transmissions; “B” donor, singular  
46 transmission; and “B” donor, multiple transmissions). We considered the ratio of the

1 number of accepted parameters as an approximation of a Bayes factor for the model.  
2 We took the marginal means of the accepted samples to be the point estimate for the  
3 model parameters and the appropriate quantiles to define the 95% credible  
4 intervals.  
5

## 6 **Study subjects**

7 To test our model and ABC framework, we analyzed a set of sequences from three  
8 HIV-1 infected subjects (MP1, MP2, MP3) that had been part of a forensic HIV  
9 transmission investigation, where male MP2 had accused female MP1 of intentional  
10 transmission, and where MP2 infected MP3, a later female sexual partner. MP1 and  
11 MP2 subjects had a history of intravenous drug use, but MP3 did not. Thus, based on  
12 the epidemiological record, MP1 and MP2 could potentially have infected each other  
13 via either sexual contact or needle injection, but transmission between MP2 and  
14 MP3 could only have been through sexual interaction. Based on maximum likelihood  
15 (ML) phylogenetic reconstruction of HIV-1 *env* DNA sequences, MP1 taxa were  
16 separated from MP2 taxa by multiple local control and database sequences (Fig S1).  
17 Hence, MP1 was highly unlikely to have infected MP2 or MP3. However, the  
18 phylogenetic reconstruction was consistent with HIV-1 transmission from MP2 to  
19 MP3. The criminal investigation concluded that MP1 had not infected MP2, in part  
20 based on the phylogenetic evidence (Fig S1). That investigation used the case  
21 sequences in this paper plus 119 *env* sequences selected from Portuguese  
22 and publicly available databases. In the general framework above, MP2 can be seen  
23 as host "A" and MP3 as host "B".  
24

## 25 **DNA sequencing**

26 Chromosomal DNA was extracted from infected PBMC's of each subject using  
27 Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer  
28 recommendations. Nested PCR was done to obtain a 534 bp fragment from the C2V3  
29 *env* region (HXB2 positions 6858-7392). Thermal cycling conditions were as  
30 previously described (BARTOLO *et al.* 2009). PCR products were cloned into the  
31 pCR™4-TOPO vector (Invitrogen), using the TOPO TA Cloning Kit (Invitrogen)  
32 according to the manufacturer's instructions. DNA sequencing was performed using  
33 the BigDye Terminator V3.1 Cycle sequencing Kit (Applied Biosystems) and an  
34 automated sequencer (3100-Avant Genetic Analyzer; Applied Biosystems, Foster  
35 City, CA). We derived 31, 20, and 19 sequences from MP1, MP2, and MP3,  
36 respectively.  
37

## 38 **Phylogenetic reconstruction**

39 HIV-1 sequences were aligned using MAFFT with the L-INS-i algorithm (KATOH AND  
40 TOH 2008). Maximum likelihood phylogenetic trees were inferred using PhyML  
41 (GUINDON *et al.* 2005) under a GTR+I+G substitution model, 4 categories Gamma  
42 optimization, with a Bio-NJ starting tree and best of NNI and SPR search, and aLRT  
43 SH-like branch support (ANISIMOVA AND GASCUEL 2006). The posterior distribution of  
44 trees was sampled using MrBayes (RONQUIST AND HUELSENBECK 2003) under the same  
45 model parameterization as the PhyML trees. Two Markov chains were run for 20  
46 million steps each. Removing 25% of the chain as burn-in, combining the chains, and

1 sampling every thousandth tree, we obtained 30,000 independent trees from the  
2 posterior distribution of trees.

3

4 **Data availability**

5 Sequences have been deposited in GenBank under accession numbers KT123041-  
6 KT123171.

7

8

9

10 **RESULTS**

11

12 **Direct transmission of multiple, diverse phylogenetic lineages**

13 Using the MP1 population as outgroup, the inferred rooted ML tree suggested that  
14 MP3 was infected with at least 7 independent phylogenetic lineages from MP2 (Fig  
15 3). The ML tree thus indicated a paraphyletic-polyplyletic (PP) topological  
16 relationship (ROMERO-SEVERSON *et al.* 2016), strongly suggesting direct transmission  
17 from MP2 to MP3.

18

19 However, the ML tree does not give any sense of the variance in the topological  
20 signal that suggests that MP2 is the likely donor. In the posterior distribution of  
21 trees (obtained using MrBayes) we found that 93% of the trees had a PP topology  
22 with an MP2 root label. Likewise, the mean number of monophyletic clades in the  
23 posterior (7.7) was close to the number of monophyletic clades in the ML tree (7),  
24 however, the range in the posterior was quite large (4-14, Fig 4A). Note that the  
25 bottom 4 monophyletic clades in the ML tree (Fig 3) are only very weakly separated  
26 considering branch lengths, and thus it is no surprise that the posterior distribution  
27 of trees display a range of possible values. It is important to point out that the  
28 number of monophyletic clades is the *minimum* number of lineages establishing an  
29 infection; the actual number of infecting lineages can be much higher due to  
30 extinction of founding clades and suboptimal sampling of extant lineages in the  
31 donor and recipient.

32

33 The high diversity amongst the taxa from MP2 and MP3 also supports the idea of a  
34 high degree of shared diversity between MP2 and MP3 (Fig 4B). Comparing the  
35 within-host diversity at times of sampling showed that MP3 had only a little less  
36 diversity than MP2 (the mean pairwise taxa distance in MP2 and MP3 was 0.088 and  
37 0.079 substitutions/site, respectively, again in agreement with the ML tree at 0.090  
38 and 0.075 substitutions/site, respectively). Furthermore, the between MP2 and MP3  
39 population distance was somewhat larger (0.124 and 0.122 substitutions/site for  
40 posterior tree sample and ML tree, respectively). Together, these distances  
41 indicated that a large amount of diversity indeed had been transmitted.

42

43 The transmission of multiple lineages and the large diversity can only occur if there  
44 was either a large diverse population transmitted from the donor or if there were  
45 multiple transmission events between MP2 and MP3.

46

1 **Evidence for direction and frequency of transmission**

2 To evaluate how so many lineages and so much diversity was transmitted, we  
3 considered 4 possible scenarios to explain the data: MP2 or MP3 as the donor, and  
4 transmission at either a singular event or at multiple occasions (Fig 1). We  
5 considered the ratio of acceptance rates of samples from the priors based on either  
6 the topological statistics or the full set of tree statistics in the 4 possible scenarios as  
7 the relative evidence of one hypothesis over the other.

8  
9 The evidence based on the full tree statistics favored MP2 as the donor in both the  
10 single and multiple transmission scenarios; Bayes factor (BF) was 24 for the  
11 singular transmission case, 9.9 for multiple transmissions, and 12 for combined. If  
12 only topological statistics were used, the evidence was weaker but still favored MP2  
13 as the donor (BF 3.6 in singular transmission, 2.5 in multiple transmissions, and 2.8  
14 when combined). Thus, regardless of whether transmission occurred once (with  
15 multiple lineages) or many times (with one lineage at each time), the evidence  
16 points to MP2 as the donor.

17  
18 Both the topological and full tree statistic very weakly favored the ongoing  
19 transmission case (BF 1.3 and 2.1, respectively). From a purely statistical point of  
20 view, we could establish MP2 as the donor, but it was unclear whether there had  
21 been a single or multiple transmission events between the MP2-MP3 pair.

22  
23 **Point estimates and credible intervals of model parameters**

24 To further analyze support for one transmission scenario over the other, we  
25 evaluated the coalescent model parameters (Tab 1). With priors informed by  
26 available clinical and epidemiological information, the infection time of MP2 and  
27 linear growth rate  $\beta$  in MP2 were robust to our singular or multiple transmission  
28 event hypotheses. Infection time of MP3 and linear growth rate in MP3 were not,  
29 however. Because  $\beta_{MP3}$  is a function of the infection time of MP3, the lack of  
30 robustness in  $\beta_{MP3}$  is due to the lack of identifiability of the infection time in the  
31 ongoing transmission case. Similarly, the initial transmission time in the multiple  
32 transmission scenario is not well identified (posterior is close to prior). This is likely  
33 due to the fact that one cannot identify where the process starts (the infection time)  
34 if the rate of additional transmissions (migration) is not well constrained.

35  
36 **Extreme transmission rate makes multiple transmission events implausible**

37 The level of ongoing transmissions that is consistent with the data is very high ( $\rho$  in  
38 Tab 1 & Fig 5). In our model MP2 and MP3 reinfect each other at rate  $\rho$  and  $\rho/2$   
39 respectively over the time period from when MP3 was first infected until sampled  
40 (corresponding to the yellow area in Fig 2). The mean posterior ( $\rho= 1.3$ ) indicates  
41 that transmission events occur more than once per day, which is implausible as it  
42 implies very high contact rates and greater than ever reported risk of HIV-1  
43 transmission per heterosexual contact (BOILY *et al.* 2009). Even the lower posterior  
44 bound ( $\rho= 0.3$ ) is almost certainly biologically implausible.

45

1 Thus, while the ongoing transmission model explains the data well, the model only  
2 works at very high levels of transmission between MP2 and MP3 (Fig 5). This is due  
3 to the fact that the number of sampled lineages (20 sequences from MP2 plus 19  
4 sequences from MP3) is much smaller than the population sizes in MP2 and MP3. To  
5 observe any transmission events in the very small subset of the population that was  
6 sampled, there must be a very high rate of ongoing transmission between MP2 and  
7 MP3 (Fig 5). Therefore, from an epidemiological point of view, the implausible  
8 posterior transmission rate that is needed to support our data under the ongoing  
9 transmission scenario rather lends support for single transmission of multiple  
10 lineages.

11

## 12 **Robust estimation of transmitted diversity**

13 Effective population size ( $N_e$ ) is a model construct based on an idealized population  
14 and is thus an abstract formalism that can be difficult to interpret. However,  $N_e$  can  
15 be linked to the more concrete and measurable population diversity. Our maximum  
16 posterior estimate of  $\alpha = 116$  does not necessarily imply that exactly 116 virions  
17 were transmitted; we can, however, estimate the diversity of the establishing  
18 inoculum by simulating an evolutionary process from the posterior distribution of  
19 parameters corresponding to the transmitted lineages. Hence, while we cannot say  
20 exactly how many physical virions established the infection, we found that the mean  
21 diversity of the establishing population after transmission was robust over the  
22 posterior distribution of parameters. Notably, the diversity of the transmitted  
23 population was nearly as diverse (0.069 substitutions/site) as was later observed  
24 among the sequence clones in MP2 and MP3 at time of sampling (see 1<sup>st</sup> results  
25 section). Note also that the total diversity at time of sampling was likely larger than  
26 that observed among the clones.

27

28 To investigate the effect of  $\alpha$  on mean diversity at transmission, we first measured  
29 diversity as  $\alpha$  increased while other parameters were fixed at their maximum  
30 posterior values (Tab 1). As the number of transmitted lineages increases linearly  
31 the expected transmitted diversity initially increases rapidly (m1 in Fig 6). Allowing  
32 all model parameters to vary, we see a similar increase in diversity with increasing  
33  $\alpha$  (m2 in Fig 6). In both situations diversity of the founding infection does not  
34 increase beyond  $\alpha > 20$ . Note that the diversity plateaus at a higher level when only  $\alpha$   
35 is varied; when all remaining parameters are sampled from the posterior they have  
36 a compensating effect (e.g. lower  $\beta_{MP2}$ ) to explain the robust diversity estimate.  
37 Naturally, as higher  $\alpha$  increases the certainty that transmission involves high  
38 diversity, a corresponding decrease in the standard deviation is observed (Fig 6).  
39 Thus, while there is a non-linear relationship between  $\alpha$  and transmitted diversity,  
40 we can be reasonably certain that MP2 infected MP3 with a highly diverse inoculum.  
41 This result is robust even if we consider only the range of the absolute minimum  
42 number of transmitted viruses implied by the posterior distribution of the number  
43 of monophyletic clades in MP3 (4-14).

44

45

46

1 **DISCUSSION**

2

3 In this study we show how to apply previously described theoretical evaluations of  
4 epidemiological linkage to a real HIV-1 transmission case that involved a highly  
5 diverse founding HIV-1 population. We show that one can simultaneously estimate  
6 direction, diversity, and frequency of the transmission event(s). We used a  
7 previously developed within-host coalescent framework (ROMERO-SEVERSON *et al.*  
8 2014), and expanded it by allowing additional transmission events (migration)  
9 between the hosts. Inference was achieved using an ABC method informed by  
10 topological and distance-based tree statistics, which allowed Bayes factor  
11 comparisons between alternative epidemiological scenarios.

12

13 The transmission between MP2 and MP3 involved many lineages, probably more  
14 than we could observe among the limited sample of HIV-1 sequences derived from  
15 the patients. It is impossible to know exactly how many lineages were transmitted  
16 with these data. Our ABC framework can however estimate the diversity that was  
17 transmitted, and arguably this measure is more important from a clinical  
18 perspective as it may relate to how difficult it is to combat the incoming virus for the  
19 immune system, antiviral drugs, and future vaccines. We estimated that the  
20 inoculum that infected MP3 had a diversity of 0.069 substitutions/site, which is very  
21 high (corresponding to years of diversification). This level of diversity is equivalent  
22 to an incoming effective population size,  $\alpha$ , of either  $\approx 100$  or alternatively there was  
23 additional transmission events between MP2 and MP3 at levels that are highly  
24 unrealistic compared to empirical estimates of heterosexual transmission rates  
25 (BOILY *et al.* 2009). Thus, the transmission of the degree of diversity in this case  
26 seems to be the result of a single transmission with a very diverse inoculum  
27 involving many phylogenetic lineages from the donor.

28

29 HIV-1 co-infection has been defined as infection of several HIV-1 genetically diverse  
30 virions before seroconversion (typically 21 days after infection (COHEN *et al.* 2011))  
31 or within a somewhat later time (3-6 months) when a strong immune response has  
32 developed to the initial inoculum, and super-infection as additional infections after  
33 the strong immune response has been established (VAN DER KUYL AND CORNELISSEN  
34 2007; RONEN *et al.* 2013). In addition, super-infection is often thought of as an  
35 additional infection from another donor than the initial one. In the transmission  
36 case we studied here, both co- and super-infection was evaluated involving only the  
37 original donor and recipient, a stable heterosexual couple. Thus, with repeated  
38 contacts over time, transmissions may span and blur the defined periods of co- and super-  
39 infection. Importantly, HIV-1 evolves significantly during any period  $>1$   
40 month (SKAR *et al.* 2011), putting later transmitted variants somewhere in between  
41 the genetic diversity possible from co-infection from the original donor and super-  
42 infection from another donor, further blurring the co- and super-infection  
43 distinctions. Thus, while super-infection involving multiple donors appears rare  
44 (VAN DER KUYL AND CORNELISSEN 2007), given the fact that 20-40% of sexual infections  
45 involve  $>1$  genetic variant (KEELE *et al.* 2008; SALAZAR-GONZALEZ *et al.* 2009; LI *et al.*  
46 2010; RIEDER *et al.* 2011), ongoing transmission between stable couples as

1 investigated here may be more common than previously realized. On the other hand,  
2 at least in the case we studied here, ongoing transmission seemed unrealistic as it  
3 implied an impracticable high transmission rate (BOILY *et al.* 2009). Our study  
4 provides the first results of modeling single versus ongoing transmission events to  
5 explain how multiple lineages could end up in the recipient. A possible extension to  
6 our framework could be to allow for transmission of >1 lineage at multiple times,  
7 but without additional data, e.g., frequent longitudinal and deep sampling, there  
8 would not be enough power to identify how many variants that were transmitted at  
9 each possible occasion.

10  
11 We have recently shown that a joint HIV-1 phylogeny of two epidemiologically  
12 linked hosts may reveal direction and directness when >1 lineage has been  
13 transmitted (ROMERO-SEVERSON *et al.* 2016). As predicted by our topological  
14 evaluation, the epidemiological record confirmed that the inferred paraphyletic-  
15 polyphyletic (PP) tree resulting from the MP2-MP3 transmission identified MP2 as  
16 donor with no intermediary link to MP3. We note that the MP2 donor assignment  
17 was independent of whether transmission occurred once with multiple lineages or  
18 many times with single lineages. We also expanded the theoretical predictions with  
19 evaluations of many possible trees that could reasonably explain the sequence data,  
20 i.e., by evaluating a posterior tree sample derived from a Bayesian MCMC tree  
21 search using MrBayes. We show that in our case a ML tree reconstructed with  
22 PhyML (using NNI+SPR search) gives a good point estimate of the number of  
23 transmitted lineages in the sample, but that this gives no idea of the possible range.  
24 We hypothesize that there are situations when a ML-based estimate may not agree  
25 with the maximum posterior estimate. This becomes especially true in more  
26 complex models when parameters can compensate for each other to explain the  
27 data.

28  
29 In conclusion, taking phylogenetic uncertainty into account, we have created a  
30 framework that can evaluate how much diversity is transmitted, and whether  
31 transmission occurs once or over a period of time. We argue that estimating the  
32 transmitted diversity in the inoculum may reveal more about how difficult a  
33 transmission would be to prevent or fight than trying to find the exact number of  
34 transmitted lineages.

35

36

## 37 **ACKNOWLEDGEMENTS**

38

39 Research reported in this publication was supported by the NIAID/NIH under  
40 award number R01AI087520 and by grants PTDC/SAU-EPI/122400/2010,  
41 VIH/SAU/0029/2011 and UID/Multi/04413/2013 from Fundação para a Ciência e  
42 Tecnologia (FCT), Portugal. Inês Bárto was supported by a post-doc fellowship  
43 (SFRH/BPD/76225/2011) from FCT, Portugal. Ingo Bulla was supported by a post-  
44 doc fellowship (BU 2685/4-1) from the Deutsche Forschungsgemeinschaft.

45

46

1 **FIGURE LEGENDS**

2

3 **Figure 1. Phylogenetic assessment of transmission scenario.** Given a joint  
4 donor-recipient HIV-1 phylogenetic tree that suggests transmission of multiple  
5 lineages, two transmission scenarios are possible: 1) Transmission of multiple  
6 lineages at a single transmission event (co-infection), or 2) transmission of single  
7 lineages at multiple events (super-infection). In this example, host A (blue) is donor  
8 and B is recipient (red). In the observed phylogeny the root host-label (A, B, or  
9 equivocal [\*]) is derived by standard maximum parsimony. At time of transmission  
10 ( $t_{trans A \rightarrow B}$ ) either multiple lineages are transmitted (single transmission event with  
11  $\alpha_B$  lineages) or the initial transmission takes place (multiple transmission events  
12 each at  $\alpha=1$ ). Additional transmissions (migration) occurs at later time points ( $t_{trans 2}$   
13 and  $t_{trans 3}$ ) at rate  $\rho$ . The effective populations grow at  $\beta_A$  and  $\beta_B$  in donor and  
14 recipient, respectively. Samples with individual HIV-1 clonal sequences are taken at  
15  $t_{sample A}$  and  $t_{sample B}$ , respectively.

16

17 **Figure 2. Principle joint donor-recipient time-scaled phylogeny.** When a donor  
18 (A, blue) infects a recipient (B, red), the possible time-interval when transmission  
19 could have occurred (yellow field) is restricted in a time-scaled topology of when  
20 the most recent donor-recipient (A-B) coalescence occurred among the sampled  
21 lineages and when the recipient was sampled at  $t_B$ . The actual transmission ( $t_{trans}$ )  
22 must have occurred in this interval. The “source population” in direct transmission  
23 exists in the donor (blue field), from which at least 2 lineages were transmitted in  
24 this example to the donor (red fields). Note that if  $t_{trans}$  occurred later at least 3  
25 lineages could have been transmitted.

26

27 **Figure 3. Maximum likelihood reconstruction of the MP2-MP3 joint HIV-1 env**  
28 **phylogeny.** MP1 (yellow) did not infect either MP2 or MP3 (Fig S1), and is used to  
29 root the MP2 (red) and MP3 (blue) HIV-1 tree. Clades with aLTR support ( $>0.90$ ) are  
30 indicated with a “S”. The topology of this tree suggested that at least 7 lineages were  
31 transmitted from MP2 to MP3. Because the branch lengths were zero or near zero in  
32 the bottom clade, we added a small distance for readability purpose to show the 4  
33 possible transmitted lineages that the topology suggested in this clade. Partially to  
34 avoid depending on this single (best) tree, we evaluated 30,000 posterior trees  
35 presented in in Figure 4.

36

37 **Figure 4. Evaluation of posterior tree sample.** 30,000 MrBayes trees were  
38 sampled after burn-in to evaluate the possible range of inferred minimum number  
39 of transmitted lineages and sampled diversity in the hosts. **(A)** The number of MP3  
40 monophyletic clades in the joint MP2+MP3 trees ranged from 4–14 in the Bayesian  
41 posterior MCMC sample with a mean of 7.7 near the ML estimate at 7. **(B)** The  
42 diversity as measured by the patristic tree distance (substitutions/site) in hosts  
43 MP2, MP3, and between the HIV-1 populations.

44

45 **Figure 5. Migration analysis in the multiple transmission events scenario.**  
46 Simulation results using our ABC coalescent-based method showed that the

1 transmission rate  $\rho$  (x-axis) implied a very high number of migration events in the  
2 modeled population to explain the number of MP3 monophyletic clades observed in  
3 the sample (y-axis). Red line, median trend of 50,000 simulations; grey envelope  
4 with blue edges, 5-95% interval; grey open circles, individual simulation results.  
5

6 **Figure 6. Inference of the transmitted diversity among lineages.** Simulation m1  
7 shows how the transmitted diversity changes as only  $\alpha$  increases (with all other  
8 model parameters fixed at the maximum posterior values in Tab 1). Simulation m2  
9 shows the same trend when all model parameters are optimized at different  $\alpha$ . s1  
10 and s2 are the corresponding standard deviations.  
11

12  
13  
14

15 **Fig S1. Time-scaled phylogeny of MP1, MP2, and MP3 HIV-1 populations**  
16 **compared to local and database control sequences.** MP1, MP2, and MP3  
17 sequences are indicated by color, and control sequences in black. Numbers at nodes  
18 indicate posterior support values.  
19

20  
21

1    **REFERENCES**

2

3    Altfeld, M., T. M. Allen, X. G. Yu, M. N. Johnston, D. Agrawal *et al.*, 2002 HIV-1  
4    superinfection despite broad CD8+ T-cell responses containing replication of  
5    the primary virus. *Nature* 420: 434-439.

6    Anisimova, M., and O. Gascuel, 2006 Approximate likelihood-ratio test for branches:  
7    A fast, accurate, and powerful alternative. *Syst Biol* 55: 539-552.

8    Bartolo, I., C. Rocha, J. Bartolomeu, A. Gama, R. Marcelino *et al.*, 2009 Highly  
9    divergent subtypes and new recombinant forms prevail in the HIV/AIDS  
10    epidemic in Angola: new insights into the origins of the AIDS pandemic.  
11    *Infect Genet Evol* 9: 672-682.

12    Boily, M. C., R. F. Baggaley, L. Wang, B. Masse, R. G. White *et al.*, 2009 Heterosexual  
13    risk of HIV-1 infection per sexual act: systematic review and meta-analysis of  
14    observational studies. *Lancet Infect Dis* 9: 118-129.

15    Carrillo, F. Y., R. Sanjuan, A. Moya and J. M. Cuevas, 2007 The effect of co- and  
16    superinfection on the adaptive dynamics of vesicular stomatitis virus. *Infect*  
17    *Genet Evol* 7: 69-73.

18    Cohen, M. S., G. M. Shaw, A. J. McMichael and B. F. Haynes, 2011 Acute HIV-1  
19    Infection. *The New England Journal of Medicine* 364: 1943-1954.

20    Dollo, 1893 Les lois de l'évolution. *Bull Soc Belge Geol Pal Hydr* VII: 164-166.

21    Gottlieb, G. S., D. C. Nickle, M. A. Jensen, K. G. Wong, J. Grobler *et al.*, 2004 Dual HIV-1  
22    infection associated with rapid disease progression. *Lancet* 363: 619-622.

23    Grobler, J., C. M. Gray, C. Rademeyer, C. Seoighe, G. Ramjee *et al.*, 2004 Incidence of  
24    HIV-1 dual infection and its association with increased viral load set point in  
25    a cohort of HIV-1 subtype C-infected female sex workers. *J Infect Dis* 190:  
26    1355-1359.

27    Guindon, S., F. Lethiec, P. Duroux and O. Gascuel, 2005 PHYML Online--a web server  
28    for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res*  
29    33: W557-559.

30    Katoh, K., and H. Toh, 2008 Recent developments in the MAFFT multiple sequence  
31    alignment program. *Briefings in bioinformatics* 9: 286-298.

32    Keele, B. F., E. E. Giorgi, J. F. Salazar-Gonzalez, J. M. Decker, K. T. Pham *et al.*, 2008  
33    Identification and characterization of transmitted and early founder virus  
34    envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 105: 7552-  
35    7557.

36    Leigh Brown, A. J., 1997 Analysis of HIV-1 env gene sequences reveals evidence for a  
37    low effective number in the viral population. *Proceedings of the National*  
38    *Academy of Sciences USA* 94: 1862-1865.

39    Leitner, T., and J. Albert, 1999 The molecular clock of HIV-1 unveiled through  
40    analysis of a known transmission history. *Proceedings of the National*  
41    *Academy of Sciences USA* 96: 10752-10757.

42    Leitner, T., and W. M. Fitch, 1999 The phylogenetics of known transmission histories  
43    in *The evolution of HIV*, edited by K. A. Crandall. Johns Hopkins Univ. Press,  
44    Baltimore, MD.

45    Li, H., K. J. Bar, S. Wang, J. M. Decker, Y. Chen *et al.*, 2010 High Multiplicity Infection  
46    by HIV-1 in Men Who Have Sex with Men. *PLoS Pathog* 6: e1000890.

1 Nijhuis, M., C. A. Boucher, P. Schipper, T. Leitner, R. Schuurman *et al.*, 1998  
2 Stochastic processes strongly influence HIV-1 evolution during suboptimal  
3 protease-inhibitor therapy. *Proceedings of the National Academy of Sciences*  
4 of the United States of America 95: 14441-14446.

5 Pennings, P. S., S. Kryazhimskiy and J. Wakeley, 2014 Loss and recovery of genetic  
6 diversity in adapting populations of HIV. *PLoS genetics* 10: e1004000.

7 Rieder, P., B. Joos, A. U. Scherrer, H. Kuster, D. Braun *et al.*, 2011 Characterization of  
8 human immunodeficiency virus type 1 (HIV-1) diversity and tropism in 145  
9 patients with primary HIV-1 infection. *Clin Infect Dis* 53: 1271-1279.

10 Romero-Severson, E., H. Skar, I. Bulla, J. Albert and T. Leitner, 2014 Timing and  
11 Order of Transmission Events Is Not Directly Reflected in a Pathogen  
12 Phylogeny. *Mol Biol Evol* 31: 2472-2482.

13 Romero-Severson, E. O., I. Bulla and T. Leitner, 2016 Phylogenetically resolving  
14 epidemiologic linkage. *Proc Natl Acad Sci U S A* 113: 2690-2695.

15 Ronen, K., C. O. McCoy, F. A. Matsen, D. F. Boyd, S. Emery *et al.*, 2013 HIV-1  
16 superinfection occurs less frequently than initial infection in a cohort of high-  
17 risk Kenyan women. *PLoS Pathog* 9: e1003593.

18 Ronquist, F., and J. P. Huelsenbeck, 2003 MrBayes 3: Bayesian phylogenetic  
19 inference under mixed models. *Bioinformatics* 19: 1572-1574.

20 Salazar-Gonzalez, J. F., M. G. Salazar, B. F. Keele, G. H. Learn, E. E. Giorgi *et al.*, 2009  
21 Genetic identity, biological phenotype, and evolutionary pathways of  
22 transmitted/founder viruses in acute and early HIV-1 infection. *The Journal*  
23 of experimental medicine 206: 1273-1289.

24 Sanborn, K. B., M. Somasundaran, K. Luzuriaga and T. Leitner, 2015 Recombination  
25 elevates the effective evolutionary rate and facilitates the establishment of  
26 HIV-1 infection in infants after mother-to-child transmission. *Retrovirology*  
27 12: 96.

28 Shattock, R. J., and J. P. Moore, 2003 Inhibiting sexual transmission of HIV-1 infection.  
29 *Nat Rev Microbiol* 1: 25-34.

30 Skar, H., R. N. Gutenkunst, K. Wilbe Ramsay, A. Alaeus, J. Albert *et al.*, 2011 Daily  
31 sampling of an HIV-1 patient with slowly progressing disease displays  
32 persistence of multiple env subpopulations consistent with neutrality. *PLoS*  
33 ONE 6: e21747.

34 Smith, D. M., M. C. Strain, S. D. Frost, S. K. Pillai, J. K. Wong *et al.*, 2006 Lack of  
35 neutralizing antibody response to HIV-1 predisposes to superinfection.  
36 *Virology* 355: 1-5.

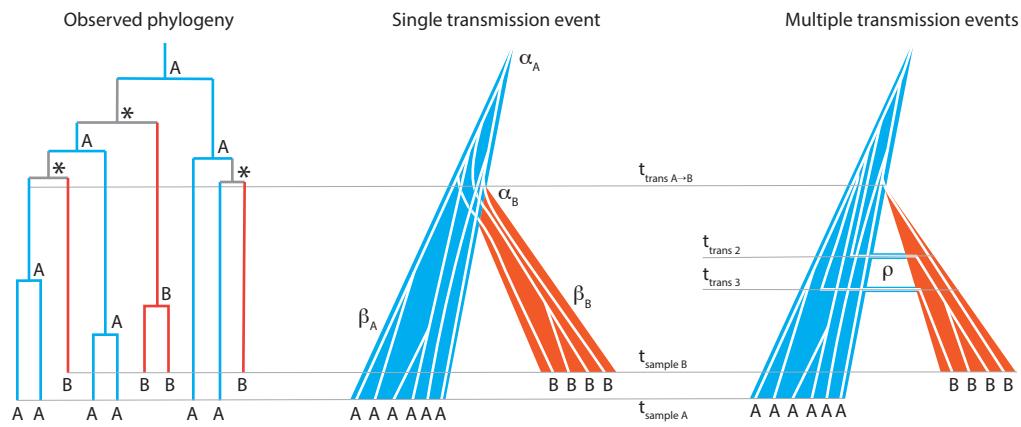
37 Smith, D. M., J. K. Wong, G. K. Hightower, C. C. Ignacio, K. K. Koelsch *et al.*, 2004  
38 Incidence of HIV superinfection following primary infection. *JAMA* 292:  
39 1177-1178.

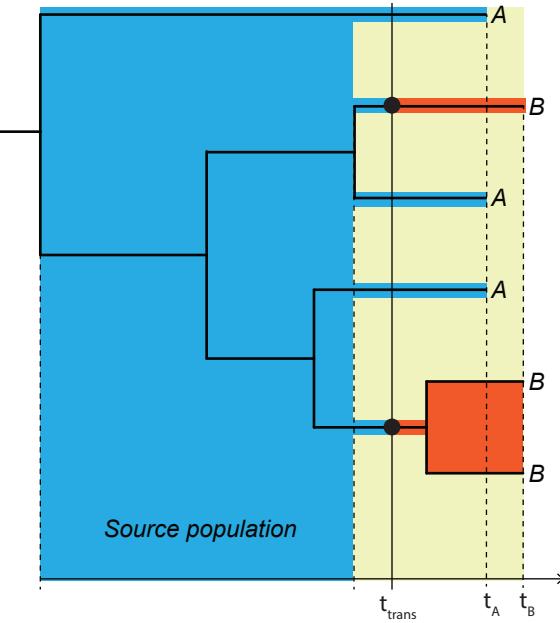
40 Smith, D. M., J. K. Wong, G. K. Hightower, C. C. Ignacio, K. K. Koelsch *et al.*, 2005 HIV  
41 drug resistance acquired through superinfection. *AIDS* 19: 1251-1256.

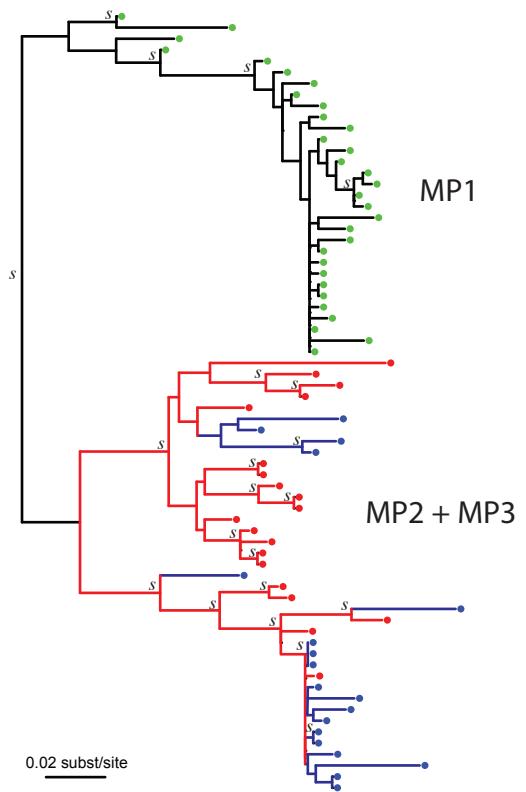
42 van der Kuyl, A. C., and M. Cornelissen, 2007 Identifying HIV-1 dual infections.  
43 *Retrovirology* 4: 67.

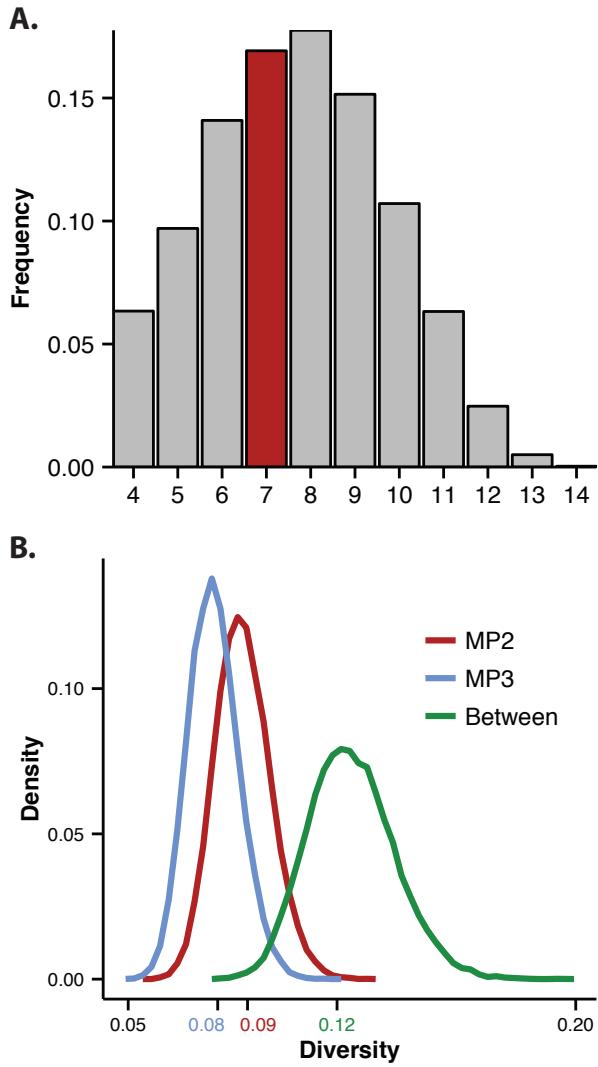
44 Wakeley, J., 2009 *Coalescent theory: an introduction*. Roberts and Company  
45 Publishers, Greenwood Village, CO.

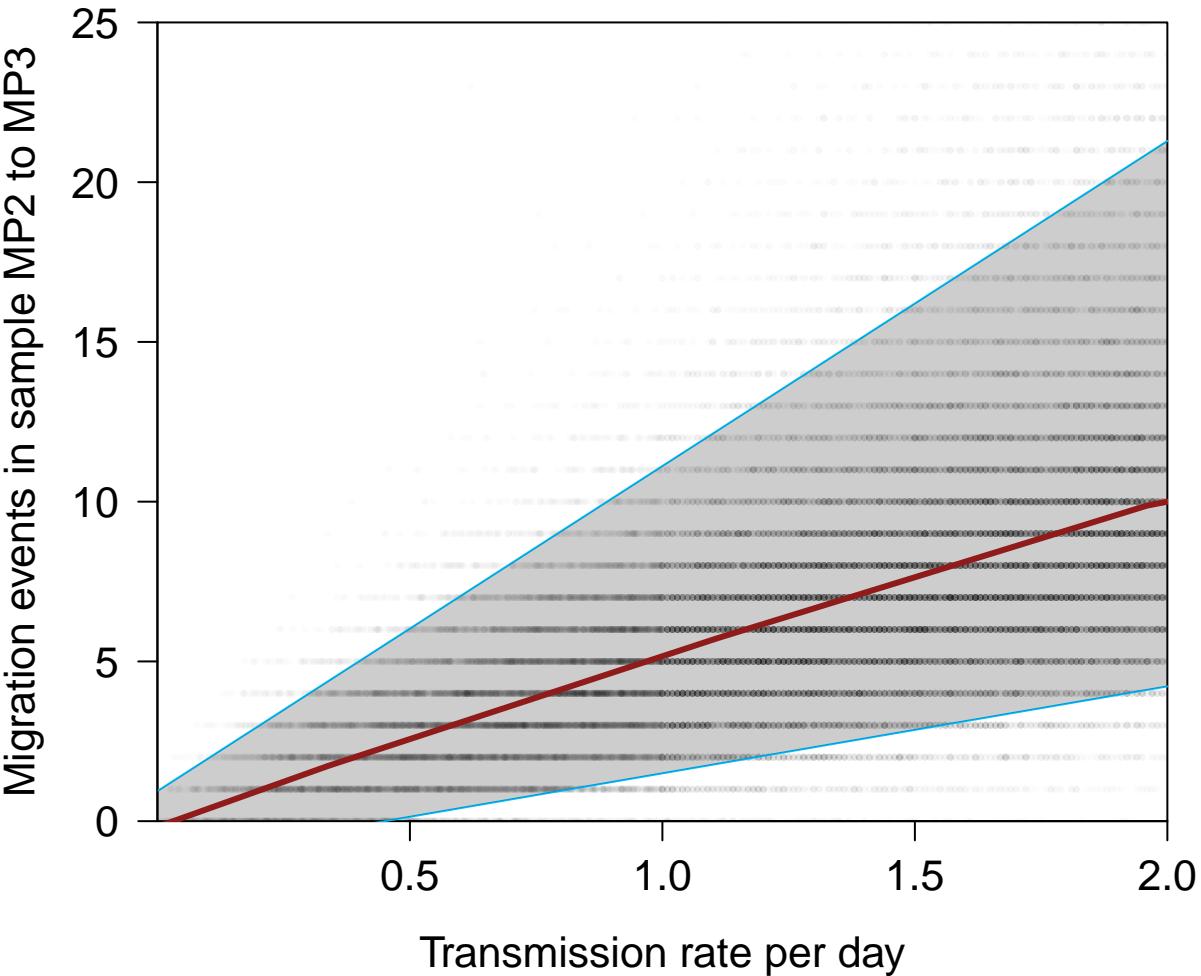
1 Yang, O. O., E. S. Daar, B. D. Jamieson, A. Balamurugan, D. M. Smith *et al.*, 2005  
2 Human immunodeficiency virus type 1 clade B superinfection: evidence for  
3 differential immune containment of distinct clade B strains. *J Virol* 79: 860-  
4 868.  
5











Diversity at transmission

