# Suplementary Text

Dataset compilation - transmission chains

To obtain a comprehensive collection of genetic data sets from HIV-1 transmission chains, we employed the search feature of the HIV database (http://www.hiv.lanl.gov/) that allows retrieving intra-host data sets. At the time of the query (November 2012), 953 sets were reported without specific selection criteria. The title and abstract of the published studies were screened and each study that potentially involved clonal sequence data together with known transmission route and/or known infection time frame was subjected to a more detailed screening. We also undertook a literature search based on the references with high potential from the selected studies.

All transmission pairs for which time-stamped clonal sequences of both donor and recipient was available together with at least an upper bound on the transmission interval were grouped under the following risk groups: heterosexual (HET), men having sex with men (MSM) and blood contact (BC). The denominator ‘BC’ groups together not only those pairs infected through blood transfusion, but also contains transmission pairs involving a bite [[1](#_ENREF_1)], a knife-fight [[2](#_ENREF_2)], surgical procedures [[3](#_ENREF_3)] and malignant injection [[4](#_ENREF_4)]. A summary overview of the found transmission chains is presented in Table 1 in the main text.

The time between infection and sampling of the recipient patients is an important variable for accurately quantifying the loss of diversity at transmission using a population genetic approach [[5](#_ENREF_5)]. The available number of samples is also important as this determines the amount of information available to the phylogenetic reconstructions. A number of descriptive statistics of these two parameters are summarized in Additional Table 1. Briefly, most recipients were sampled at one time point and both MSM and HET are more frequently sampled earlier since time of infection as compared to BC.

Dataset compilation - risk group data sets for evolutionary rate estimation

Subtypes have been shown to evolve at different rates [[6](#_ENREF_6)]. To compare the evolutionary rate between risk groups within subtypes, we downloaded near complete genome datasets for subtype A1, B and CRF01\_AE from the Los Alamos HIV database. The other subtypes lacked sufficient data of at least 2 risk groups to allow for a meaningful comparative analysis among risk groups. This discrepancy is particularly present for subtype C where, in contrast to the large amount of HET sequence data, only three MSM near full genomes are available, and where MSM data from smaller genomic fragments such as the *pol* and *env* CDS lack temporal signal (R2 values from a regression of root-to-tip divergence vs. sampling time are <0.1). The characteristics of the subtype-specific risk group datasets are given in Table 2 in the main text. An exploration of whether patterns of geographical spread may affect the within-subtype evolutionary is provided below.

Incorporating the uncertainty of sampling dates and transmission times

In order to apply the same time-scale for all analyses, all sampling and transmission dates were specified in units of days. Often however, sampling dates or transmission intervals are reported only approximately. Arbitrarily choosing for example the midpoint of the potential interval can introduce biases in the bottleneck size estimations because, especially in situations of recipient sampling close to the time of infection, there can be an interaction between the bottleneck size parameter and transmission/divergence times [[5](#_ENREF_5)]. To avoid this, we made use of the flexibility offered by BEAST [[7](#_ENREF_7)] to integrate out the transmission and sampling dates constrained by a time interval [[8](#_ENREF_8)]. This required an extension of the standard approach [[8](#_ENREF_8)] to accommodate sampling of a single date for a set of taxa representing a single clonal sample. We therefore extended BEAST [[7](#_ENREF_7)] to handle such situations, and specified uniform priors over a transmission or sampling interval with known boundaries by making use of the Fiebig stage [[9](#_ENREF_9)], time to seroconversion or symptoms of primary infection (see below).

Determination of transmission interval width

For some transmission chain datasets it was possible to approximate the boundaries for the transmission interval using the communicated Fiebig stage [[9](#_ENREF_9)]. Specifically, the earliest and latest possible infection dates were taken as the cumulative lower and upper boundaries of the 95% confidence intervals of the duration of the stages up to (lower boundary) or up to and including (upper boundary) the stage the recipient was in at the moment of sampling (Table 1). As an example, suppose the first sample was taken while the recipient was in Fiebig stage II. Our approach assumes that stage II could have begun at the earliest 10 days after infection (5 days eclipse phase + 5 days phase I) and could have ended at the latest 18 days later (10 days eclipse phase + 10 days phase I + 8 days phase II). Therefore, infection could have taken place 10 to 28 days before the sampling date.

When the time to seroconversion was known, we interpreted this as Fiebig stage III/IV because it is not always communicated which antibody detection test(s) was/were used and sensitivity of the tests has increased.

When the timing of symptoms of primary or acute HIV infection was given, we calculated the transmission time interval using Fiebig stage I/II boundaries (infection anywhere between 5 and 28 days before the onset of symptoms) [[10](#_ENREF_10)].

Table 1: Timings of the Fiebig stages used to specify the transmission interval boundaries

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage | shortest duration (95% CI) (days) | longest duration (95% CI) (days) | earliest start (days) | latest end (days) |
| eclipse | 5 | 10 | 0 | 10 |
| Fiebig I | 5 | 10 | 5 | 20 |
| Fiebig II | 4 | 8 | 10 | 28 |
| Fiebig III | 2 | 5 | 14 | 33 |
| Fiebig IV | 4 | 8 | 16 | 41 |
| Fiebig V | 40 | 122 | 20 | 163 |

Model selection

For all datasets, the log marginal likelihood was estimated for all demographic functions currently available under the transmission model (constant population size, exponential and logistic growth) [[5](#_ENREF_5)]. To this end, we made use of the path sampling (PS) [[11](#_ENREF_11)] and stepping-stone (SS) sampling estimators [[12](#_ENREF_12)] of the (log) marginal likelihood as implemented in BEAST [[13-15](#_ENREF_13)]. All (log) marginal likelihood estimates were checked for convergence by performing multiple runs with different computational settings. We consider a model to be better supported by the data than the competing hypothesis when the difference in log Bayes factor (logBF) support exceeds 5 [[16](#_ENREF_16)].

Selection of informative transmission chains for the fixed effects analysis

The parameterization of the population genetic dynamics throughout transmission in the transmission chain model may influence the bottleneck size estimations. Although we can select the best fitting parameterization using marginal likelihood estimation, the preference for a particular coalescent model may depend on the amount of available information: whereas a simple model may be required for independent estimation, a more complex model may still be suitable when sharing information in a hierarchical modeling setting. We therefore assess the robustness of the results with respect to demographic model specification, by first performing the HPM + fixed effects analysis with the within-host population dynamics described by the function obtained by model selection (‘best fit’). Next, we also ran the fixed effects analyses on these datasets consistently applying either a logistic or an exponential model for the demographic process to each transmission chain (‘logistic’ and ‘exponential’). In order to avoid a higher weight for transmission chains for which multiple genomic regions are available, the genomic regions were analyzed separately.

Our previous results using the transmission chain model [[5](#_ENREF_5)] indicate that the model is best informed by data that capture the early population dynamics. The poor mixing for the *gag* and *env* genomic regions for many datasets revealed that not all the transmission chain data sets can be used to properly inform the model. For the *pol* datasets, mixing was less of an issue but here the bottleneck size estimates were highly dependent on the specified model. Because of these issues we only focused on those datasets with first sampling close to the time of transmission and with good mixing properties for further analyses. This filtering step retained 17 HET and 11 MSM chains with *env* data for the HPM + fixed effects analysis.

Patterns of geographical spread do not affect the within-subtype evolutionary rate.

Various subtypes are known to have arisen through founder effects and by explicitly distinguishing between those, we take into account these major founder effects when comparing risk group evolutionary rates. However, also within a subtype dataset samples are from various geographical areas, and hence concerns may remain that the patterns of geographical spread may affect the evolutionary rate. To investigate this in detail, we have performed exploratory linear regression analyses of root-to-tip divergence as a function of sampling time for all data sets and visualised the contribution of samples from the different locations.

As an example we present the root-to-tip regression model for one dataset (the CRF01\_AE HET dataset) in Figure 1 and the associated residuals from this regression in Figure 2. The residual plots for all other datasets are also provided (Figures 3-7).

These plots illustrate that samples from different locations contribute in a roughly uniform manner to the divergence through time, or in other words, there is no clear pattern of noticeable positive or negative residuals by location, which reflects a roughly uniform rate across all locations.



Figure 1: Plot of the root-to-tip regression of divergence versus sampling date for the CRF01\_AE HET dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The full red line represents the fitted regression model.



Figure 2: Residual plot of the root-to-tip regression of divergence versus sampling date for the CRF01\_AE HET dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The dashed red line indicates that data points that perfectly fall on the fitted regression model have zero residual.



Figure 3: Residual plot of the root-to-tip regression of divergence versus sampling date for the CRF01\_AE MSM dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The dashed red line indicates that data points that perfectly fall on the fitted regression model have zero residual.



Figure 4: Residual plot of the root-to-tip regression of divergence versus sampling date for the subtype A1 HET dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The dashed red line indicates that data points that perfectly fall on the fitted regression model have zero residual.



Figure 5: Residual plot of the root-to-tip regression of divergence versus sampling date for the subtype A1 IDU dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The dashed red line indicates that data points that perfectly fall on the fitted regression model have zero residual.



Figure 6: Residual plot of the root-to-tip regression of divergence versus sampling date for the subtype B HET dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The dashed red line indicates that data points that perfectly fall on the fitted regression model have zero residual.



Figure 7: Residual plot of the root-to-tip regression of divergence versus sampling date for the subtype B MSM dataset. Data points are colored according to sampling location. The 2-letter country codes follow the international ISO 3166 Country names standard. The dashed red line indicates that data points that perfectly fall on the fitted regression model have zero residual.

Impact of donor viral diversity on the estimated loss of diversity

The amount of evolution in the donor is important when estimating the loss of diversity at transmission. Shankarappa *et al*. [[17](#_ENREF_17)], among others, have elegantly shown that the diversity of the HIV-1 viral population increases with the length of infection, and a higher diversity is usually observed in chronically infected patients [[18](#_ENREF_18)]. Consequently, the effective population size in donors with longstanding infections will likely yield higher estimates, leading to a higher *absolute* loss of diversity when compared to recently infected donors.

 To address this in more detail, we investigated the available clinical information on the donor for the informative subset of *env* datasets, and present this information in Table 2. The estimated donor population size estimates are, as expected, larger for chronically infected donors (left panel of Figure 8). There is however no strong association between the loss of diversity at transmission and the donor stage of infection in our sample (one-sided t-test p=0.345) despite the trend towards a more severe bottleneck in chronically infected sources (right panel of Figure 8). When plotting the estimates of the bottleneck against donor population size (Figure 9), the clustering in the 99%-100% range also illustrates that, irrespective of the donor’s estimated population size or stage of infection (both relate to viral diversity), there is generally an equally strong loss of diversity.

This can be explained by the fact that the bottleneck is measured as the *proportion* of the donor’s population size that survives transmission [[5](#_ENREF_5)]. A strong founder effect is expected to result in small proportions more or less independent of the absolute viral population size in the donor. For example, with transmission over 1 branch (connecting the donor and recipient population) and a relatively small population of Ne=100 (homogenous population, recent infection) and a larger population of Ne=1000 (diverse population, chronic infection), the loss of diversity will yield proportions of 0.99 and 0.999 respectively. Given the uncertainty associated with coalescent estimates, differences in these ranges are difficult to pick up. In addition, with individual estimates generally ranging around these values, a small fraction of multi-variant transmissions, as observed in SGA- based surveys [[19](#_ENREF_19)], will have a limited impact on the overall estimates of loss in diversity in each risk group. We see an example of this in the HET risk group, where the transmission chain with the smallest bottleneck (54.13%) only lowers the overall estimate with about 2.67%. The same holds true for the transmission chain in the MSM group (lowering the overall estimate with about 2.48%).

In summary, there is a trend for the HET donors to be chronically infected, while some of the MSM donors were recently infected at the time of transmission. Although this information is only available for a limited number of data sets in our analysis, it suggests that our sampling may reflect reality in the sense that a higher proportion of transmissions in the MSM risk group occur during recent infection [[20](#_ENREF_20)]. Even if a bias would exist in our sampling with respect to the donor’s length of infection, not finding a clear difference on transmitted diversity implies that this did not bias our results in either direction. A similar line of reasoning holds for the impact of the treatment history of the donor.

Table 2: Overview of the information on the donors's stage of infection of the informative subset of *env* data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | bottleneck size (%)a | population sizea | risk group | stageb |
| Frost-004-007 | 99.98 | 655 | MSM | recent |
| Frost-206-201 | 99.95 | 1307 | MSM | chronic |
| Frost-206-204 | 99.7 | 1318 | MSM | chronic |
| Frost-512-558 | 99.93 | 353 | MSM | recent |
| Frost-512-559 | 99.92 | 386 | MSM | recent |
| Frost-551-550 | 98.65 | 749 | MSM | recent |
| Frost-564-557 | 99.37 | 3584 | MSM | chronic |
| Lawson | 97.67 | 1242 | MSM | NA |
| Li-AD83 | 99.14 | 386 | MSM | recent |
| Herbeck2 | 71.71 | 1055 | MSM | recent |
| Herbeck4 | 95.93 | 27510 | MSM | chronic |
| Boeras-RW36 | 99.97 | 5731 | HET | chronic |
| Boeras-RW221 | 99.98 | 8803 | HET | chronic |
| Boeras-RW242 | 99.9 | 7856 | HET | chronic |
| Boeras-RW292 | 99.99 | 8109 | HET | chronic |
| Derdeyn-53 | 99.92 | 5870 | HET | chronic |
| Derdeyn-55 | 96.3 | 4988 | HET | chronic |
| Derdeyn-71 | 99.96 | 4040 | HET | chronic |
| Derdeyn-83 | 99.6 | 3858 | HET | chronic |
| Derdeyn-109 | 99.56 | 3704 | HET | chronic |
| Haaland-RW41 | 99.99 | 11704 | HET | chronic |
| Haaland-RW53 | 99.98 | 5246 | HET | chronic |
| Haaland-RW57 | 99.77 | 7299 | HET | chronic |
| Haaland-RW66 | 54.13 | 13493 | HET | NA |
| Haaland-ZM229 | 99.78 | 10221 | HET | chronic |
| Haaland-ZM243 | 99.95 | 10116 | HET | chronic |
| Liu-J | 99.98 | 5620 | HET | chronic |
| Wolfs-A11A12 | 99.44 | 1062 | HET | NA |

a Both the bottleneck and population size are given by their mean estimates.  b When the donor was infected >6 months before transmission, he/she was labeled as chronically infected. The population size refers to the estimated population size in the donor at the moment of transmission from the ‘best fit’ analysis. For the transmission couple reported by Lawson et al. [2002] it is only known the donor was infected ≥5 months, which is why the stage of infection at transmission is left blank. Similarly, it is only given that the source of couple ZM229 was at least 3 months infected. For this reason, this entry was also left blank.



Figure 8: Left panel: boxplot of the mean estimated donor population size by donor stage of infection. Right panel: boxplot of the mean estimated bottleneck size by donor stage of infection. Data are from Table 2.



Figure 9. Left panel: plot of the mean estimated bottleneck size versus mean estimated donor population size, colored by donor stage of infection. Right panel: same plot as in the left panel, but only with average bottleneck sizes ≥90%. Data are from Table 2.

References:

1. Andreo, S.M.S., et al., *HIV type 1 transmission by human bite.* AIDS Res Hum Retroviruses, 2004. **20**(4): p. 349-50.

2. Kao, C.-F., et al., *An uncommon case of HIV-1 transmission due to a knife fight.* AIDS Res Hum Retroviruses, 2011. **27**(2): p. 115-22.

3. Blanchard, A., et al., *Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients.* J Virol, 1998. **72**(5): p. 4537-40.

4. Metzker, M.L., et al., *Molecular evidence of HIV-1 transmission in a criminal case.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14292-7.

5. Vrancken, B., et al., *The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates.* PLoS Computational Biology, 2014. **10**(4): p. e1003505.

6. Abecasis, A.B., A.-M. Vandamme, and P. Lemey, *Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution.* J Virol, 2009. **83**(24): p. 12917-24.

7. Drummond, A.J., et al., *Bayesian phylogenetics with BEAUti and the BEAST 1.7.* Mol Biol Evol, 2012. **29**(8): p. 1969-73.

8. Shapiro, B., et al., *A Bayesian phylogenetic method to estimate unknown sequence ages.* Mol Biol Evol, 2011. **28**(2): p. 879-87.

9. Fiebig, E.W., et al., *Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection.* AIDS, 2003. **17**(13): p. 1871-9.

10. Cohen, M.S., et al., *The detection of acute HIV infection.* J Infect Dis, 2010. **202 Suppl 2**: p. S270-7.

11. Lartillot, N. and H. Philippe, *Computing Bayes factors using thermodynamic integration.* Syst Biol, 2006. **55**(2): p. 195-207.

12. Xie, W., et al., *Improving marginal likelihood estimation for Bayesian phylogenetic model selection.* Syst Biol, 2011. **60**(2): p. 150-60.

13. Baele, G. and P. Lemey, *Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency.* Bioinformatics, 2013. **29**(16): p. 1970-1979.

14. Baele, G., et al., *Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty.* Mol Biol Evol, 2012. **29**(9): p. 2157-67.

15. Baele, G., et al., *Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics.* Mol Biol Evol, 2012.

16. Kass, R.E. and A.E. Raftery, *Bayes Factors.* journal of the american Statistical Association, 1995. **90**(430): p. 773-795.

17. Shankarappa, R., et al., *Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection.* J Virol, 1999. **73**(12): p. 10489-502.

18. Frost, S.D.W., et al., *Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B.* J Virol, 2005. **79**(10): p. 6523-7.

19. Shaw, G.M. and E. Hunter, *HIV transmission.* Cold Spring Harb Perspect Med, 2012. **2**(11).

20. Brenner, B., M.A. Wainberg, and M. Roger, *Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions.* AIDS, 2013. **27**(7): p. 1045-57.