Supplementary Information for manuscript
"The genomic signature of population reconnection following isolation: From theory to HIV"

Nicolas Alcala*,§,†, Jeffrey D. Jensen†, Amalio Telenti† and Séverine Vuilleumier*,†,**
* Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland. § Department of Biology, Stanford University, Stanford, CA 94305-5020, USA. † School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland. ‡ The J.Craig Venter Institute, 4120 Capricorn Lane, La Jolla, CA 92037, USA. ** Institute of Microbiology, University Hospital and University of Lausanne, CH-1011 Lausanne, Switzerland.

Contents

1 Bimodality on pairwise nucleotide differences as a signal of population reconnection after isolation 3

2 The signature of population reconnection after isolation on the Site Frequency Spectrum (SFS) 3

2.1 Impact of the sampling scheme on the signature of population reconnection after isolation on the SFS 3

2.2 Text A. Derivation of optimal tests to detect the signature of population reconnection after isolation from the site frequency spectrum (SFS) 4

2.2.1 Method to derive $T_{\Omega}$ and $T_{\Omega}^t$ 5

2.2.2 Derivation of the expected SFS under the population reconnection scenario 5

2.2.3 General properties of test statistics $T_{\Omega}^l$ and $T_{\Omega}^l$ 5

2.2.4 Power of the test statistics to detect departure from neutrality 5

3 Robustness of the signature 7

4 Detection of past isolation and current reconnection of HIV-1 subtypes 10
1 Bimodality on pairwise nucleotide differences as a signal of population reconnection after isolation

In this section, we assess the power to detect population reconnection after isolation from the distribution of pairwise nucleotide differences, by testing for the bimodality pattern derived in Appendix A, illustrated in Fig. 1 and observed in the case of HIV-1 subtypes in Fig. 4.

![Diagram](image1)

Figure A Number of samples \( n \) needed to reach a high power to detect bimodality (95%) in the distribution of pairwise nucleotide differences (A) within-population, \( \pi_w \), and (B) between-population, \( \pi_b \), as a function of the duration of the reconnection period \( T_{reco} \). The gray areas correspond to different scaled migration rates \( M \); areas are plotted above each other and sorted by increasing \( M \) values; dashed lines represent the limits of areas which are overlapped by other areas. The probabilities to draw a sample from each mode of the distributions of \( \pi_w \) and \( \pi_b \) are computed in Appendix A (eqs. A6.a and A6.b).

2 The signature of population reconnection after isolation on the Site Frequency Spectrum (SFS)

2.1 Impact of the sampling scheme on the signature of population reconnection after isolation on the SFS

In this section, we assess the impact of the sampling scheme on the signature of population reconnection after isolation on the SFS. We consider two sampling scheme: unbalanced sampling, where the sample sizes are different in each population (Fig. B), and local sampling, where all samples come from a single population (Fig. C). These figures are to be compared with that under a balanced sampling scheme (same sample size from each population) presented in the main text (Fig. 2A).
Figure B Temporal signature of reconnection after isolation on the total site frequency spectrum (SFS), for an unbalanced sampling scheme (sample sizes in each of the four populations are 16, 18, 22 and 24), as opposed to the signature under a balanced sampling scheme presented Fig. 2A. The gray solid line represents the expected SFS in equilibrium connected populations. The total SFS is represented as a function of the number of generations since the reconnection event ($T_{reco}$). The maximum number of peaks corresponds to the sum of the number of combinations of $i$ populations among $d$ (i.e., $\binom{d}{i}$), for $i$ ranging from 1 to $d-1$ (i.e., a total of $2^d-2$ peaks); we subtract 1 from this number for each two sets of populations which sample sizes sum to the same value (i.e., for each two sets $\{i_1, i_2, ..., i_i\}$ and $\{j_1, j_2, ..., j_j\}$ for which $n_{i1}+n_{i2}+...+n_{i_i}=n_{j1}+n_{j2}+...+n_{j_j}$, with $i_1 \neq i_2, ..., j_1 \neq j_2, ..., j_j$). For example, in the figure we have $d=4$ populations, and 2 sets of populations which sum of sample sizes are equal ($16+24$ and $18+22$), so we have $2^4-2-1=13$ peaks. Results are for $d=4$ populations, $\theta=2$, and $M=5$ during connection periods. Means are over 5,000 replicate simulations.

Figure C Temporal signature of population reconnection after isolation under a local sampling scheme (local site frequency spectrum), as opposed to the signature under a balanced sampling scheme presented Fig. 2A. The gray solid and dashed lines represent the expected SFS in equilibrium isolated and connected populations, respectively. The local SFS is represented as a function of the number of generations since the reconnection event ($T_{reco}$). The signature of population reconnection after isolation on the local SFS could be confounded with the signature of other events. First, the SFS is "U-shaped", similar to what is expected when variants are under positive selection. Second, we observe an excess of variants at intermediate frequencies, similar to the SFS observed under balancing selection. Parameters are $\theta=2$, $T_{iso}=6$, and $M=5$ during connection periods. Means are over 5,000 replicate simulations.

2.2 Text A. Derivation of optimal tests to detect the signature of population reconnection after isolation from the site frequency spectrum (SFS)

In this section, we derive, using the method to detect departures from neutrality from Ferretti et al. (2010), two test statistics to detect the signature of population reconnection after isolation from the local and total SFS, respectively denoted $T_{\Omega}$ and $T_{\Omega}^l$. We first present the method of Ferretti et al. (2010), which requires the expected SFS under both scenarios. Second, we present a method to derive the expected SFS. Finally, we present the properties of the resulting test statistics and we assess their power. Statistic $T_{\Omega}^l$ is denoted $T_{\Omega}$ in the main text and used in the section Detection of past isolation and current reconnection of HIV-1 subtypes of the SI and the corresponding section in the results of the main text.
2.2.1 Method to derive $T_{\Omega}$ and $T_{t\Omega}$

Achaz (2009) demonstrates that all neutrality tests based on the SFS can be written as:

$$T_{\Omega} = \sum_{i=1}^{n-1} \Omega_i \xi_i \sqrt{\text{Var} \left( \sum_{j=1}^{n-1} \Omega_j \xi_j \right)}$$  \hspace{5cm} (A.1)

where $\Omega_i$, with $i = 1, 2, ..., n - 1$ are the weights which uniquely define the test (for example, taking $\Omega_i = 2(n - i)/[n(n - 1)] - 1/(i \sum_j 1/j)$ leads to Tajima’s $D$) and the $\xi_i$ are the classes of the SFS (i.e., the number of variants at frequency $i/n$). The variance in the denominator can be computed using equation 9 from Achaz (2009). As shown in Ferretti et al. (2010), the optimal test against a given demographic scenario can be built by choosing the $\Omega_i$ that maximizes the expected value of $T_{\Omega}$ under the given scenario. By denoting $\xi^l_i$ and $\xi^t_i$ the expected values of the classes of the local and total SFS under the population reconnection scenario, the optimal test statistics $T^l_{\Omega}$ and $T^t_{\Omega}$ have the following weights, denoted $\Omega^l_i$ and $\Omega^t_i$, respectively (Ferretti et al. 2010):

$$\Omega^l_i = \frac{\xi^l_i}{\sum_j \xi^l_j} - \frac{1}{ia_n}$$

$$\Omega^t_i = \frac{\xi^t_i}{\sum_j \xi^t_j} - \frac{1}{ia_n}$$  \hspace{5cm} (A.2)

where $a_n = \sum_{i=0}^{n-1} 1/i$.

$T^l_{\Omega}$ and $T^t_{\Omega}$ have positive values under the population reconnection scenario and approach 0 under neutrality. Indeed, the positive terms in equation A.2 correspond to the expected value of the SFS under the population reconnection scenario, while the negative terms correspond to the expected value under neutrality. Consequently, the statistics should be used in one-sided tests.

2.2.2 Derivation of the expected SFS under the population reconnection scenario

Values $\xi^l_i$ and $\xi^t_i$ can be estimated from coalescent simulations under each scenario with fixed parameters. For our study, we used the values presented in Figures 2 (second column) and C (second column). Indeed, these values present the most prominent features of the signature of population reconnection after isolation on the SFS (large excess of high frequencies in the local SFS and large peaks in the total SFS) and are thus have the strongest power to reject neutrality.

2.2.3 General properties of test statistics $T^l_{\Omega}$ and $T^t_{\Omega}$

Weights of the tests of neutrality from the local ($\Omega^l_i$) and total SFS ($\Omega^t_i$) under population reconnection are presented in Figure D. As expected from the signature on the local SFS of population reconnection after isolation (Figure C), positive $\Omega^l_i$ values are present around classes $i = 1$ and $i = n - 1$, where $n$ is the sample size in each population. As expected from the signature on the total SFS of population reconnection after isolation (Figure 2), positive $\Omega^t_i$ values are present around classes $i = n$ and $i = 2n$, where $n$ is the sample size in each population.

2.2.4 Power of the test statistics to detect departure from neutrality

In this section, we compute using coalescent simulations the power of the optimal statistics $T^l_{\Omega}$ and $T^t_{\Omega}$ to detect the reconnection of previously isolated populations.
Figure D Weights of the optimal neutrality tests to detect the signature of population reconnection after isolation from (A) the local ($\Omega_l$) and (B) the total SFS ($\Omega_t$), as a function of the class ($i$) of the variant in the SFS. The weights were computed for $d = 3$ populations with samples sizes $n = 16$, with an isolation period $T_{iso} = 6$ and a scaled migration rate $M = 5$ during the connection periods.

Figure E Power of test statistics to detect departure from the constant population size model as a function of time since the reconnection event ($T_{reco}$). Test statistics are either computed from the local SFS (A), or computed from the total SFS (B). Test statistics are Tajima's $D$, Fu and Li's $D^*$ and $F^*$, Fay and Wu's $H$, Zeng et al.'s $E$ and optimal test statistics derived using the method from (Ferretti et al. 2010) (see SI file 1) that detects the signature of population reconnection after isolation ($T_{\Omega_l}$ and $T_{\Omega_t}$) from the local and total SFS, respectively. (A) All statistics have a low power to detect departure from the constant population size model from the local SFS, (B) $T_{\Omega_l}^l$, and to a lesser degree, $H$, have a high power to detect departure from the constant population size model from the total SFS, but high power is restricted in the domain where time since the reconnection event is short. Parameters are $d=3$, $\theta=100$, $n = 16$ sampled genes of 9719bp per population, and $T_{iso} = 6$. Means are over 5,000 replicate simulations. Test $T_{\Omega_l}^l$ is used in Fig. L to detect temporal variations in the SFS of HIV-1 subtypes.
3 Robustness of the signature

In this section, we assess the robustness of the signature of population reconnection after isolation (Figs. 1 and 2) to homoplasy (Fig. F), and to alternative demographic scenarios (Figs. G and H).

Figure F Sensitivity of the total SFS (Fig. 2A) to homoplasy, for different scaled mutation rates ($\theta$). SFS are simulated under a finite sites model and assuming population isolation ($T_{reco} = 0$), where ancestral populations split $2N$ generations ago (using software fastsimcoal (Excoffier and Foll 2011)). Sample sizes are $n_1 = 13$ in the first population and $n_2 = 22$ in the second; the number of sites is 9729 per sequence.

Figure G Effect of population expansion on the signature of reconnection between isolation populations in the distribution of pairwise nucleotide differences (Fig. 1B-C) and the total SFS (Fig. 2A). Simulations consider that a population expansion started with the reconnection event (e.g., due to rapid growth following colonization by a small population). The representation of the total SFS of Nawa and Tajima (2008) is used: the expectation in a single panmictic population is a straight line, and deviations from a straight line indicate an excess of variants at a given frequency. The exponential rate of increase is 5 (in units of $2N$ generations), corresponding to a 150 fold increase in size in the past $2N$ generations. Other parameters are $d = 3$, $\theta = 100$, $T_{iso} = 6$; we consider no recombination. Each plot represent the mean over 1000 replicate coalescent simulations (software ms; Hudson 2002).
Figure H Effect of skewed offspring distribution (beta-coalescent process of parameter $\alpha$) on the signature of reconnection between isolated populations in the total SFS (Fig. 2A). $\alpha = 0$ corresponds to the classic Kingman coalescent (Wright-Fisher model), $\alpha = 1$ corresponds to the Bolthausen-Sznitman coalescent. Coalescence processes follow the beta-coalescent at all times – before isolation, during isolation and after reconnection –. The representation of the total SFS of Nawa and Tajima (2008) is used: the expectation in a single panmictic population is a straight line, and deviations from a straight line indicate an excess of variants at a given frequency. The relative proportion of sites are represented on the Y-axis (i.e., number of sites divided by the total number of segregating sites), to allow the comparison between different $\alpha$ values. Parameters are $d = 3$, $\theta = 100$, $T_{iso} = 6$; we consider no recombination. Each plot represent the mean over 1000 replicate coalescent simulations (software $ms$; Hudson 2002).

The signature of the Isolation with Migration model

In this section, we simulate the signature of the Isolation with Migration (IM) model on the distribution of pairwise nucleotide differences and on the site frequency spectrum. Under the IM model, a panmictic population of size $dN$ split into $d$ populations of size $N$ at time $T_{split}$ in the past, and after the split, the $d$ populations exchange migrants at rate $m$ (so the scaled migration rate is $M = 4Nm$), following an island model of migration. To match the HIV-1 application, we consider that $d = 3$.

Fig. I shows the signature of the IM model. We can see that, although the signature at some time points is similar to that of a reconnection event (Figs. 1 and 2), the dynamics of the signature are different. Indeed, the IM model is similar to a model where migration decreases (from panmixia, to a finite rate $M$), while in the reconnection model we present in the main text, migration increases at time $T_{reco}$. 
Figure I Signature of the Isolation with Migration (IM) model on the distribution of pairwise nucleotide differences (A) and the total SFS (B), as a function of the scaled migration rate $M$. Simulations consider that a population split into 3 populations $T_{split}$ generations ago, and that the scaled migration rate is $M$ between each of the 3 populations. The representation of the total SFS of Nawa and Tajima (2008) is used: the expectation in a single panmictic population is a straight line, and deviations from a straight line indicate an excess of variants at a given frequency. We consider no recombination, a finite sites model with 9719 sites (HIV-1 sequence length) and a scaled mutation rate of $\theta = 100$; each plot represent the mean over 1000 replicate simulations. We can see that the IM model can generate bimodality in the distribution of pairwise nucleotide differences and peaks in the SFS under certain circumstances (weak migration, $M < 1$, and old split, $T_{split} > 1$). Nevertheless, contrary to the signature of a reconnection of isolated populations in which we observe a decrease in peaks size with time, with the IM model the size of the modes in the distribution of pairwise nucleotide differences and the size of the peaks in the SFS increase with time.
4 Detection of past isolation and current reconnection of HIV-1 subtypes

In this section, we illustrate the HIV-1 subtype clusters used in the result section "Detection of past isolation and current reconnection of HIV-1 subtypes" (Fig. J), and assess the sensitivity of the SFS pattern presented Fig. 3A to the choice of outgroup (Fig. K). We also use the test statistic $T_{11}$ to assess the significance of the SFS patterns observed in the case of HIV-1 subtypes from China (Fig. 3(a)-(b)) and South America (Fig. 3(g)-(h)). Finally, we present into three tables the HIV-1 sequences analysed (all subtypes in table A, Chinese subtypes in table B, South American subtypes in table C).

Figure J Phenetic tree of all HIV sequences presented in the results section and Fig. 3, (a) in China and (b) in South America and the corresponding heatmap of pairwise nucleotide differences between sequences. The phenetic tree was obtained using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm on the euclidian genetic distance matrix. The colour in the heatmap correspond to the genetic distances between all pairs of sequences (white for dissimilar sequences, orange for similar sequences). The colors above each column correspond to cluster memberships inferred using Discriminant Analysis of Principal Components –DAPC– (blue for cluster CRF01_AE, green for cluster B, red for cluster C and orange for cluster F1). The grey colors left of each row corresponds to the sampling year of each sequence. We can see that clusters identified using DAPC match very well the phylogenetic clusters and are not associated with the sampling year.
Figure K Sensitivity of the total SFS (Fig. 3(a),(b),(g),(h)) to the outgroup choice for ancestral state, for HIV-1 sequences sampled in China (left panel) and South America (right panel). Each row corresponds to a different outgroup sequence (subtype J –used in the main text–, and subtypes G and K). We can see an excess of variants at high frequency, as expected from ancestral state misinference (Baudry and Depaulis 2003). However, the excess of intermediate frequency variants (“peaks”) is robust to the chosen outgroup, and is independent from ancestral state misinference.
Figure L Bootstrap test of the temporal changes of the Site Frequency Spectrum (SFS) of HIV subtypes (displayed in Fig. 3(a),(b),(g),(h)), captured by test statistic $T_{Ω}$. Results are presented for China (left major panel) and for South America (right major panel). Two time points are considered in China (table B) and in South America (table C). Boxplots in (a) and (b) represent the value of $T_{Ω}$ computed from the total (pooled) SFS for 1000 bootstraps, respectively in China and South America. For each bootstrap, we randomly resampled sequences with replacement from each cluster (16 sequences per subtype cluster in China, 8 sequences in South America), then computed the SFS and the corresponding $T_{Ω}$ value. Histograms in (c) and (d) represent the differences in $T_{Ω}$ values between the two time points, respectively in China and South America. The red dashed lines in (c) and (d) represent the $5\%$ quantiles of these distributions; these quantiles are used to assess the significativity of the temporal changes: a $5\%$ quantile larger than 0 indicates a significant temporal decrease. We can see in (c) that $T_{Ω}$ values significantly decreased in China between 2006-2007 and 2008-2009, while $T_{Ω}$ values decreased overall but non-significantly in South America between 2001-2002 and 2004-2005.
Table A Number and name of HIV-1 pure subtypes and recombinant forms (CRF: circulant recombinant forms; URF: unique recombinant forms) considered for the analysis (Fig. 5). To avoid sampling bias, we excluded redundant samples (from the same patient) and samples with identity larger than 98%. We excluded the gaps in sequences alignment and segregating sites with more than 2 states (due to homoplasy), resulting in 3676 segregating sites. Alignments are available upon request.

<table>
<thead>
<tr>
<th>pure subtypes</th>
<th>CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>619</td>
<td>378</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>URF</th>
<th>BC</th>
<th>BCF1</th>
<th>BF1</th>
<th>01B</th>
<th>01BC</th>
<th>01C</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>1</td>
<td>61</td>
<td>52</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table B Number and name of HIV-1 subtypes and recombinant forms from China sampled between 2001 and 2010 (subset from table S1) considered for the analysis (Figs. 3A-F and 4A).

<table>
<thead>
<tr>
<th>year</th>
<th>pure subtypes</th>
<th>CRF</th>
<th>URF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>C</td>
<td>01_AE</td>
</tr>
<tr>
<td>2005-2007</td>
<td>24</td>
<td>3</td>
<td>91</td>
</tr>
<tr>
<td>2008-2010</td>
<td>15</td>
<td>2</td>
<td>58</td>
</tr>
</tbody>
</table>

Table C Number and name of HIV-1 subtypes and recombinant forms from South America (Brazil and Argentina) sampled between 2001 and 2007 (subset from table S1) considered for the analysis (Figs. 3G-L and 4B).

<table>
<thead>
<tr>
<th>year</th>
<th>pure subtypes</th>
<th>CRF</th>
<th>URF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>C</td>
<td>F1</td>
</tr>
<tr>
<td>2001-2007</td>
<td>58</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>
References

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics 183: 249–258.


