The Genomic Signature of Population Reconnection Following Isolation: From Theory to HIV

Nicolas Alcala,*1 Jeffrey D. Jensen,2 Amalio Telenti,3 and Séverine Vuilleumier*1,2,4
*Department of Ecology and Evolution, University of Lausanne, Lausanne CH-1015, Switzerland, 1Department of Biology, Stanford University, California 94305-5020, 2School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland, 3Genomic Medicine, The J. Craig Venter Institute, 4120 Capricorn Lane, La Jolla, California 92037, and 4**Institute of Microbiology, University Hospital and University of Lausanne, Lausanne CH-1011, Switzerland
ORCID ID: 0000-0002-5961-5064 (N.A.)

ABSTRACT Ease of worldwide travel provides increased opportunities for organisms not only to colonize new environments but also to encounter related but diverged populations. Such events of reconnection and secondary contact of previously isolated populations are widely observed at different time scales. For example, during the quaternary glaciation, sea water level fluctuations caused temporal isolation of populations, often to be followed by secondary contact. At shorter time scales, population isolation and reconnection of viruses are commonly observed, and such events are often associated with epidemics and pandemics. Here, using coalescent theory and simulations, we describe the temporal impact of population reconnection after isolation on nucleotide differences and the site frequency spectrum, as well as common summary statistics of DNA variation. We identify robust genomic signatures of population reconnection after isolation. We utilize our development to infer the recent evolutionary history of human immunodeficiency virus 1 (HIV-1) in Asia and South America, successfully retrieving the successive HIV subtype colonization events in these regions. Our analysis reveals that divergent HIV-1 subtype populations are currently admixing in these regions, suggesting that HIV-1 may be undergoing a process of homogenization, contrary to popular belief.
diverged populations after isolation remains difficult, as models of past isolation with recent contact are challenging to distinguish from models of ancient population separation with a continuous exchange of migrants. Also, parameter inference requires a priori knowledge of a population split (Hey 2010), and the time at which migration event(s) occurred has thus far been difficult to estimate (Sousa et al. 2011; Strasburg and Riesenberg 2011). Nielsen and Wakeley (2001) first proposed a method to infer time of divergence (IM model), the migration rate between the two populations, as well as the relative sizes of the populations. This approach was extended to account for the divergence of multiple populations (Hey 2010) and temporal reduction of gene flow following an initial population split (Wilkinson-Herbots 2012). Though others have discussed the expected signature of a reconnection event following a past isolation period (Beckert and Przeworski 2009; Sousa and Hey 2013), detailed theoretical formalization is still lacking. Besides this, the use of the aforementioned methods for general inference is also challenging as they are difficult to apply when the number of sampled populations is large (Beaumont et al. 2002; Drummond et al. 2005; Gutenkunst et al. 2009; Lobste et al. 2011; Excoffier et al. 2013), and when multiple sampled time-points are available (Excoffier et al. 2013; Foll et al. 2014). Further, they often rely on a specific evolutionary model, such as the Wright-Fisher model (e.g., Gutenkunst et al. 2009; Lobste et al. 2011; Excoffier et al. 2013) which might not be representative of, for example, the skewed offspring distributions found for marine species or viruses (Tellier and Lemaire 2014).

Past population isolation and subsequent reconnection or secondary contact are common in natural populations and can have a drastic impact on species’ evolutionary histories. Anatomically, modern humans are thought to have admixed with Neanderthal populations after a period of isolation between the African and European continents (Green et al. 2010; Patterson et al. 2012; Sankararaman et al. 2012). Similarly, African and non-African populations of *Drosophila melanogaster* also experienced a period of isolation, followed by subsequent admixture (Pool et al. 2012). Domesticated species of both plants and animals are isolated from their wild relatives, and often experience subsequent genetic exchange (e.g., in crops; Ellstrand et al. 1999). Importantly, isolation and subsequent reconnection events are also common features in viral histories, including the Influenza A virus, Human Immunodeficiency Virus (HIV) and Human Cytomegalovirus (HCMV; Renzette et al. 2013), where dependency on the host cell and other features of the life cycle isolate viral populations within hosts, groups of hosts, or host species interbreeding. For example, reassortment between avian and human influenza viruses caused the pandemic outbreaks in 1957 and 1968, and a reassortment of swine, avian and human influenza viruses was responsible for the 2009 pandemic (Hsieh et al. 2006; Garten et al. 2009; Flahault and Zylberman 2010).

As in influenza, events of HIV isolation and reconnection are common and can be observed at different spatial and temporal scales. Indeed, tracing the origin of HIV has confirmed that cross-species transmission of SIVs from other primates to humans, acting as sources of HIV, occurred several times over the past 100 years (Liégeois et al. 2014). The virus remained isolated for a very long period and adapted rapidly in humans, which ultimately rose to epidemic levels (Korber et al. 2001; Tebit and Arts 2011). Circulating HIV populations in humans are strongly divergent. Two major types of HIV exist (HIV-1 and HIV-2) and both have diversified into several groups or subtypes (Robertson et al. 2000). The HIV pandemic is mainly caused by HIV-1 Group M, which is composed by divergent (up to 35% sequence divergence) subtypes (named A–D, F–H, J, and K; Gaschen et al. 2002). The HIV epidemic is associated with the worldwide movements of people that allowed HIV-1 subtype colonization events (Tebit and Arts 2011; Hamelaar et al. 2011). In the course of their evolution in human populations, HIV-1 subtypes experienced several large-scale isolation events within both transmission risk groups (Su et al. 2000; Pang et al. 2012; Han et al. 2013; Li et al. 2013) and countries (Bello et al. 2011; Castro-Nallar et al. 2012). Subsequent reconnection then occurred through co-infection, super-infection, and recombination (Tebit and Arts 2011; Vuliémeier and Bonhoeffer 2015). These reconnection events are the source of emergence of many new recombinant forms which constitute a major challenge for vaccine development and are at the origin of new epidemics (Tebit and Arts 2011; Feng et al. 2013). The multiple colonization events of HIV-1 that occurred in Asia and in South America are representative of this trend (Hamelaa et al. 2011). Indeed, with an estimated 4.8 million people living with HIV as of 2011, Asia is the second most HIV-affected region worldwide (Zhang et al. 2013). Historical epidemics in Asia are associated with transmission events to various risk groups; HIV-1 subtypes B and C were first detected in the mid-1980s and early 1990s, respectively, in the Injecting Drug Users (IDUs) risk group (Lu et al. 2008), and then spread to other populations. Subtype CRF01_AE was first detected in the Men who have Sex with Men ( MSM) risk group in the 1990s, then spread to other groups and became the most prevalent HIV-1 subtype in many parts of China. Similarly, the epidemic clade circulating in South America is derived from subtype B viruses that migrated out of Haiti around 1969 (1966–1972) and spread through the world (Bello et al. 2011). Later, a single-founder subtype F1 strain, introduced in Brazil, spread and recombined with local subtype B viruses to form the HIV-1 BF1 and F1 epidemics (Aulicino et al. 2007).

The present study identifies specific temporal signatures of reconnection of diverged populations in two broadly used summaries of DNA polymorphism: the distribution of pairwise nucleotide differences (via coalescent theory), and the Site Frequency Spectrum (SFS, via simulation). We also describe how past isolation can influence commonly used SFS-based test statistics (Tajima’s D, Fu and Li’s D* and F*, Fay and Wu’s H, Zeng et al.’s E), and discuss the specificity and the robustness of the signature relative to other demographic and selective processes. Then, using theoretical results and full genome polymorphism data, we reconstruct the successive invasions of HIV-1 subtypes in China and in South America and analyze the subsequent dynamics of genome composition. We find extensive admixture between HIV-1 subtypes, potentially leading in the long term to homogenization of HIV-1 genomes in these regions.

**MATERIALS AND METHODS**

Using coalescent theory, we first present methods for detecting the signature of population reconnection after isolation on the distribution of pairwise nucleotide differences. Then, in the second section, we use coalescent simulations to derive the signature of population reconnection after isolation on the SFS. The third section highlights the robustness of these signatures to alternative demographic scenarios. Finally, the fourth section presents a data application for these methods: inference of recent HIV-1 evolution in China and in South America.

**Theoretical signature**

In our model, we consider $d$ previously connected populations that became isolated at time $T_{iso}$ and then reconnected at time $T_{rec}$. The scaled migration rate $M$ between populations (i.e., twice the number of migrant genes per population per generation) is the same before and after the isolation period and mutations follow a Poisson process of rate $\theta/2$ (Kimura 1969). We used an infinite sites model for the analytical investigations, and for the simulations we considered a finite sites model with a number of sites corresponding to HIV-1 genome size (9719bp, from the reference HXB1 sequence). Our developments are based on the Wright-Fisher model (Fisher 1930; Wright 1931), where
each population has a constant size $N$. For analytical simplicity, we considered that times $T_{iso}$ and $T_{rec}$ are scaled in units of $N$ generations, so $T_{iso} = 1$ corresponds to an isolation event that occurred $N$ generations ago. We derived an exact theoretical formula for the distribution of pairwise nucleotide differences in nonrecombining genomic regions (Appendix A), and simulated this distribution using coalescent simulations (software ms; Hudson 2002) in the case of recombining regions. We computed the power to sample sequences from each mode of the distribution using equations A.6a–A.6b (Supporting Information, Figure A in File S1).

We generated the SFS through time following isolation and reconnection events using coalescent simulations. We studied the total SFS (considering pooled samples from all populations); in the main text, we present results for equal sample sizes in each population, and we present results for alternative sampling schemes (Figures B and C in File S1). We also studied the joint SFS between pairs of populations. We assessed the temporal changes of the total SFS using the visual test of Nawa and Tajima (2008), which is based on the difference between the expected SFS under neutrality in a population with constant size, and the observed SFS. For each frequency $i/n$, where $n$ is the sample size, the transformation represents $\theta_i = i\xi$, instead of the corresponding number of variants $\xi_i$. This simple test enables one to easily detect which frequencies present an excess or a deficit of variants. We also derived an optimal test statistic to detect the signature of population reconnection from the SFS (Text A and Figure D in File S1) and compute the power of common neutrality tests to detect such events (Figure E in File S1).

### Robustness analysis of the theoretical signature

We then analyzed the sensitivity of our results to departures from the infinite sites model (Figure F in File S1), from the constant population size model (Figure G in File S1) and from the Wright-Fisher model (Figure H in File S1). To account for repeated bottlenecks (e.g., during transmission events), extinction-recolonization dynamics, and rapid adaptation, we simulated the distribution of pairwise nucleotide differences under a scenario of reconnection of isolated populations under a beta-coalescent model. Beta-coalescent models have recently been proposed to model the genealogies of many organisms with a large number of variants because they can follow the beta-coalescent at all times: before isolation, during isolation and after reconnection. In addition to mutation and migration rates, beta-coalescents rely on a parameter $\alpha$ that gives the probability of multiple lineages coalescing at the same time. We present the results for four different parameter values: $\alpha = 0$ (equivalent to the classic Kingman coalescent), $\alpha = 0.5$, $\alpha = 1$ [corresponding to the Bolthausen-Sznitman coalescent, used to model strong positive selection (e.g., Neher and Hallatschek 2013)], and $\alpha = 1.5$.

We also compared the signature of population reconnection after isolation with that of a closely related model: the Isolation with Migration model [IM (e.g., Takahata 1995; Wakeley 1996; Rosenberg and Feldman 2002)], using coalescent simulations (Figure I in File S1).

### HIV data analysis

We used 1646 whole genome HIV-1 samples from subtypes B, C, F1 and CRF01_AE and recombinant forms between these subtypes (Table A in File S1). We used the subset of genomes from China sequenced between 2005 and 2009 (297 sequences; Table B in File S1). For the study in South America (Brazil and Argentina), we used data available between 1999 and 2005 (128 sequences; Table C in File S1). Finally, for the worldwide PCA analysis, we used all 1646 genomes available.

For the Chinese and South American data, we identified HIV-1 populations (subtype clusters) by Discriminant Analysis of Principal Components (DAPC) (Jombart et al. 2010). Our investigations considered dominant subtypes B, C, CRF01_AE and F1 and their associated recombinants. Each cluster encompassed a dominant subtype and some recombinant forms between subtypes. We found strong support for four groups using the Bayesian Information Criterion (ranging from 2 to 100 PCA axes) that are independent of sampling year (See Figure J in File S1 for a neighbor joining tree of the sequences where the different clusters and sampling years are represented). We then computed the distribution of pairwise nucleotide differences and the total and joint SFS for the major circulating HIV-1 subtypes in China and South America.

### Computing the temporal patterns of genetic variation of HIV subtypes

We computed the distribution of pairwise nucleotide differences, and the total and joint SFS for the major circulating HIV-1 subtypes in China and South America. The distribution of pairwise nucleotide differences within (within-population pairwise differences) and between clusters (between-population pairwise differences) were generated for each nonrecombinant block between subtypes that we identified with ihHMM (Schultz et al. 2009). In total, 19 and 24 blocks were identified in sequences sampled in China and South America, respectively; the 5%, median and 95% quantiles of block length were 111bp, 220bp and 1333bp, respectively. Detection of bimodality in the distribution, the signature of a past isolation event, was performed by fitting a Gaussian Mixture Model using the Expectation-Maximization algorithm from Meng and Rubin (1993) as implemented in the R package mixtools (Benaglia et al. 2009). Gaussian distributions are used to estimate positions of the mode as distributions are symmetric (the 5% and 95% quantiles of the skewness of all components are -1.05 and 1.05, respectively). The algorithm also provides posterior probabilities of membership into modes, which is used to identify genomes that have bimodal distribution (i.e., indicating admixture). We analyzed the signatures in the pairwise nucleotide differences by comparing HIV genome signatures with the expected results, assuming either small linked sequences or large recombining sequences.

The total and joint SFS were computed using subtype J as an outgroup. Results were robust to the chosen outgroup (Figure K in File S1). Note that the distinction between recombining and nonrecombining sequences does not apply for the SFS, as the expectation (and the maximum likelihood estimate) of the SFS is not affected by recombination: only its variance is affected (Gutenkunst et al. 2009). We tested the significance of the temporal changes of the SFS using a bootstrap test (Figure L in File S1).

### Data availability

All data used are from the Los Alamos HIV database (http://www.hiv.lanl.gov/) (Table A in File S1).

### RESULTS

#### Bimodality in pairwise nucleotide differences as a signal of population reconnection after isolation

Two distinct signatures of population reconnection after isolation were found in the temporal distribution of pairwise nucleotide differences: bimodality and increased variance of the first mode (Figure 1). Bimodality was observed in these distributions (both within and between populations, $\pi_w$ and $\pi_h$) only when a population reconnection...
occurred (Figure 1A). This bimodality reflects the different possible origins of the two sequences sampled: the two sequences either have a recent common ancestor (i.e., after the reconnection event; first mode in Figure 1A) or an ancient common ancestor (before the isolation event; second mode in Figure 1A). Bimodality was also observed in the distributions of total pairwise nucleotide difference (considering all populations, \( \pi \)) after a reconnection event but, in this case, bimodality was also observed when populations were completely isolated (\( T_{iso} = 0 \), Figure 1, B and C). To differentiate a connection from an isolation event, bimodality needs to be associated with a temporal increase of the variance of the modes in the distribution (which depends on the recombination rate; see Figure 1). Although testing for increase in bimodality can be challenging, a common approach is to compare the variance of the modes before and after the expected reconnection event, with an increase indicating a connection event.
Population reconnection after isolation can thus be detected by testing for bimodality and a temporal increase of the variance of the modes. To detect bimodality, with high probability (95%) (Figure A in File S1), depends on the sample size, the duration of the reconnection period, and the amount of gene flow among populations (scaled migration rate). As shown in Figure A in File S1, it is easier to detect a reconnection event from the distribution of pairwise nucleotide differences between populations than within populations. Indeed, the second mode in the distribution of pairwise nucleotide differences was larger between population (σ_ν_2) than that of within population (σ_ν_1, Figure 1A). This difference in mode size increases when migration is weak or intermediate (M ≤ 1) and the reconnection event is not recent (Figure A in File S1). Ancient reconnection events can be detected even with small sample sizes (n=10) when migration is moderate or strong (M ≥ 1; Figure A in File S1). To evaluate the temporal increases in variance of the mode, no further development is required as the Bartlett test (Bartlett 1937) can be used.

Thus, we show that distinctive signatures of past isolation and reconnection events can be identified on the distribution of the number of nucleotide differences between pairs of genomes sampled (within, between, or in the total population). We also determine the parameter ranges for which this signature can be detected (sample size, migration rate, and time since reconnection).

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

When a population is at mutation-drift equilibrium, the distribution of its alleles, or SFS, has a characteristic shape (dotted line Figure 2A). When populations experience demographic changes, departures from this characteristic shape are expected and the size and the position of those changes are informative as to the demographic change. Consistent with the expectation, we found a large excess of variants at intermediate frequencies on the SFS following reconnection events between previously isolated populations (Figure 2). This excess was due to exchanges of variants among populations that were once fixed in one or several populations during isolation. The size and the position of this excess is indicative of the number of populations that have reconnected and of the timing of the reconnection event (Figure 2A), and can be identified by the visual test of equilibrium neutrality proposed by Nawa and Tajima (2008). We predict that in the total SFS (SFS obtained with merged samples from all populations), shortly after a reconnection event one would expect narrow, high peaks of specific variants. These peaks are found at frequencies 1/d, where d is the number of sampled populations and 1 ≤ i < d, when populations have equal sample sizes. In Figure 2A, peaks can be seen at frequencies 1/d and 2/d (d = 3). With unbalanced sample sizes, the number of observed peaks can be higher (up to 2d – 2 peaks; see Figure B in File S1).

The temporal patterns in the joint SFS (Figure 2B) are also strongly informative as to the occurrence of reconnection events and as to the amount of gene flow among populations since the reconnection event. When populations are isolated, each has its own segregating variants. This translates, in the joint SFS, into variants being restricted to the axes of the joint SFS [classes (20i), (i,20), with i ranging from 1 to 20 in Figure 2B]. Then, following the reconnection event, gene flow distributes the variants among populations to the point that they occur at similar frequencies in each population and the genetic compositions of populations are fully homogenized. Consequently, in the joint SFS,
several variants move progressively away from the axes toward the diagonal of the joint SFS, as expected in a structured population at equilibrium (last column Figure 2B).

To robustly detect the excess of variants on the SFS generated by a reconnection event, we derived an optimal test statistic, denoted $T_V$, using developments provided by Ferretti et al. (2010) and Achaz (2009) (File S1). Although use of this test is restricted to recent reconnection events that follow long periods of isolation, it performs better than commonly used statistics such as Tajima’s $D$ (Tajima 1989), Fu and Li’s $D^*$ and $F^*$ (Fu and Li 1993), Fay and Wu’s $H$ (Fay and Wu 2000), Zeng et al.’s $E$ (Zeng et al. 2006). Indeed, even with long isolation periods, those statistics have less than 60% power to detect reconnection events (Figure E in File S1). The results presented in this section apply to both recombining and nonrecombining sequences, as the expected SFS is the same in both cases (Gutenkunst et al. 2009). Also, the power estimates from the tests based on nonrecombining sequences are conservative as they correspond to the case with the highest variance (Achaz 2009) and thus the lowest power.

Thus, we show that distinctive signatures of past isolation and reconnection events can be identified in the SFS as well as in the joint SFS. Namely, the SFS has an excess of variants, the size and position of which are indicative of the number of previously isolated populations and of the timing since the reconnection event. The temporal dynamic of the joint SFS is informative as to the level of gene flow between populations.

Figure 3: Signatures of HIV-1 subtype reconnection events on the Site Frequency Spectrum (SFS). Results are presented for China (left major panel) and for South America (right major panel). Two time points of HIV-1 sequences sampled are considered in China (Table B in File S1) and in South America (Table C in File S1) (see method section ‘HIV-1 genome analysis’s). The total (pooled) SFS is presented in the first row following representation of Nawa and Tajima (2008), the joint SFSs of each pairwise group of HIV subtypes are presented in the second, third, and fourth row and, in the fifth row, the Principal Component Analyses (PCAs) are presented. Subtype J is used as an outgroup (see Figure K in File S1, results for alternative outgroups). Original subtypes (subtypes found early in the epidemics in Africa; at the corners) are used to define a triangle in the PCA space. The position of the HIV-1 genomes on the straight edges of the triangle suggests strong admixture between subtypes. For the total and joint SFS, we consider the mean value of 500 random samples of sequences in each subtypes cluster (see Figure J in File S1 for the different clusters). A sample of 16 sequences per subtype cluster is considered in China, and of eight sequences per subtype cluster in South America. The significance of the trends in the dynamics is tested using the optimal test statistic $T_V$ derived in the SI (Figure L in File S1).

112 | N. Alcala et al.
Robustness of identified signatures to alternative demographic scenarios

The signatures of population reconnection identified here, i.e., bimodality and temporal increase of variance of peaks in the SFS, are robust to population expansion (Figure G in File S1), alternative evolutionary and population dynamic models such as repeated bottlenecks, rapid extinction-recolonization dynamics, recurrent positive selection or highly skewed offspring distributions (Figure H in File S1). Moreover, the signature is distinct from the signature of a closely related model: the isolation with Migration model (IM) (Figure I in File S1).

The observed signatures are robust even when a massive population expansion occurs concomitantly with the reconnection event (Figure G in File S1). Such events modify the expected signature only by presenting an excess of low frequency variants compared to what is expected with populations of constant size (as expected from the results of Tajima (Tajima 1989) in a single panmictic population). Replacing the classic Kingman coalescent model (Wright-Fisher model) with a beta-coalescent model does not alter our results for values of the model parameter $\alpha$ (Figure H in File S1). Beta-coalescent models account for large variance in the number of offspring per individual, as expected under repeated bottlenecks, fast extinction-recolonization dynamics, or recurrent beneficial fixations (Neher and Hallatschek 2013; Tellier and Lemaire 2014). Such variance leads to an excess of low frequency variants and smaller peaks in the total SFS compared to Kingman’s coalescent model; this effect is larger for small $\alpha$ values (light-gray lines Figure H in File S1). However, having an excess of variants at low frequency does not alter the identified signature of reconnection on the total SFS (dynamics of the peaks). Also the temporal dynamics of the signature distinguish this pattern from that of an IM model. Indeed under the IM model, the size of the peaks in the SFS increases with time (Figure LB in File S1), whereas the signature of a reconnection of isolated populations is characterized by a decrease in peak size with time (Figure 2A).

Thus, we show that, even though the signature at specific time-points may be confounded with that of other selective or demographic processes, the temporal signature of past isolation and reconnection identified is unlikely to be confounded, and provides a robust signal.

Detection of past isolation and current reconnection of HIV-1 subtypes

We applied the conceptual and theoretical approaches via an analysis of the colonization history of different HIV-1 subtypes in Asia and in South America. Using data from the Los Alamos HIV database, we detected a strong signal of reconnection events between previously isolated HIV-1 subtypes in these populations.

Past isolation between HIV-1 subtypes in China and South America is signaled by the excess of variants, with peaks at intermediate frequencies in the total SFS (Figure 3, A and B and G and H). These peaks reflect the presence of variants from the sampled population that diverged during isolation and were present at high frequencies in each previously isolated subtype population identified with cluster analysis (see Results).

The reconnection of HIV-1 subtype populations in China and South America is indicated by the bimodality in the distribution of pairwise nucleotide differences between-populations (considering sequences of small length Figure 4). We found that a large proportion of the genome of HIV-1 displays bimodality (Figure 4). In China, this proportion was initially low but rapidly increased through time. Between 2006 and 2007, the proportion of bimodality increased fourfold (Figure 4), and increased a further 40% between 2008 and 2009. This suggests that the time series analyzed here captures a burst of HIV-1 subtype admixture. The dynamics of genomic admixture between HIV-1 subtypes in China is also highlighted by the temporal changes in the distribution of pairwise nucleotide differences and the SFS. The peaks formed by the excess variants (resulting from population admixture in the total SFS) decreased in size, and the values of optimal test statistics $T_0$ significantly decreased (Figure L in File S1). We find a significant increase in the variance of the first mode in the distribution of pairwise nucleotide differences between 2005 and 2009 ($p < 0.01$ two-sided Bartlett test). The joint SFSs for the two time points also indicate that the reconnection event and genome admixture is either recent or that genome admixture is limited but ongoing. Indeed, we can observe the excess of variants along the axes of the joint SFS shifting to the diagonal (Figure 3, C and D). In South America, the proportion of bimodality in HIV-1 samples from 1999 is large (60%) and remains approximately
constant from 1999 to 2004, signaling a prior reconnection event. Patterns observed in the joint SFS suggest that HIV-1 subtype genomes are well homogenized overall; further, this process of homogenization seems to have either slowed or ceased. Indeed, no increase of variance in the modes of the distribution of pairwise nucleotide differences was detected (two-sided Bartlett test), and the total SFS shows only small changes between 2001 and 2005, with nonsignificant changes in values of optimal test statistics $T_3$ (Figure L in File S1).

Genome admixture in China and in South America is also visible when the sampled genomes are projected together in a PCA space defined by original subtypes i.e., which originate from the beginning of the HIV epidemic in Africa (Figure 3, E and F and K and L). Several HIV-1 genomes are present on the straight edges of the triangles defined by the ones at the origin. Temporal PCA patterns also suggest homogenization between subtypes B and C in China and between B and F1 in South America, as they move away from the pure subtypes in PCA space.

To provide insights on the generality and global importance of these results, we performed a PCA analysis of worldwide genomes from HIV-1 subtypes B, C, F1, and CRF01_AE (http://www.hiv.lanl.gov). Again we projected them in the PCA with original subtypes B, C, F1, and CRF01_AE as references (Figure 5). Consistent with the observations for Chinese and South American samples, we observed that HIV-1 genomes are positioned on the straight edges of the triangle formed by original subtypes, positioned at the corners, forming a continuum along the axes rather than discrete clusters within. Very few genomes are localized outside the edges of the triangle defined by the 3 original subtypes. This pattern not only signals that most of the circulating subtypes. When the sampled genomes are projected together in a PCA space defined by original subtypes reconnected, a transient excess of genetic diversity is predicted in the resulting population (Alcala et al. 2013; Alcala and Amos 2012). Standard country codes are used: Argentina (AR), Brazil (BR), Botswana (BW), China (CN), Cyprus (CY), Spain (ES), Japan (JP), Thailand (TH), USA (US), Vietnam (VN), South Africa (ZA). Misrepresented sequences [squared cosine of the PC plane with the observation < 0.05 (Abdi and Williams 2010)] were excluded.

Figure 5 Projection of worldwide distributed HIV-1 sequences from subtypes B, C, CRF01_AE and F1, and URF and CRF recombinant forms between these subtypes. 1646 genome sequences (see Table A in File S1). (A) The first two axes of the principal component analysis (PC1 and PC2), and (B) the second and third axes of the principal component analysis (PC2 and PC3) defined by pure subtypes B, C, CRF01_AE and F1 (see method section ‘HIV-1 genome analysis’). The percentage of variance explained by each PC is indicated between parentheses. The position of a HIV-1 genome sequence along the axes of the triangle formed by three pure subtypes reflect admixture (proportion) between the subtypes.

DISCUSSION

We present an analytical framework to understand the evolutionary histories of organisms that are characterized by periods of isolation. Our theoretical investigations identify specific genomic signatures of past population isolation. We show, for example, that the number, size, and dynamics of modes in the distributions of the pairwise nucleotide differences in populations are informative on the existence of past isolation. Our results extend previous work on the distribution of pairwise coalescence times and pairwise nucleotide differences under a model of population split with or without migration (IM model). We show, for example, that under the reconnection model, the distribution of pairwise nucleotide differences has a temporal evolution which is distinct from the one expected under an IM model (Takahata 1995; Wakeley 1996; Rosenberg and Feldman 2002): the distribution is first bimodal and then tends to become unimodal with time. We also found that in the Site Frequency Spectrum (SFS), the presence, the position and the value variants in excess are informative as to the numbers of previously isolated populations, and the joint SFS characterizes the amount of genetic overlap between populations. Signatures identified are robust to the presence of recombination, extinction-colonization processes, population bottlenecks, and expansion as well as skewed offspring distributions. Our developments add to the current literature in the field as we demonstrate that population reconnection after isolation is not fully captured by commonly used test statistics based on SFS classes of variants at high and/or low frequencies such as $D$, $H$, and $E$ (Ferretti et al. 2010). Taken together, this provides critical information on the evolution of the genomic composition of populations.

We apply our theory in order to analyze the genome signature of successive colonization of HIV-1 subtypes in Asia and South America. In both regions, we reconstruct the successive colonization of HIV-1 subtypes, and find that they are followed by strong HIV-1 subtype admixture. We find, in these regions, that HIV diversification is followed by a temporal homogenization of HIV-1 genome composition. Our analysis highlights a large degree of admixture between circulating HIV-1 subtype genomes in China and South America, and most probably worldwide. These results suggest that genomic exchanges may indeed be relatively strong, and that HIV-1 is currently in a process of homogenization subsequent to isolation and diversification. HIV genome homogenization may seem to contradict the idea that HIV is undergoing a diversification process (Hamelaar et al. 2011; Tebit and Arts 2011; Feng et al. 2013). However, this homogenization is consistent with observed diversity in the HIV-1 population and the increased number of new HIV recombinant forms described worldwide (Hamelaar et al. 2011; Tebit and Arts 2011; Feng et al. 2013). Indeed, when isolated and diverged populations reconnect, a transient excess of genetic diversity is predicted in the resulting population (Alcala et al. 2013; Alcala and Amos 2012). Standard country codes are used: Argentina (AR), Brazil (BR), Botswana (BW), China (CN), Cyprus (CY), Spain (ES), Japan (JP), Thailand (TH), USA (US), Vietnam (VN), South Africa (ZA). Misrepresented sequences [squared cosine of the PC plane with the observation < 0.05 (Abdi and Williams 2010)] were excluded.
Vuilleumier 2014) as all the genetic material that accumulated during the isolation phase is shared among populations. However, this excess of genetic diversity is transient and with time, the diversity decreases (Alcala et al. 2013). In HIV-1, such processes occur through multiple recombination events among subtypes through coinfection or superinfection and are the source of novel recombinant forms. In the long term, successive recombination events will ultimately lead to the homogenization of the genomic composition of HIV-1. Homogenization might be further enhanced by selective forces imposed by the immune response or treatment acting on the HIV genome (Fellay et al. 2007; Pelak et al. 2011; Chikata et al. 2014). As many world regions are currently being newly colonized by different HIV-1 subtypes, it is expected that new recombinants may still emerge and HIV diversity will further increase. However, as demonstrated here, this is expected to be a transient state toward global homogenization.

We find signatures of recombination events among populations of HIV-1 subtypes that were previously isolated both in China and South America. The signatures on HIV-1 DNA polymorphism suggest recombination events of three previously isolated HIV-1 populations in both regions. These results align with the known epidemiologic history of HIV-1 in the two regions (Su et al. 2000; Yang et al. 2002; Aulicino et al. 2007; Lu et al. 2008; Takebe et al. 2010; Bello et al. 2011; An et al. 2012; Pang et al. 2012; Han et al. 2013; Feng et al. 2013) and are supported by recently identified recombinants in China (Feng et al. 2014; Wang et al. 2014; Wei et al. 2014; Hsi et al. 2014). To confirm this, we identified the genome sequences that signal the recombination event. As expected, they reflect first the HIV outbreak in China in 1989: colonization of B and C subtypes (variance of the first peak in the SFS is generated by B and C subtypes). The subsequent colonization of CRF01_AE at the origin of the HIV outbreak in the late 1990s is also signaled in the SFS (a high proportion of CRF01_AE in the highest mode). We obtained similar concordance for the origin of the sequences forming the peaks and the successive subtypes’ colonization events in South America.

Our methods also appear robust to high mutation rates and the presence of selection. Indeed, we detect past HIV-1 history despite its extraordinary defining features: HIV has a small genome (about 9.8 kb in length), a short generation time (24 hr; Mohammedi et al. 2013), and a high mutation rate (about 0.002 substitutions/site/year; Korber et al. 2000; Schlub et al. 2014). It is also under strong selective pressure from the immune system (e.g., Snoeck et al. 2011) and undergoes frequent bottlenecks followed by population expansion [during transmission (e.g., Keele et al. 2008)]. Nevertheless, the signature observed here in the distributions of pairwise nucleotide difference and on SFS only slightly deviates from theoretical expectations. The robustness of these results may have a simple explanation: a separation of time-scales of the processes involved. Selection and population demographic changes drive divergence of populations during isolation. Migration and recombination during population reconnection allows a rapid – relative to divergence – genome admixture. Both impact genome polymorphism, but the latter does so more drastically and at much shorter time-scales.

The evolutionary history of an organism is a powerful resource for understanding its current evolutionary trajectory. The analytical framework developed can be applied to understanding the potential evolutionary pathway of many pathogens and other species that are characterized by periods of isolation. These events are common features of pathogens (Karlsson et al. 2014), but also occur in numerous other species, including humans and domesticated plants and animals (Ellstrand et al. 1999; Lecis et al. 2006; Caicedo et al. 2007; Green et al. 2010; Patterson et al. 2012). Reconnection following a period of isolation is increasingly recognized as an effective mechanism to trigger fast and complex adaptation (Seehausen 2002, 2004; Aulicino et al. 2013). Following reconnection, accumulated diversity (from previously isolated populations) is shared, and the resulting high level of genetic diversity may increase the potential for adaptation (Alcala and Vuilleumier 2014). Admixture among divergent populations may also promote complex adaptation by bringing together new functions acquired in different environments that have evolved independently but are compatible with the genome. Such processes have been observed in influenza and in HIV, for example. Indeed, recombination between strains previously isolated in their nonhuman hosts preceded the HIV epidemic in humans (Bailes et al. 2003; Sharp and Hahn 2010; Vuilleumier and Bonhoeffer 2015). Such events are expected to be more frequent in the future due to the increase in human mobility (Karlsson et al. 2014). Finally, our analytical framework could also be applied to evolutionary trajectories of other zoonotic viruses, such as Ebola virus, coronaviruses, hantaviruses, or henipaviruses.

It is important to keep in mind that viral population isolation and reconnection events can occur at a much smaller time scale within hosts. There is increasing evidence suggesting that genetic compartmentalization of viral populations occurs in different organs, tissues, or cells, and might play a critical role for virus evolution and more importantly for its control. This has been demonstrated for HIV (Schnell et al. 2011; Buzón et al. 2011) as well as human cytomegalovirus (Renzette et al. 2011, 2013, 2014, 2015). Applying the analytical framework detailed here could provide a powerful tool to dissect the dynamics and genetic composition of virus across different genetic compartments, in addition to those dynamics between hosts.

ACKNOWLEDGMENTS

We thank the editor Stephen Wright, John Wakeley, and three anonymous reviewers, whose comments and suggestions greatly improved the manuscript. We would also like to thank Kristen Irwin and Nicholas Renzette for careful reading of the manuscript. This project was funded by the Swiss National Science Foundation (SNSF) grants PZ00P3_139421 and 31003A-130065 and FMPDP3_158381 (S.V.), 31003A_1497241 (A.T.) and an interdisciplinary grant from the Faculty of Biology and Medicine (FBM) of the University of Lausanne (S.V. and A.T.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

LITERATURE CITED

Abdi, H., and L. J. Williams, 2010 Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2: 433–459.

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

Aulicino, P. C., G. Bello, C. Rocco, H. Romero, A. Mangano et al., 2007 Description of the Hybrid origin of SIV in chimpanzees. Science 300: 1713.

Aulicino, P. C., G. Bello, C. Rocco, H. Romero, A. Mangano et al., 2007 Description of the first full-length HIV type 1 subtype f1 strain in argentina: implications for the origin and dispersal of this subtype in South America. AIDS Res. Hum. Retroviruses 23: 1176–1182.

Bailes, E., F. Gao, F. Bibollet-Ruche, V. Courgnaud, M. Peeters et al., 2003 Hybrid origin of SIV in chimpanzees. Science 300: 1713.

Bartlett, M. S., 1937 Properties of sufficiency and statistical tests. Proc. R. Soc. Lond. A. Math. Phys. Sci. 160: 268–282.
Naduvilezath, L. L. E. Rose, and D. Metzler, 2011 Jaatha: a fast composite-likelihood approach to estimate demographic parameters. Mol. Ecol. 20: 2709–2723.

Nawa, N., and F. Tajima, 2008 Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. Genes Genet. Syst. 83: 353–360.

Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proc. Natl. Acad. Sci. USA 110: 437–442.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a markov chain monte carlo approach. Genetics 158: 885–896.

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29: 59–75.

Pang, W., C. Zhang, L. Duo, Y.-H. Zhou, Z.-H. Yao et al., 2012 Extensive and complex HIV-1 recombination between b, c and crf01_ae among Idus in south-east Asia. AIDS 26: 1121–1129.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland et al., 2012 Ancient admixture in human history. Genetics 192: 1065–1093.

Pelak, K., A. C. Need, J. Fellay, K. V. Shianna, S. Feng et al., 2011 Copy number variation of kis genes influences HIV-1 control. PLoS Biol. 9: e1001208.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno et al., 2012 Population genomics of sub-saharan Drosophila melanogaster: African diversity and non-african admixture. PLoS Genet. 8: e1003080.

Reich, D., N. Patterson, D. Campbell, A. Tandon, S. Mazieres et al., 2012 Reconstructing native american population history. Nature 488: 340–374.

Renzette, N., B. Bhattacharjee, J. D. Jensen, L. Gibson, and T. F. Kowalik, 2011 Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. PLoS Pathog. 7: e1001344.

Renzette, N., L. Gibson, B. Bhattacharjee, D. Fisher, M. R. Schleiss et al., 2013 Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. PLoS Genet. 9: e1003735.

Renzette, N., L. Gibson, J. D. Jensen, and T. F. Kowalik, 2014 Human cytomegalovirus intrahost evolution—a new avenue for understanding and controlling herpesvirus infections. Curr. Opin. Virol. 8: 109–115.

Renzette, N., C. Pokalyuk, L. Gibson, B. Bhattacharjee, M. R. Schleiss et al., 2015 Limits and patterns of cytomegalovirus genomic diversity in humans. Proc. Natl. Acad. Sci. USA 112: E4120–E4128.

Robertson, D., J. Anderson, J. Bradac, J. Carr, B. Foley et al., 2000 HIV-1 nomenclature proposal. Science 288: 55–56.

Rosenberg, N. A., and M. W. Feldman, 2002 The relationship between coalescence times and population divergence times, pp. 130–164 in Modern Developments in Theoretical Population Genetics: The legacy of Gustave Malécot, edited by Slatkin, M. W., and M. Veuille. Oxford University Press, Oxford.

Sankaramanan, S., N. Patterson, H. Li, S. Pääbo, and D. Reich, 2012 The date of interbreeding between neandertals and modern humans. PLoS Genet. 8: e1002947.

Schlab, T. E., A. J. Grimm, R. P. Smyth, D. Cromer, A. Chopra et al., 2014 Fifteen to twenty percent of HIV substitution mutations are associated with recombination. J. Virol. 88: 3837–3849.

Schneider, S., and L. Excoffier, 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics 152: 1079–1089.

Schnell, G., S. Joseph, S. Spudich, R. W. Price, and R. Swanson, 2011 HIV-1 replication in the central nervous system occurs in two distinct cell types. PLoS Pathog. 7: e1002286.

Schultz, A.-K., M. Zhang, I. Bulla, T. Leitner, B. Korber et al., 2009 jphmm: improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res. 37: W647–W651.

Seehausen, O., 2002 Patterns in fish radiation are compatible with pleis-tocene desiccation of lake victoria and 14,600 year history for its cichlid species flock. Proc. Biol. Sci. 269: 491–497.

Seehausen, O., 2004 Hybridization and adaptive radiation. Trends Ecol. Evol. 19: 198–207.

Sharp, P. M., and B. H. Hahn, 2010 The evolution of HIV-1 and the origin of AIDS. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 2487–2494.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413–429.

Slatkin, M., 1996 Gene genealogies within mutant allelic classes. Genetics 144: 579–587.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129: 555–562.

Snoeck, J., J. Fellay, I. Bartha, D. C. Douek, and A. Telenti, 2011 Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology 8: 87.

Soua, V., and J. Hey, 2013 Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14: 404–414.

Soua, V. C., A. Grelaud, and J. Hey, 2011 On the nonidentifiability of migration time estimates in isolation with migration models. Mol. Ecol. 20: 3956–3962.

Strasburg, J. L., and L. H. Rieseberg, 2011 Interpreting the estimated timing of migration events between hybridizing species. Mol. Ecol. 20: 2353–2366.

Strimmer, K., and O. G. Pybus, 2001 Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. 18: 2298–2305.

Su, L., M. Graf, Y. Zhang, H. von Briesen, H. Xing et al., 2000 Characterization of a virtually full-length human immunodeficiency virus type I genome of a prevalent intersubtype (c/b) recombinant strain in China. J. Virol. 74: 11367–11376.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Takahata, N., 1995 A genetic perspective on the origin and history of humans. Annu. Rev. Ecol. Syst. 26: 343–372.

Takebe, Y., H. Liao, S. Hase, R. Uenishi, Y. Li et al., 2010 Reconstructing the epidemic history of HIV-1 circulating recombinant forms crf07_bc and crf08_bc in east asia: the relevance of genetic diversity and phylodynamics for vaccine strategies. Vaccine 28: B39–B44.

Tebit, D. M., and E. J. Arts, 2011 Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect. Dis. 11: 45–56.

Tellier, A., and C. Lemaire, 2014 Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Mol. Ecol. 23: 2637–2652.

Vuilleumier, S., and S. Bonhoeffer, 2015 Contribution of recombination to the evolutionary history of HIV. Curr. Opin. HIV AIDS 10: 77–127.

Wakeley, J., 1996 Pairwise differences under a general model of population subdivision. J. Genet. 75: 81–89.

Wakeley, J., 1999 Nonequilibrium migration in human history. Genetics 153: 1863–1871.

Wang, N., H. Wei, R. Xiong, H. Zhang, J. H. Hsi et al., 2014 Near full-length genome characterization of a new crf01_ae/crf08_bc recombinant transmitted between a heterosexual couple in Guangxi, China. AIDS Res. Hum. Retroviruses 30: 380–383.

Wilkinson-Herbots, K. and O. G. Pybus, 2001 Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. 18: 2298–2305.

Communicating editor: S. I. Wright
APPENDIX

Derivation of the Theoretical Signature of Population Reconnection on Pairwise Nucleotide Differences

Derivation of the distribution of pairwise coalescence times: We here derive the distribution of within- and between-population pairwise coalescence times, \( T_w \) and \( T_b \), assuming a long isolation period \((T_{iso} - T_{reco} > 3)\) and small sequences, i.e., without recombination. These coalescence times \( T_w \) and \( T_b \) correspond to the time to the most recent common ancestor of two lineages sampled in the same and in different populations, respectively, scaled by the number of genes per population \( N \).

The coalescence of two genes under the finite island model can be described by a three states continuous time Markov chain (Notohara 1990): the two genes can be present (1) in the same population, (2) in different populations or (3) be coalesced. The transition probabilities from one state to another depend on the scaled migration rate \( M \) between populations and on the number of populations \( d \). Time \( T \) is counted in units of \( N \) generations. The corresponding Markov chain can be summarized by the following transition rate matrix (Notohara 1990):

\[
Q = \begin{pmatrix}
-M - 1 & M & 1 \\
M/(d-1) & -M/(d-1) & 0 \\
0 & 0 & 0
\end{pmatrix}
\]  

(A.1)

The probability that two genes were in state (1), (2) or (3) \( T \) generations ago (row vector \( P_T \)), given that they are in state (1), (2) or (3) at the current generation is given by \( P_T = P_0 e^{QT} \) (\( P_0 \) at generation 0), where \( e^{QT} \) is a \( 3 \times 3 \) matrix. The elements of matrix \( e^{QT} \) are denoted \( q_{ij}(T) \) and represent the probability that a gene in state \( i \) at the current generation was in state \( j \) \( T \) generations ago \((cumulative \ distribution \ function)\). Thus, the probability distribution \( f_{T_w}(T) \) and \( f_{T_b}(T) \) of coalescence of two genes \( T \) generations ago from within- and between-populations, respectively.

\[
f_{T_w}(T) = \begin{cases}
f_w.coal(T) e^{-(T-T_{reco})} , & T < T_{reco} \\
f_1(T_{reco}) e^{-(T-T_{reco})} , & T_{reco} < T < T_{iso} \\
q_1(T_{reco}) e^{-(T-T_{iso})} f_w.coal(T-T_{iso}) + q_2(T_{reco}) f_b.coal(T-T_{iso}) , & T_{iso} < T,
\end{cases}
\]

(A.2a)

\[
f_{T_b}(T) = \begin{cases}
f_b.coal(T) e^{-(T-T_{reco})} , & T < T_{reco} \\
q_2(T_{reco}) e^{-(T-T_{reco})} f_w.coal(T-T_{iso}) + q_2(T_{reco}) f_b.coal(T-T_{iso}) , & T_{iso} < T,
\end{cases}
\]

(A.2b)

These distributions have a simple interpretation. In the most recent period, the reconnection period, \( T < T_{reco} \) \( f_{T_w}(T) \) and \( f_{T_b}(T) \) correspond to the probability of coalescence of 2 lineages under the finite island model, \( f_w.coal(T) \) and \( f_b.coal(T) \). During the past isolation period, \( T_{reco} < T < T_{iso} \), two lineages can only coalesce if they are in the same population after \( T_{reco} \), two lineages can only coalesce if they were in the same population after isolation (probabilities \( q_1(T_{reco}) \) and \( q_2(T_{reco}) \) within- and between-populations, respectively). Their probability of coalescence at generation \( T \) in an isolated population is \( e^{-(T-T_{reco})} \). During the older connection period, \( T > T_{iso} \), two genes can coalesce if they were in the same population during isolation (probabilities \( q_1(T_{reco}) \) and \( q_2(T_{reco}) \), within- and between-populations, respectively) and if they did not coalesce during isolation (probability \( e^{-(T-T_{iso})} \)). Their probability of coalescence at generation \( T \) is \( f_w.coal(T-T_{iso}) \). Alternatively, they could have coalesced if they were in different populations during isolation (probabilities \( q_2(T_{reco}) \) and \( q_2(T_{reco}) \)). Their probability of coalescence at generation \( T \) is \( f_b.coal(T-T_{iso}) \).

Functions \( f_w.coal(T) \), \( f_b.coal(T) \), \( q_1(T) \), \( q_1(T) \), \( q_1(T) \) and \( q_2(T) \) all depend on the eigenvalues of matrix \( Q \),

\[
\lambda_1 = -\frac{1}{2} \left( 1 + \sqrt{1 + \frac{d}{d-1}M} \right) \quad \lambda_2 = -\frac{1}{2} \left( 1 + \sqrt{1 - \frac{d}{d-1}M} \right),
\]

where \( \Delta = (d-1 + dM)|T-4(d-1)M \). We obtain:

\[
f_w.coal(T) = A_w.1 \lambda_1 e^{\lambda_1 T} + A_w.2 \lambda_2 e^{\lambda_2 T}
\]

(A.3a)

\[
f_b.coal(T) = A_b.1 \lambda_1 e^{\lambda_1 T} + A_b.2 \lambda_2 e^{\lambda_2 T}
\]

(A.3b)
Where \( A_{w,1} = \frac{1 + \lambda_2}{\lambda_2 - \lambda_1}, \ A_{w,2} = -\frac{(1 + \lambda_1)}{\lambda_2 - \lambda_1}, \ A_{b,1} = \frac{\lambda_2}{\lambda_2 - \lambda_1}, \ A_{b,2} = -\frac{\lambda_1}{\lambda_2 - \lambda_1} \) and \( B = \frac{M}{\lambda_2 - \lambda_1} \).

**Derivation of the distribution of pairwise nucleotide differences:** From equations A.2a and A.2b, we can derive the distributions of within- (\( \pi_w \)) and between-population pairwise nucleotide differences (\( \pi_b \)). Given that mutations follow a Poisson process of mean \( \theta \), we have \( P(\pi_w = k/G) = \int_0^\infty e^{-\theta T} \frac{\theta^k}{k!} f_{\pi_w}(T)dT \) and \( P(\pi_b = k) = \int_0^\infty e^{-\theta T} \frac{\theta^k}{k!} f_{\pi_b}(T)dT \), where \( k \) is a positive integer and \( G \) is the length of the sequence (under the infinite size model, we assume that \( G \) is much larger than the number of segregating sites). We can decompose \( P(\pi_w = k) \) into a sum of smaller integrals (respectively for \( P(\pi_b = k) \)):

\[
P(\pi_w = \frac{k}{G}) = \int_0^\infty e^{-\theta T} \frac{\theta^k}{k!} f_{\pi_w}(T)dT + \int_{T_{su}}^\infty e^{-\theta T} \frac{\theta^k}{k!} f_{\pi_w}(T)dT + \int_0^\infty e^{-\theta T} \frac{\theta^k}{k!} f_{\pi_b}(T)dT.
\]

(A.4)

Substituting the value of \( f_{\pi_w}(T) \) in each term of equation A.4 by the corresponding value in equation A.2a and using the result \( f_{\pi_w}(T) = e^{-\theta T} \lambda^k \frac{\theta^k}{k!} \), we obtain:

\[
P(\pi_w = \frac{k}{G}) = \sum_{i=0}^{2} A_{w,i} \left( \frac{-\theta}{\lambda_i} \right)^k \frac{1}{k!} \sum_{j=0}^{k} \frac{T_{su}^j}{j!} \left( -\lambda_i + \theta \right)^j
\]

(A.5)

The resulting distributions of pairwise nucleotide differences (from equation A.5) are presented in Figure 1A. Further assuming that \( T_{iso} - T_{reco} > 3 \) (thus \( e^{-T_{su} + T_{reco} = 0} \)) leads to:

\[
P(\pi_w = \frac{k}{G}) = \sum_{i=0}^{2} A_{w,i} \left( \frac{-\theta}{\lambda_i} \right)^k \frac{1}{k!} \sum_{j=0}^{k} \frac{T_{su}^j}{j!} \left( -\lambda_i + \theta \right)^j + q_{1,1}(T_{reco}) e^{-T_{su} + T_{reco}} A_{w,i} + q_{1,2}(T_{reco}) A_{b,i}
\]

(A.6a)

\[
P(\pi_b = \frac{k}{G}) = \sum_{i=0}^{2} A_{b,i} \left( \frac{-\theta}{\lambda_i} \right)^k \frac{1}{k!} \sum_{j=0}^{k} \frac{T_{iso}^j}{j!} \left( -\lambda_i + \theta \right)^j + q_{2,1}(T_{reco}) \left( \frac{-\theta}{\lambda_i} \right)^k \frac{1}{k!} \sum_{j=0}^{k} \frac{T_{su}^j}{j!} \left( -\lambda_i + \theta \right)^j + q_{2,2}(T_{reco})
\]

(A.6b)

Terms in equation A.6a (resp. A.6b) have important interpretations in terms of coalescent events. The first term is the contribution of sequences which coalesce during the recent connection period; indeed, \( \sum_{i=0}^{2} A_{w,i} \left( \frac{-\theta}{\lambda_i} \right)^k \frac{1}{k!} \) (resp. \( \sum_{i=0}^{2} A_{b,i} \left( \frac{-\theta}{\lambda_i} \right)^k \frac{1}{k!} \)) corresponds to the distribution of \( \pi_w \).
(resp. $\pi_b$) under the finite island model at equilibrium, while terms \( 1 - e^{(\lambda - \theta)T_{\text{reco}}} \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \) are due to the truncation of the distribution between $T_{\text{reco}}$ and infinity. The second term corresponds to sequences that coalesce during the isolation period. Indeed, with probability $q_{1,1}(T_{\text{reco}})$ (resp. $q_{2,1}(T_{\text{reco}})$), sequences did not coalesce during the reconnection period and were in the same population during isolation; \( \frac{\theta^k}{(1+\theta)^{k+1}} \) correspond to the distribution of $\pi_n$ in an isolated population, and the term $e^{-\theta T_{\text{iso}}} \sum_{k=0}^{\infty} \frac{T_{\text{iso}}(1+\theta)^k}{k!}$ is due to the truncation between 0 and $T_{\text{iso}}$. Finally, the third term corresponds to sequences that coalesce during the prior connection period. With probability $q_{1,2}(T_{\text{reco}})$ (resp. $q_{2,2}(T_{\text{reco}})$), sequences coalesce during the prior connection period (as they did not coalesce during the reconnection period and were in different populations during isolation). Then, $\sum_{i=0}^{n} A_{b,m} \frac{(-\theta)^i}{(1-k)^{k+1}}$ corresponds to the distribution of $\pi_b$ under the finite island model at equilibrium, while terms $e^{-\theta T_{\text{iso}}} \sum_{k=0}^{\infty} \frac{T_{\text{iso}}(1+\theta)^k}{k!}$ are due to the truncation of the distribution between 0 and $T_{\text{iso}}$.

**Derivation of the power to detect bimodality:** We can derive the power to detect bimodality from the distributions of $\pi_n$ and $\pi_b$ by computing the probability that, in a sample of $n$ independent pairwise differences sequences, at least one is from the first mode and at least one is from the second. Given that the probability for a difference $\pi_n$ to be drawn from the second mode of the distribution is $q_{1,2}(T_{\text{reco}})$, the probability that among $n$ differences, at least one is from the first mode and one from the second mode is $1 - q_{1,2}(T_{\text{reco}})^n - (1 - q_{1,2}(T_{\text{reco}}))^n$. Similarly, the probability to observe a bimodal distribution of $\pi_b$ from a sample of size $n$ is $1 - q_{2,2}(T_{\text{reco}})^n - (1 - q_{2,2}(T_{\text{reco}}))^n$. These results are used to draw Figure A in File S1. Note that with $n$ sequences, we can only compute $n - 1$ independent pairwise differences, although the total number of pairwise differences we can compute is \( \frac{n(n-1)}{2} \).