G3: Genes|Genomes|Genetics Early Online, published on November 6, 2015 as doi:10.1534/g3.115.024208

The genomic signature of population reconnection following isolation: From theory to HIV

Nicolas Alcala[*,§,1], **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

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 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.

KEYWORDS
Admixture; migration; coalescent; site frequency spectrum; HIV

INTRODUCTION

Past population demographic events leave distinctive signatures in DNA sequence polymorphisms and gene genealogies (Tajima 1989; Simonsen *et al.* 1995; Schneider and Excoffier 1999; Excoffier 2004; Zeng *et al.* 2006; Gutenkunst *et al.* 2009; Naduvilezhath *et al.* 2011). For example, a population bottleneck increases the variance of coalescence times and reduces the number of variants at low frequency in the site frequency spectrum (SFS) (Tajima 1989). Population expansion creates star-shaped genealogies and increases the mean number of variants at low frequency in the SFS (Tajima 1989; Slatkin 1996). Population sub-division leads to long internal branches in the genealogy and increases the number of fixed variants (Harpending *et al.* 1998). It also results in "structured" genealogies in which the means and variances of coalescence times (e.g., time to the most recent common ancestor) are large, resulting in an excess of intermediate frequency variants in the SFS (Wakeley 1999). Balancing selection leads to a similar structure, and to an excess of variants at intermediate frequency (Tajima 1989; Fay and Wu 2000; Barton and Etheridge 2004; Zeng *et al.* 2006). Directional selection leads to genealogies with both star shapes (leading to the common ancestor of the selected lineages) and long branches (due to remaining ancestral lineages and/or recombination with the selected haplotype), and generates an excess of variants at both low and high frequency (Barton 1998).

Consequently, DNA sequence polymorphisms are increasingly used to infer past demographic events and have successfully reconstructed past histories of many organisms, including humans (Gravel *et al.* 2011; Reich *et al.* 2012). Some of the first methods proposed were developed to infer population growth rates (Slatkin and Hudson 1991) or for specific demographic scenarios (e.g., Nielsen and Wakeley 2001; Hey 2010). Recently developed approaches allow for the inference of a wide-range of demographic parameters (e.g., Strimmer and Pybus 2001; Beaumont *et al.* 2002; Drummond *et al.* 2005; Gutenkunst *et al.* 2009; Excoffier *et al.* 2013). However, inferring reconnection of diverged populations after isolation remains difficult, as models of past isolation with re-

Manuscript compiled: Wednesday 4th November, 2015%
[1]Department of Biology, Stanford University, Stanford, CA 94305-5020, USA. email: nalcala@stanford.edu

cent 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 Rieseberg 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 (Becquet 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; Lohse *et al.* 2011; Excoffier *et al.* 2013), and when multiple sampled time-points are available (but see 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; Lohse *et al.* 2011; Excoffier *et al.* 2013) which might not be representative of, for example, the skewed offspring distributions found 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 Neandertal 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, often to 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 in-between infections. 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 *et al.* 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 subtypes 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, superinfection, and recombination (Tebit and Arts 2011; Vuilleumier and Bonhoeffer 2015). Those 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 (Hamelaar *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: first, 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 analyse 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.

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_{reco}. 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 use an infinite sites model for the analytical investigations and for the simulations we consider 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 consider that times T_{iso} and T_{reco} are scaled in units of N generations, so $T_{iso} = 1$ corresponds to an isolation event which occurred N generations ago. We derive an exact theoretical formula for the distribution of pairwise nucleotide differences in non-recombining genomic regions (Appendix A), and simulate this distribution using coalescent simulations (software *ms*; Hudson 2002) in the case of recombining regions. We compute the power to sample sequences from each mode of the distribution using eqs. A.6a-A.6b (Fig. A in File S1).

We generate the SFS through time following isolation and reconnection events using coalescent simulations. We study the *total SFS* (considering pooled samples from all populations); in the main text, we present results for equal sample sizes in each populations, and we present results for alternative sampling schemes in the SI (Fig. B and C in File S1). We also study the joint SFS between pairs of populations. We assess the temporal changes of the total SFS using the visual test of Nawa and Tajima (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 $\hat{\theta}_i = i\xi_i$, instead of the corresponding number of variants ξ_i. This simple test enables one to easily detect which frequencies present an excess or a deficit of variants. We also derive an optimal test statistic to detect the signature of population reconnection from the SFS (Text A and Fig. D in File S1) and compute the power of common neutrality tests to detect such events (Fig. 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 (Fig. F in File S1), from the constant population size model (Fig. G in File S1) and from the Wright-Fisher model (Fig. H in File S1). To account for repeated bottlenecks (e.g., during transmission event), 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 variance in the number of offspring per individual (Birkner and Blath 2008; Tellier and Lemaire 2014), in particular viruses (Neher and Hallatschek 2013). We extended the algorithm from Birkner and Blath (Birkner and Blath 2008) – which generates samples under a beta-coalescent at equilibrium – to include an island model of migration (each lineage migrates to another population at rate M) and an isolation period. We assume that coalescence processes 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 α ($0 \leq \alpha \leq 2$) 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; see e.g. Takahata 1995; Wakeley 1996; Rosenberg and Feldman 2002), using coalescent simulations (Fig. I in File S1).

HIV data analysis

We use 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 use 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 consider dominant subtypes B, C, CRF01_AE and F1 and their associated recombinants. Each cluster encompasses a dominant subtype and some recombinant forms between subtypes. We find strong support for four groups using the Bayesian Information Criterion (ranging from 2 to 100 PCA axes) that are independent of sampling year (See Fig. J in File S1 for a neighbour 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 cluster (within-population pairwise differences) and between clusters (between-population pairwise differences) are generated for each non-recombinant block between subtypes that we identified with *jpHMM* (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 are 111bp, 220bp and 1333bp, respectively. Detection of bimodality in the distribution, the signature of a past isolation event, is performed by fitting a Gaussian Mixture Model using the Expectation-Maximization algorithm from Meng and Rubin (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 analyze 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 are computed using subtype J as an outgroup. Results are robust to the chosen outgroup (Fig. K in File S1). Note that the distinction between recombining and non-recombining 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 (Fig. 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 (Fig. 1). Bimodality is observed in these distributions (both within- and between-populations, π_w and π_b) only when a population reconnection occurred (Fig. 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 Fig. 1A) or an ancient common ancestor (before the isolation event; second mode in Fig. 1A). Bimodality is also observed in the distributions of total pairwise nucleotide difference (considering all populations, π) after a reconnection event but, in this case, bimodality is also observed when populations are completely isolated ($T_{reco} = 0$, Fig. 1B-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 Fig. 1). Although testing for increase in variance requires data from different time-points, it has the advantage of not requiring *a priori* knowledge of the origin of the samples (i.e. being within- or between-populations).

Population reconnection after isolation can thus be detected by testing for bimodality and a temporal increase of the variance of the modes. Power to detect bimodality, with high probability (95%) (Fig. 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 Fig. 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 is larger between-population (π_b) than that of within-population (π_w, Fig. 1A). This difference in mode size increases when migration is weak or intermediate ($M \leq 1$) and the reconnection event is not recent (Fig. A in File S1). Ancient reconnection events can be detected even with small sample sizes ($n \simeq 10$) when migration is moderate or strong ($M \geq 1$; Fig. 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 Fig. 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 (Fig. 2). This excess is 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 (Fig. 2A), and can be identified by the visual test of equilibrium neutrality proposed by Nawa and Tajima (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 i/d, where d is the number of sampled populations and $1 \leq i < d$, when populations have equal sample sizes. In Fig. 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 2^d-2 peaks; see Fig. B in File S1).

The temporal patterns in the joint SFS (Fig. 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(20,i), (i,20), with i ranging from 1 to 20 in Fig. 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 Fig. 2B).

To robustly detect the excess of variants on the SFS generated by a reconnection event, we derive an optimal test statistic, denoted T_Ω, using developments provided by Ferretti *et al.* (2010) and Achaz (2009) (see Supporting Information file 1). 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 (Fig. E in File S1). The results presented in this section apply to both recombining and non-recombining sequences, as the expected SFS is the same in both cases (Gutenkunst *et al.* 2009). Also, the power estimates from the tests based on non-recombining 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.

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 (Fig. 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 (Fig. H

in File S1). Moreover, the signature is distinct from the signature of a closely related model: the Isolation with Migration model (IM) (Fig. I in File S1).

The observed signatures are robust even when a massive population expansion occurs concomitantly with the reconnection event (Fig. 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 populations). 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 α (Fig. 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 α values (light-gray lines Fig. 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 (Fig. I.B in File S1), whereas the signature of a reconnection of isolated populations is characterized by a decrease in peak size with time (Fig. 2A).

Thus, we show that, even though the signature at specific timepoints 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 (Fig. 3(a)-(b) and (g)-(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 Fig. 4). We found that a large proportion of the genome of HIV-1 displays bimodality (Fig. 4). In China, this proportion is initially low but rapidly increases through time. Between 2006 and 2007 the proportion of bimodality increased 4-fold (Fig. 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_Ω significantly decreased (Fig. 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 (Fig. 3(c)-(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 non-significant changes in values of optimal test statistics T_Ω (Fig. 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 HIV epidemic in Africa (Fig. 3(e)-(f) and (k)-(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 (Fig. 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 HIV-1 genomes experience constant gene flow from other subtypes but also represent current spatial distribution of HIV-1 populations. It reflects for example the extensive admixture between subtypes B and CRF01_AE in Brazil and China, between B and CRF01_AE in Thailand, and B and F1 in Brazil, Argentina and Spain.

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 here 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 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 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 co-infection 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 reconnection 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 reconnection 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 reconnection 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 (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 hours; Mohammadi *et al.* 2013), and a high mutation rate (about 0.002 substitution/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 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; Aguilée *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 non-human 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). Appling 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 a careful reading of the manuscript. This project was funded by the Swiss National Science Foundation (SNSF) grants PZ00P3_139421 and 31003A-130065 and PMPDP3_158381 (to SV), 31003A_149724/1 (to AT) and an interdisciplinary grant from the Faculty of Biology and Medicine (FBM) of the University of Lausanne (to SV and AT). 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 Interdisciplinary Reviews: Computational Statistics **2**: 433–459.

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

Aguilée, R., D. Claessen, and A. Lambert, 2013 Adaptive radiation driven by the interplay of eco-evolutionary and landscape dynamics. Evolution **67**: 1291–1306.

Alcala, N., D. Streit, J. Goudet, and S. Vuilleumier, 2013 Peak and persistent excess of genetic diversity following an abrupt migration increase. Genetics **193**: 953–971.

Alcala, N. and S. Vuilleumier, 2014 Turnover and accumulation of genetic diversity across large time-scale cycles of isolation and connection of populations. Proceedings of the Royal Society B: Biological Sciences **281**: 20141369.

An, M., X. Han, J. Xu, Z. Chu, M. Jia, H. Wu, L. Lu, Y. Takebe, and H. Shang, 2012 Reconstituting the epidemic history of hiv strain crf01_ae among men who have sex with men (msm) in liaoning, northeastern china: Implications for the expanding epidemic among msm in china. Journal of virology **86**: 12402–12406.

Aulicino, P. C., G. Bello, C. Rocco, H. Romero, A. Mangano, M. G. Morgado, and L. Sen, 2007 Description of the first full-length hiv type 1 subtype f1 strain in argentina: implications for the origin and dispersion of this subtype in south america. AIDS Res Hum Retroviruses **23**: 1176–1182.

Bailes, E., F. Gao, F. Bibollet-Ruche, V. Courgnaud, M. Peeters, P. A. Marx, B. H. Hahn, and P. M. Sharp, 2003 Hybrid origin of siv in chimpanzees. Science **300**: 1713–1713.

Bartlett, M. S., 1937 Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences **160**: 268–282.

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genetics Research **72**: 123–133.

Barton, N. H. and A. M. Etheridge, 2004 The effect of selection on genealogies. Genetics **166**: 1115–1131.

Baudry, E. and F. Depaulis, 2003 Effect of misoriented sites on neutrality tests with outgroup. Genetics **165**: 1619–1622.

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate bayesian computation in population genetics. Genetics **162**: 2025–2035.

Becquet, C. and M. Przeworski, 2009 Learning about modes of speciation by computational approaches. Evolution **63**: 2547–2562.

Bello, G., M. A. Soares, and C. G. Schrago, 2011 The use of bioinformatics for studying hiv evolutionary and epidemiological history in south america. AIDS Res Treat **2011**: 154945.

Benaglia, T., D. Chauveau, D. R. Hunter, D. S. Young, *et al.*, 2009 mixtools: An r package for analyzing finite mixture models. Journal of Statistical Software **32**: 1–29.

Birkner, M. and J. Blath, 2008 Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. Journal of mathematical biology **57**: 435–465.

Buzón, M. J., F. M. Codoñer, S. D. W. Frost, C. Pou, M. C. Puertas, M. Massanella, J. Dalmau, J. M. Llibre, M. Stevenson, J. Blanco, B. Clotet, R. Paredes, and J. Martinez-Picado, 2011 Deep molecular characterization of hiv-1 dynamics under suppressive haart. PLoS Pathog **7**: e1002314.

Caicedo, A. L., S. H. Williamson, R. D. Hernandez, A. Boyko, A. Fledel-Alon, T. L. York, N. R. Polato, K. M. Olsen, R. Nielsen, S. R. McCouch, C. D. Bustamante, and M. D. Purugganan, 2007 Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet **3**: 1745–1756.

Castro-Nallar, E., M. Pérez-Losada, G. F. Burton, and K. A. Crandall, 2012 The evolution of hiv: inferences using phylogenetics. Mol Phylogenet Evol **62**: 777–792.

Chikata, T., J. M. Carlson, Y. Tamura, M. A. Borghan, T. Naruto, M. Hashimoto, H. Murakoshi, A. Q. Le, S. Mallal, M. John, *et al.*, 2014 Host-specific adaptation of hiv-1 subtype b in the japanese population. Journal of virology **88**: 4764–4775.

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus, 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol **22**: 1185–1192.

Ellstrand, N. C., H. C. Prentice, and J. F. Hancock, 1999 Gene flow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics **30**: pp. 539–563.

Excoffier, L., 2004 Patterns of dna sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Mol Ecol **13**: 853–864.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and snp data. PLoS Genet **9**: e1003905.

Excoffier, L. and M. Foll, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics **27**: 1332–1334.

Fay, J. C. and C. I. Wu, 2000 Hitchhiking under positive darwinian selection. Genetics **155**: 1405–1413.

Fellay, J., K. V. Shianna, D. Ge, S. Colombo, B. Ledergerber, M. Weale, K. Zhang, C. Gumbs, A. Castagna, A. Cossarizza, *et al.*, 2007 A whole-genome association study of major determinants for host control of hiv-1. Science **317**: 944–947.

Feng, Y., X. He, J. H. Hsi, F. Li, X. Li, Q. Wang, Y. Ruan, H. Xing, T. T.-Y. Lam, O. G. Pybus, Y. Takebe, and Y. Shao, 2013 The rapidly expanding crf01_ae epidemic in china is driven by multiple lineages of hiv-1 viruses introduced in the 1990s. AIDS **27**: 1793–1802.

Feng, Y., H. Wei, J. Hsi, H. Xing, X. He, L. Liao, Y. Ma, C. Ning, N. Wang, Y. Takebe, and Y. Shao, 2014 Identification of a novel hiv type 1 circulating recombinant form (crf65_cpx) composed of crf01_ae and subtypes b and c in western yunnan, china. AIDS Res Hum Retroviruses **30**: 598–602.

Ferretti, L., M. Perez-Enciso, and S. Ramos-Onsins, 2010 Optimal neutrality tests based on the frequency spectrum. Genetics **186**: 353–365.

Fisher, R. A., 1930 *The genetical theory of natural selection*. Clarendon

Press.
Flahault, A., P. Zylberman, *et al.*, 2010 Influenza pandemics: past, present and future challenges. Public Health Reviews **32**: 319–340.
Foll, M., H. Shim, and J. D. Jensen, 2014 Wfabc: a wright–fisher abc-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular Ecology Resources pp. n/a–n/a.
Fu, Y. X. and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.
Garten, R. J., C. T. Davis, C. A. Russell, B. Shu, S. Lindstrom, A. Balish, W. M. Sessions, X. Xu, E. Skepner, V. Deyde, M. Okomo-Adhiambo, L. Gubareva, J. Barnes, C. B. Smith, S. L. Emery, M. J. Hillman, P. Rivailler, J. Smagala, M. de Graaf, D. F. Burke, R. A. M. Fouchier, C. Pappas, C. M. Alpuche-Aranda, H. López-Gatell, H. Olivera, I. López, C. A. Myers, D. Faix, P. J. Blair, C. Yu, K. M. Keene, P. D. Dotson, Jr, D. Boxrud, A. R. Sambol, S. H. Abid, K. St George, T. Bannerman, A. L. Moore, D. J. Stringer, P. Blevins, G. J. Demmler-Harrison, M. Ginsberg, P. Kriner, S. Waterman, S. Smole, H. F. Guevara, E. A. Belongia, P. A. Clark, S. T. Beatrice, R. Donis, J. Katz, L. Finelli, C. B. Bridges, M. Shaw, D. B. Jernigan, T. M. Uyeki, D. J. Smith, A. I. Klimov, and N. J. Cox, 2009 Antigenic and genetic characteristics of swine-origin 2009 a(h1n1) influenza viruses circulating in humans. Science **325**: 197–201.
Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, *et al.*, 2002 Diversity considerations in hiv-1 vaccine selection. Science **296**: 2354–2360.
Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, T. . G. Project, and C. D. Bustamante, 2011 Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences **108**: 11983–11988.
Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo, 2010 A draft sequence of the neandertal genome. Science **328**: 710–722.
Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genet **5**: e1000695.
Hamelaar, J., E. Gouws, P. D. Ghys, S. Osmanov, and *WHO-UN AIDS Network for HIV Isolation and Characterisation*, 2011 Global trends in molecular epidemiology of hiv-1 during 2000-2007. AIDS **25**: 679–689.
Han, X., M. An, B. Zhao, S. Duan, S. Yang, J. Xu, M. Zhang, J. M. McGoogan, Y. Takebe, and H. Shang, 2013 High prevalence of hiv-1 intersubtype b /c recombinants among injecting drug users in dehong, china. PloS one **8**: e65337.
Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers, and S. T. Sherry, 1998 Genetic traces of ancient demography. Proc Natl Acad Sci U S A **95**: 1961–1967.
Hey, J., 2010 Isolation with migration models for more than two populations. Mol Biol Evol **27**: 905–920.
Hsi, J., H. Wei, H. Xing, Y. Feng, X. He, L. Liao, M. Jia, N. Wang, C. Ning, and Y. Shao, 2014 Genome sequence of a novel hiv-1 circulating recombinant form (crf64_bc) identified from yunnan, china. AIDS Res Hum Retroviruses **30**: 389–393.
Hsieh, Y.-C., T.-Z. Wu, D.-P. Liu, P.-L. Shao, L.-Y. Chang, C.-Y. Lu, C.-Y. Lee, F.-Y. Huang, and L.-M. Huang, 2006 Influenza pandemics: past, present and future. J Formos Med Assoc **105**: 1–6.
Hudson, R. R., 2002 Generating samples under a wright–fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.
Jombart, T., S. Devillard, and F. Balloux, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC genetics **11**: 94.
Karlsson, E. K., D. P. Kwiatkowski, and P. C. Sabeti, 2014 Natural selection and infectious disease in human populations. Nature Reviews Genetics **15**: 379–393.
Keele, B. F., E. E. Giorgi, J. F. Salazar-Gonzalez, J. M. Decker, K. T. Pham, M. G. Salazar, C. Sun, T. Grayson, S. Wang, H. Li, *et al.*, 2008 Identification and characterization of transmitted and early founder virus envelopes in primary hiv-1 infection. Proceedings of the National Academy of Sciences **105**: 7552–7557.
Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893–903.
Korber, B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours, 2001 Evolutionary and immunological implications of contemporary hiv-1 variation. British medical bulletin **58**: 19–42.
Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya, 2000 Timing the ancestor of the hiv-1 pandemic strains. Science **288**: 1789–1796.
Lecis, R., M. Pierpaoli, Z. Biro, L. Szemethy, B. Ragni, F. Vercillo, and E. Randi, 2006 Bayesian analyses of admixture in wild and domestic cats (felis silvestris) using linked microsatellite loci. Molecular Ecology **15**: 119–131.
Li, X., C. Ning, X. He, Y. Yang, H. Xing, K. Hong, Y. Shao, and R. Yang, 2013 Near full-length genome sequence of a novel hiv type 1 second-generation recombinant form (crf01_ae/crf07_bc) identified among men who have sex with men in jilin, china. AIDS research and human retroviruses .
Liégeois, F., F. Schmidt, V. Boué, C. Butel, F. Mouacha, P. Ngari, B. M. Ondo, E. Leroy, J. L. Heeney, E. Delaporte, *et al.*, 2014 Full-length genome analyses of two new simian immunodeficiency virus (siv) strains from mustached monkeys (c. cephus) in gabon illustrate a complex evolutionary history among the sivmus/mon/gsn lineage. Viruses **6**: 2880–2898.
Lohse, K., R. Harrison, and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics **189**: 977–987.
Lu, L., M. Jia, Y. Ma, L. Yang, Z. Chen, D. D. Ho, Y. Jiang, and L. Zhang, 2008 The changing face of hiv in china. Nature **455**: 609–611.
Ma, J. and C. I. Amos, 2012 Principal components analysis of population admixture. PLoS One **7**: e40115.
Meng, X.-L. and D. B. Rubin, 1993 Maximum likelihood estimation via the ecm algorithm: A general framework. Biometrika **80**: 267–278.
Mohammadi, P., S. Desfarges, I. Bartha, B. Joos, N. Zangger, M. Muñoz, H. F. Günthard, N. Beerenwinkel, A. Telenti, and A. Ciuffi, 2013 24 hours in the life of hiv-1 in a t cell line. PLoS pathogens **9**: e1003161.

Naduvilezhath, 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 & Genetic Systems **83**: 353–360.

Neher, R. A. and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences **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, F.-L. Liu, H. Li, Y.-Q. Tu, and Y.-T. Zheng, 2012 Extensive and complex hiv-1 recombination between b′, c and crf01_ae among idus in southeast asia. AIDS **26**: 1121–1129.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, 2012 Ancient admixture in human history. Genetics **192**: 1065–1093.

Pelak, K., A. C. Need, J. Fellay, K. V. Shianna, S. Feng, T. J. Urban, D. Ge, A. De Luca, J. Martinez-Picado, S. M. Wolinsky, *et al.*, 2011 Copy number variation of kir genes influences hiv-1 control. PLoS biology **9**: e1001208.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley, 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, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, *et al.*, 2012 Reconstructing native american population history. Nature **488**: 370–374.

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, J. D. Jensen, and T. F. Kowalik, 2013 Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. PLoS genetics **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. Current opinion in virology **8**: 109–115.

Renzette, N., C. Pokalyuk, L. Gibson, B. Bhattacharjee, M. R. Schleiss, K. Hamprecht, A. Y. Yamamoto, M. M. Mussi-Pinhata, W. J. Britt, J. D. Jensen, *et al.*, 2015 Limits and patterns of cytomegalovirus genomic diversity in humans. Proceedings of the National Academy of Sciences pp. E4120–E4128.

Robertson, D., J. Anderson, J. Bradac, J. Carr, B. Foley, R. Funkhouser, F. Gao, B. Hahn, M. Kalish, C. Kuiken, *et al.*, 2000 Hiv-1 nomenclature proposal. Science **288**: 55–55.

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

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

Schlub, T. E., A. J. Grimm, R. P. Smyth, D. Cromer, A. Chopra, S. Mallal, V. Venturi, C. Waugh, J. Mak, and M. P. Davenport, 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. Swanstrom, 2011 Hiv-1 replication in the central nervous system occurs in two distinct cell types. PLoS pathogens **7**: e1002286.

Schultz, A.-K., M. Zhang, I. Bulla, T. Leitner, B. Korber, B. Morgenstern, and M. Stanke, 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 pleistocene 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. Philosophical Transactions of the Royal Society B: Biological Sciences **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 **143**: 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.

Sousa, V. and J. Hey, 2013 Understanding the origin of species with genome-scale data: modelling gene flow. Nature Reviews Genetics **14**: 404–414.

Sousa, 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, J. Köstler, H. Melzl, H. Wolf, Y. Shao, and R. Wagner, 2000 Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (c/b) recombinant strain in china. Journal of virology **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, X.-J. Li, X. Han, H. Shang, A. Kamarulzaman, N. Yamamoto, *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. Genetics **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, C. Ning, L. Zhang, J. Wang, Y. Feng, and Y. Shao, 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**: 484–488.

Wei, H., J. His, Y. Feng, H. Xing, X. He, L. Liao, S. Duan, C. Ning, N. Wang, Y. Takebe, and Y. Shao, 2014 Identification of a novel hiv-1 circulating recombinant form (crf62_bc) in western yunnan of china. AIDS Res Hum Retroviruses **30**: 380–383.

Wilkinson-Herbots, H. M., 2012 The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. Theoretical Population Biology **82**: 92 – 108.

Wright, S., 1931 Evolution in mendelian populations. Genetics **16**: 97–159.

Yang, R., X. Xia, S. Kusagawa, C. Zhang, K. Ben, and Y. Takebe, 2002 On-going generation of multiple forms of hiv-1 intersubtype recombinants in the yunnan province of china. Aids **16**: 1401–1407.

Zeng, K., Y.-X. Fu, S. Shi, and C.-I. Wu, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics **174**: 1431–1439.

Zhang, L., E. P. F. Chow, J. Jing, X. Zhuang, X. Li, M. He, H. Sun, X. Li, M. Gorgens, D. Wilson, L. Wang, W. Guo, D. Li, Y. Cui, L. Wang, N. Wang, Z. Wu, and D. P. Wilson, 2013 Hiv prevalence in china: integration of surveillance data and a systematic review. Lancet Infect Dis **13**: 955–963.

APPENDIX A: 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):

$$\mathbf{Q} = \begin{pmatrix} -M-1 & M & 1 \\ M/(d-1) & -M/(d-1) & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{A.1}$$

The probability that two genes were in state (1), (2) or (3) T generations ago (row vector $\mathbf{P}_T$), given that they are in state (1), (2) or (3) at the current generation is given by $\mathbf{P}_T = \mathbf{P}_0 e^{\mathbf{Q}T}$ ($\mathbf{P}_0$ at generation 0), where $e^{\mathbf{Q}T}$ is a 3×3 matrix. The elements of matrix $e^{\mathbf{Q}T}$ are denoted $q_{i,j}(T)$ 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 (or probability *density* function) of within-population and between-population pairwise coalescence times are given by the derivatives of corresponding $q_{i,j}(T)$ functions: $f_{w,coal}(T) = \frac{d}{dT} q_{1,3}(T)$ and $f_{b,coal}(T) = \frac{d}{dT} q_{2,3}(T)$. Populations are considered isolated when $T_{reco} < T < T_{iso}$ and no migration occurs, with $M = 0$ in the transition matrix $\mathbf{Q}$. Populations exchange migrants when $T < T_{reco}$ and $T > T_{iso}$, thus $M > 0$ in the transition matrix $\mathbf{Q}$. Consequently, the distributions of within- and between-population pairwise coalescence times, $f_{T_w}(T)$ and $f_{T_b}(T)$, following a past isolation event, are:

$$f_{T_w}(T) = \begin{cases} f_{w,coal}(T), & T < T_{reco} \\ q_{1,1}(T_{reco}) e^{-(T-T_{reco})}, & T_{reco} < T < T_{iso} \\ q_{1,1}(T_{reco}) e^{-(T_{iso}-T_{reco})} f_{w,coal}(T - T_{iso}) & \\ \quad + q_{1,2}(T_{reco}) f_{b,coal}(T - T_{iso}), & T_{iso} < T, \end{cases} \tag{A.2a}$$

$$f_{T_b}(T) = \begin{cases} f_{b,coal}(T), & T < T_{reco} \\ q_{2,1}(T_{reco}) e^{-(T-T_{reco})}, & T_{reco} < T < T_{iso} \\ q_{2,1}(T_{reco}) e^{-(T_{iso}-T_{reco})} f_{w,coal}(T - T_{iso}) & \\ \quad + q_{2,2}(T_{reco}) f_{b,coal}(T - T_{iso}), & T_{iso} < T. \end{cases} \tag{A.2b}$$

These distributions have a simple interpretation. In the most recent period, the reconnection period, when $T < T_{reco}$, $f_{T_w}(T)$ 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} generations of connection (probabilities $q_{1,1}(T_{reco})$ and $q_{2,1}(T_{reco})$ 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 $q_{1,1}(T_{reco})$ and $q_{2,1}(T_{reco})$, within- and between-populations, respectively) and if they did not coalesce during isolation (probability $e^{-(T_{iso}-T_{reco})}$). Their probability of coalescence at generation T are $f_{w,coal}(T - T_{iso})$. Alternatively, they could have coalesced if they were in different populations during isolation (probabilities $q_{1,2}(T_{reco})$ and $q_{2,2}(T_{reco})$). Their probability of coalescence at generation T is $f_{b,coal}(T - T_{iso})$.

Functions $f_{w,coal}(T)$, $f_{b,coal}(T)$, $q_{1,1}(T)$, $q_{1,2}(T)$, $q_{2,1}(T)$ and $q_{2,2}(T)$ all depend on the eigenvalues of matrix $\mathbf{Q}$, $\lambda_1 = -\frac{1}{2}(1 + \frac{d}{d-1}M + \sqrt{\Delta})$ and $\lambda_2 = -\frac{1}{2}(1 + \frac{d}{d-1}M - \sqrt{\Delta})$, where $\Delta = (d - 1 + dM)^2 - 4(d-1)M$. We obtain:

$$\begin{aligned} f_{w,coal}(T) &= A_{w,1}\lambda_1 e^{\lambda_1 T} + A_{w,2}\lambda_2 e^{\lambda_2 T} \\ f_{b,coal}(T) &= A_{b,1}\lambda_1 e^{\lambda_1 T} + A_{b,2}\lambda_2 e^{\lambda_2 T} \end{aligned} \tag{A.3a}$$

$$
\begin{aligned}
q_{1,1}(T) &= (B + A_{w,1})e^{\lambda_1 T} - (B - A_{w,2})e^{\lambda_2 T} \\
q_{1,2}(T) &= -Be^{\lambda_1 T} + Be^{\lambda_2 T} \\
q_{2,1}(T) &= \frac{-B}{d-1}e^{\lambda_1 T} + \frac{B}{d-1}e^{\lambda_2 T} \\
q_{2,2}(T) &= (-B + A_{w,2})e^{\lambda_1 T} + (B + A_{w,1})e^{\lambda_2 T}.
\end{aligned} \tag{A.3b}
$$

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- (π_w) and between-population pairwise nucleotide differences (π_b). Given that mutations follow a Poisson process of mean θ, we have $P(\pi_w{=}k/G) = \int_0^\infty e^{-\theta T}\frac{(\theta T)^k}{k!}f_{T_w}(T)dT$ and $P(\pi_b{=}k) = \int_0^\infty e^{-\theta T}\frac{(\theta T)^k}{k!}f_{T_b}(T)dT$, 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 site). We can decompose $P(\pi_w{=}k)$ into a sum of smaller integrals (respectively for $P(\pi_b{=}k)$):

$$
\begin{aligned}
P(\pi_w{=}\frac{k}{G}) = &\int_0^{T_{reco}} e^{-\theta T}\frac{(\theta T)^k}{k!}f_{T_w}(T)dT + \int_{T_{reco}}^{T_{iso}} e^{-\theta T}\frac{(\theta T)^k}{k!}f_{T_w}(T)dT \\
&+ \int_{T_{iso}}^{\infty} e^{-\theta T}\frac{(\theta T)^k}{k!}f_{T_w}(T)dT.
\end{aligned} \tag{A.4}
$$

Substituting the value of $f_{T_w}(T)$ in each term of equation A.4 by the corresponding value in equation A.2a and using the result $\int_{T_1}^{T_2} e^{-\theta T}\frac{(\theta T)^k}{k!}\lambda e^{\lambda T}dT = \frac{(-\frac{\theta}{\lambda})^k}{(1-\frac{\theta}{\lambda})^{k+1}}[e^{(\lambda-\theta)T_2}\sum_{l=0}^{k}\frac{T_2^l(-\lambda+\theta)^l}{l!} - e^{(\lambda-\theta)T_1}\sum_{l=0}^{k}\frac{T_1^l(-\lambda+\theta)^l}{l!}]$, we obtain:

$$
\begin{aligned}
P(\pi_w{=}\frac{k}{G}) = &\sum_{i=0}^{2} A_{w,i}\frac{(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}(1-e^{(\lambda_i-\theta)T_{reco}}\sum_{j=0}^{k}\frac{T_{reco}^l(-\lambda_i+\theta)^l}{l!}) \\
&+ q_{1,1}(T_{reco})e^{T_{reco}}\frac{\theta^k}{(1+\theta)^{k+1}}[e^{-(1+\theta)T_{reco}}\sum_{l=0}^{k}\frac{T_{reco}^l(1+\theta)^l}{l!} \\
&- e^{-(1+\theta)T_{iso}}\sum_{l=0}^{k}\frac{T_{iso}^l(1+\theta)^l}{l!}] \\
&+ \sum_{i=0}^{2}[\frac{(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}e^{-\theta T_{iso}}\sum_{l=0}^{k}\frac{T_{iso}^l(-\lambda_i+\theta)^l}{l!}) \\
&(q_{1,1}(T_{reco})e^{-T_{iso}+T_{reco}}A_{w,i} + q_{1,2}(T_{reco})A_{b,i})]
\end{aligned} \tag{A.5}
$$

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

$$
\begin{aligned}
P(\pi_w{=}\frac{k}{G}) = &\sum_{i=0}^{2} A_{w,i}\frac{(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}(1-e^{(\lambda_i-\theta)T_{reco}}\sum_{j=0}^{k}\frac{T_{reco}^l(-\lambda_i+\theta)^l}{l!}) \\
&+ q_{1,1}(T_{reco})\frac{\theta^k}{(1+\theta)^{k+1}}e^{-\theta T_{reco}}\sum_{l=0}^{k}\frac{T_{reco}^l(1+\theta)^l}{l!} \\
&+ q_{1,2}(T_{reco})\sum_{i=0}^{2}\frac{A_{b,i}(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}e^{-\theta T_{iso}}\sum_{l=0}^{k}\frac{T_{iso}^l(-\lambda_i+\theta)^l}{l!},
\end{aligned} \tag{A.6a}
$$

$$
\begin{aligned}
P(\pi_b{=}\frac{k}{G}) = &\sum_{i=0}^{2} \frac{A_{b,i}(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}(1-e^{(\lambda_i-\theta)T_{reco}}\sum_{j=0}^{k}\frac{T_{reco}^l(-\lambda_i+\theta)^l}{l!}) \\
&+ q_{2,1}(T_{reco})\frac{\theta^k}{(1+\theta)^{k+1}}e^{-\theta T_{reco}}\sum_{l=0}^{k}\frac{T_{reco}^l(1+\theta)^l}{l!} \\
&+ q_{2,2}(T_{reco})\sum_{i=0}^{2}\frac{A_{b,i}(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}e^{-\theta T_{iso}}\sum_{l=0}^{k}\frac{T_{iso}^l(-\lambda_i+\theta)^l}{l!}.
\end{aligned} \tag{A.6b}
$$

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, $\sum_{i=0}^{2} A_{w,i}\frac{(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}$ (resp. $\sum_{i=0}^{2} A_{b,i}\frac{(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}$) corresponds to the distribution of π_w (resp. π_b) under the finite island model at equilibrium, while terms $(1 - e^{(\lambda_i-\theta)T_{reco}}\sum_{j=0}^{k}\frac{T_{reco}^l(-\lambda_i+\theta)^l}{l!})$ are due to the truncation of the distribution between T_{reco} and infinity. The second term corresponds to sequences that coalesce during the isolation period. Indeed, with probability $q_{1,1}(T_{reco})$ (resp. $q_{2,1}(T_{reco})$), 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 π_w in an isolated population, and the term $e^{-\theta T_{reco}}\sum_{l=0}^{k}\frac{T_{reco}^l(1+\theta)^l}{l!}$ is due to the truncation between 0 and T_{reco}. Finally, the third term corresponds to sequences that coalesce during the prior connection period. With probability $q_{1,2}(T_{reco})$ (resp. $q_{2,2}(T_{reco})$), sequences coalesce during the prior connection period (as they did not coalesce during the reconnection period and were in different populations during isolation). Then, $\sum_{i=0}^{2} A_{b,i}\frac{(-\frac{\theta}{\lambda_i})^k}{(1-\frac{\theta}{\lambda_i})^{k+1}}$ corresponds to the distribution of π_b under the finite island model at equilibrium, while terms $e^{-\theta T_{iso}}\sum_{j=0}^{k}\frac{T_{iso}^l(-\lambda_i+\theta)^l}{l!}$ are due to the truncation of the distribution between 0 and T_{iso}.

Derivation of the power to detect bimodality

We can derive the power to detect bimodality from the distributions of π_w and π_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 π_w to be drawn from the second mode of the distribution is $q_{1,2}(T_{reco})$, the probability that among n differences, at least one is from the first mode and one from the second mode is $1 - q_{1,2}(T_{reco})^n - (1 - q_{1,2}(T_{reco}))^n$. Similarly, the probability to observe a bimodal distribution of π_b from a sample of size n is $1 - q_{2,2}(T_{reco})^n - (1 - q_{2,2}(T_{reco}))^n$. These results are used to draw Fig. 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 $\frac{n(n-1)}{2}$.

SUPPORTING INFORMATION

File S1

Combined supporting information file, including all supporting figures, text and tables mentioned in the manuscript. File S1 is divided into 4 sections:

- *Bimodality on pairwise nucleotide differences as a signal of population reconnection after isolation*, which includes **Supporting Figure A**
- *The signature of population reconnection after isolation on the Site Frequency Spectrum (SFS)*, which includes **Supporting Figures B** and **C**, **text A**, and **Supporting Figures D** and **E**.
- *Robustness of the signature*, which includes **Supporting Figures F**, **G**, **H** and **I**.
- *Detection of past isolation and current reconnection of HIV-1 subtypes*, which includes **Supporting Figures J**, **K** and **L**, and **Supporting Tables A**, **B** and **C**.

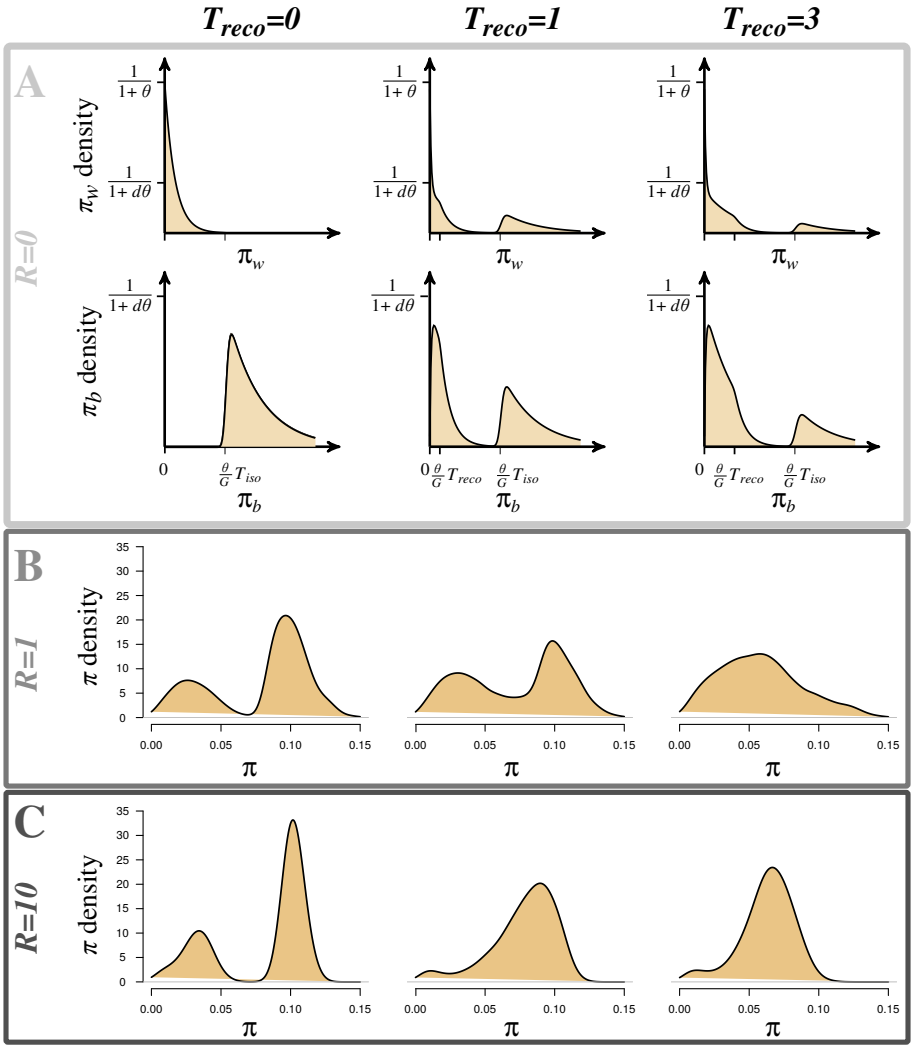

Figure 1 Signature of reconnection of previously isolated populations on the distribution of pairwise nucleotide differences as a function of time since the reconnection event T_{reco} (columns), (A) within- π_w and between-population π_b, (B-C) total π pairwise nucleotide differences. In panel A, we consider a sequence of size G, with G larger than the number of segregating sites without recombination; plots in panel A are computed using equation A.5. In panels B and C, we consider a sequence of 9719bp (length of the reference sequence for HIV-1, HXB1) with various mean number of recombination events per generation, $R = 1$ and $R = 10$. Nucleotide differences are considered from a sample of 16 sequences and each plot represents the mean distribution over 2000 replicate simulations. Parameters are $d = 3$ populations, duration of the isolation period $T_{iso} = 6$, scaled migration rate $M = 5$, scaled mutation rate $\theta = 100$.

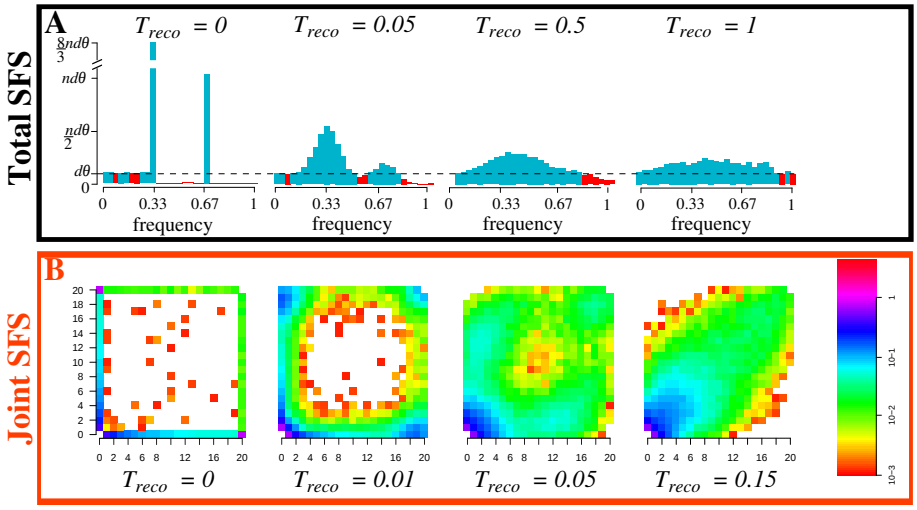

Figure 2 Signature of a reconnection event between previously isolated populations on (A) the total and (B) joint site frequency spectrum (SFS), as a function of the number of generations since the reconnection event, T_{reco}. (A) The SFS representation of Nawa and Tajima (Nawa and Tajima 2008) is used; the expected value of a single panmictic population at equilibrium is given by a straight line (dashed line), and deviations from a straight line indicate an excess (in blue) or deficit (in red) of variants at a given frequency. Parameters are $d = 3$ populations, $n = 16$ sampled genes per population (see Figs. B and C in File S1 for alternative sampling schemes), with $\theta = 100$. Means are over 2,000 replicates.

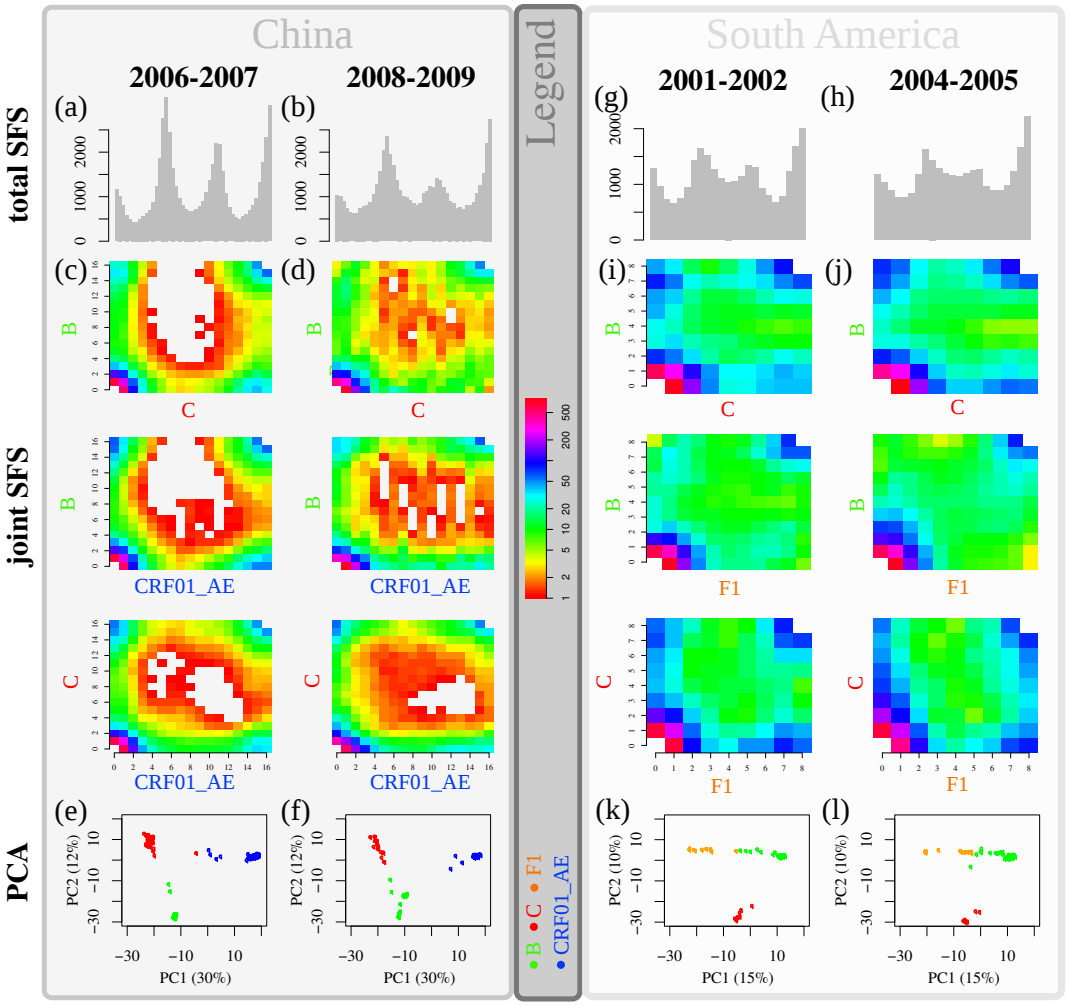

Figure 3 Signatures of HIV-1 subtype reconnection events on the Site Frequency Spectrum (SFS). Results are presented for China (left major panel) and for South America (right major panel). Two time points of HIV-1 sequences sampled are considered in China (Table B in File S1) and in South America (Table C in File S1) (see method section 'HIV-1 genome analysis'). The total (pooled) SFS is presented in the first row following representation of Nawa and Tajima (Nawa and Tajima 2008), the joint SFSs of each pairwise group of HIV subtypes are presented in the 2nd, 3rd and 4th row and in the 5th row are presented the Principal Component Analyses (PCAs). Subtype J is used as an outgroup (see Fig. 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 Fig. J in File S1 for the different clusters). A sample of 16 sequences per subtype cluster is considered in China, and of 8 sequences per subtype cluster in South America. The significance of the trends in the dynamics is tested using the optimal test statistic T_Ω derived in the SI (Fig. L in File S1).

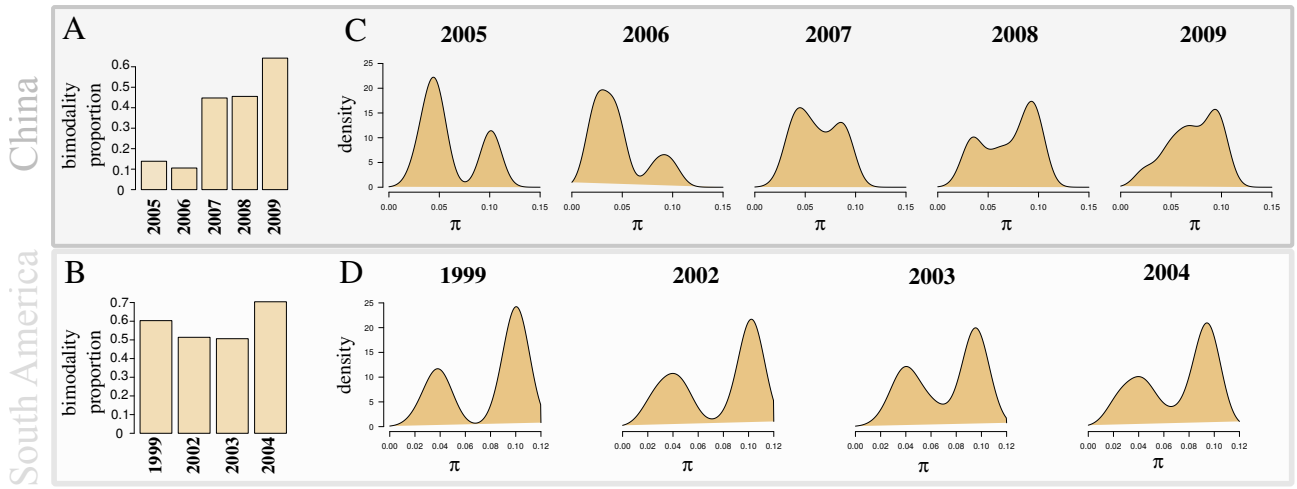

Figure 4 (A-B) Proportion of bimodality in pairwise nucleotide differences (between populations) detected in the genomes and (C-D) distribution of total pairwise nucleotide differences (all populations) of HIV-1 sequences for different time points in China (see Table B in File S1) and South America (see Table B in File S1). As expected under genomic admixture (Fig. 1), the proportion of bimodality detected in the pairwise nucleotide differences between populations in HIV-1 genomes increased and the variance of modes significantly increased with time in China ($p < 0.01$, two-sided Bartlett test). We also see a (non-significant) trend of temporal changes in South America.

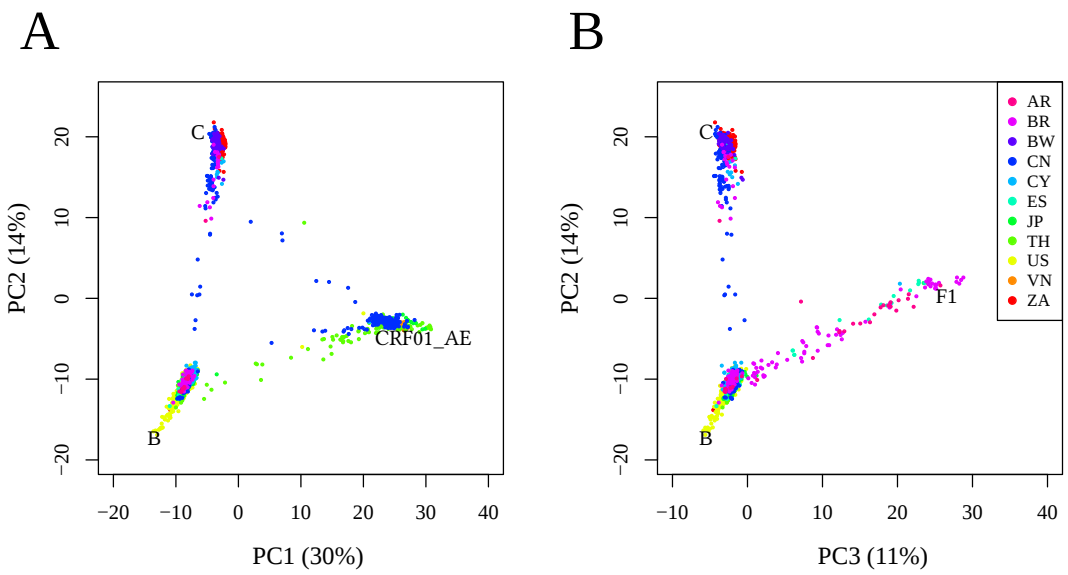

Figure 5 Projection of worldwide distributed HIV-1 sequences from subtypes B, C, CRF01_AE and F1, and URF and CRF recombinant forms between these subtypes. 1646 genome sequences (see Table A in File S1). (A) the first two axes of the principal component analysis (PC1 and PC2), and (B) the second and 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 (Ma 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.