Genetic diversity and molecular epidemiology of HIV transmission

Eduardo Castro-Nallar[1], Keith A Crandall[1] & Marcos Pérez-Losada*
[1]Department of Biology, 401 Widtsoe Building, Brigham Young University, Provo, UT 84602-5181, USA
*Author for correspondence: CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal ■ mlosada323@gmail.com

The high genetic diversity of HIV is one of its most significant features, as it has consequences in global distribution, vaccine design, therapy success, disease progression, transmissibility and viral load testing. Studying HIV diversity helps to understand its origins, migration patterns, current distribution and transmission events. New advances in sequencing technologies based on the parallel acquisition of data are now used to characterize within-host and population processes in depth. Additionally, we have seen similar advances in statistical methods designed to model the past history of lineages (the phylodynamic framework) to ultimately gain better insights into the evolutionary history of HIV. We can, for example, estimate population size changes, lineage dispersion over geographic areas and epidemiological parameters solely from sequence data. In this article, we review some of the evolutionary approaches used to study transmission patterns and processes in HIV and the insights gained from such studies.

Recent advances in DNA sequencing as well as new approaches to analyzing these data allow researchers to study the impact of epidemiological factors on the evolutionary dynamics of HIV at global (worldwide), regional (single epidemics), local (transmission chains) and individual (intrahost) scales by examining genetic variation across its genome and over geographic space and time. The purpose of this review is to present an overview of HIV genetic diversity (GD) and its estimation as well as the kind of insights these new methodologies can provide and how they can improve disease control and treatment.

GD of HIV

GD is probably one of the most important concepts in biology. In its most simple definition, GD refers to any and every kind of genetic variation at the individual, population, interpopulation or species level. GD has a large impact on conservation biology [1] and the study of human origins [2], as well as molecular epidemiology [3], domestication [4], fitness [5] and disease [6]. In HIV-1, higher levels of GD have been associated with clinical outcomes such as immune escape of selected variants [7], emergence of drug resistance mutations and the consequent therapy failure [8], and even with disease progression [9,10]. GD has also been used to study the geographic and temporal spread of HIV-1, shedding light on global and regional population dynamics.

HIV-1 GD stems from at least three different sources: multiple introductions of HIV-1 into the human population [11–13], the low fidelity and high recombinogenic power [14,15] of its reverse transcriptase [16] and its high virus turnover [17]. HIV-1 and -2 GD has been classified within discrete groups and subtypes that largely correspond to geographic regions (Figure 1) [18]. HIV-1 includes four groups, which represent different introductions into human populations, namely group M (Main, Major), N (New, Non-major), O (Outlier) and recently P [19]. HIV-1 group M is the most commonly detected variant, which, in turn, is further divided into subtypes – that is, A–D, F–H, J and K. Also, circulating recombinant forms (CRFs) carrying genetic information from two or more subtypes have been detected in infected populations (49 to date) [20]. HIV-2 includes two groups (A and B) and it is worth noting that HIV-2 CRFs have been reported only once [21]. Genetic variation among HIV subtypes is tremendous, with within-subtype divergences reaching up to 17% and between-subtype divergences of 17–35%. For comparison, human and chimpanzee divergence can reach up to 3.9% if substitutions and insertions/deletions are considered [22].

Early on, researchers noted differences in transmissibility between HIV-1 and -2 [23]. Some transmission differences are explained on the basis of structural and evolutionary differences in *env* genes [24]. It is clear that HIV-2 exhibits lower rates of transmission, almost no vertical transmission and long incubation periods [25]. Moreover, HIV-2-infected patients have reduced immune activation, low viremia and

Review
Future Virology

Keywords
■ drug resistance ■ genetic diversity ■ HIV ■ phylodynamics ■ transmission ■ vaccines

Future Medicine part of fsg

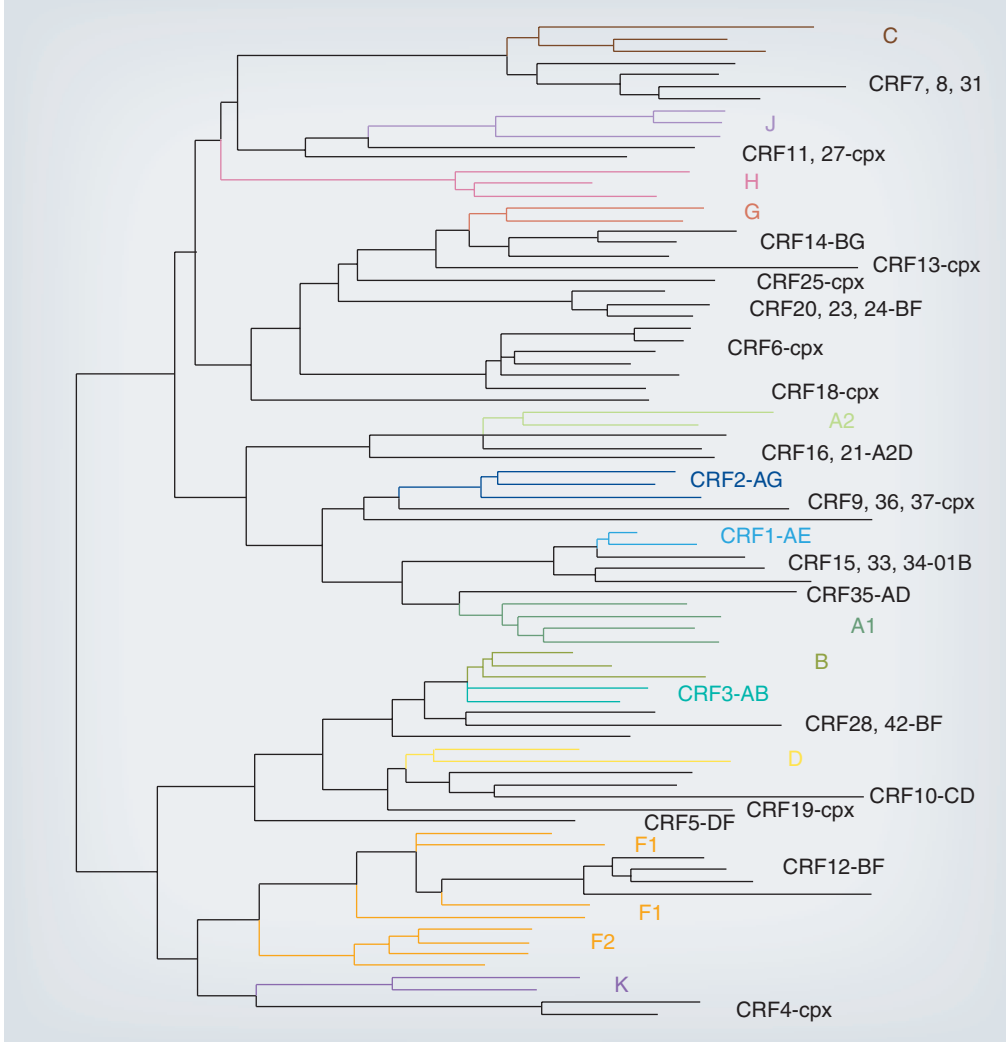

Figure 1. HIV-1 recombinants and subtypes. Phylogenetic tree representation of HIV-1 recombinants and discrete subtypes. A–D, F–H, J and K denote HIV-1 subtypes. No subtypes E or I are shown since they were found to be recombinant forms of other subtypes.
cpx: Complex recombinant pattern; CRF: Circulating recombinant form.
Reproduced with permission from [51].

rarely develop disease [26]. Within HIV-1, some researchers have proposed biological correlates to different subtypes; for example, HIV-1 subtype A-infected women are less likely to develop AIDS than non-subtype A-infected women [27]. Also, differences among subtypes have been reported in relation to chemokine coreceptor usage (tissue tropism) [28] and transmission [29]. Other researchers have suggested that the 'subtypes' are simply artifacts of poor or selective sampling of the overall HIV-1 GD [30].

Measures of GD

Given the interaction between HIV genetic and epidemiological dynamics, accurately estimating GD becomes particularly important. GD can be directly estimated from nucleotide sequence data using a variety of approaches.

Traditional measures of GD

The most intuitive way of measuring GD would be to simply count the pairwise differences between sequences in a sequence alignment or the number of polymorphic sites [31]. This approach, however, has some limitations, such as the existence of multiple hits in the same position, different probabilities of change in coding sequences or transition/transversion bias. In general, measures of GD based on sequence data can be classified into two basic categories: summary statistics and coalescent estimators. Some commonly used summary statistic approaches are: nucleotide diversity [32], haplotype diversity [32], allelic diversity, gene diversity and theta (Θ) [33]. More recently, Θ, expressed as substitution rate-scaled effective population size, has been estimated under a coalescent framework

explicitly taking into account evolutionary history [34–37], which differentiates this model from approaches based on summary statistics. The coalescent model has been further generalized to account for varying population sizes, different time scales, structure, recombination and selection, which has probably made it the most used method to estimate GD. Detailed descriptions regarding algorithms for Θ estimation can be found in other reviews [38–40].

Novel approaches

Newly developed approaches intend to estimate diversity parameters by taking advantage of the massive amounts of data that next-generation sequencing (NGS) technologies can deliver [41]. New methodologies have focused on characterizing intrahost diversity by capturing low-frequency variants [42]. Recent implementations take advantage of Bayesian inference to correct errors [43] and infer haplotypes and their frequencies (as low as 0.1% [44]) [45]. Also, the determination of HIV full genomic sequences and related measures of diversity are now feasible, which opens unexplored possibilities to comprehensively address how HIV mutates under selective pressures [46]. To date, most of the applications of NGS technologies to 'ultra-deep sequencing' of viral populations have focused on drug resistance characterization, fluctuations in GD through disease progression, and on certain events in HIV biology, such as tropism switch, transmission bottlenecks, immune escape [47,48], epistasis [49] and superinfections [50]. Drug resistance characterization has been performed primarily on target genes such as *RT* and *Integrase* (both within *pol*; Figure 2). In turn, epidemiological and subtyping studies primarily focus on the capsid proteins encoded within the *env* reading frame [51,52]. In the future, we expect to see more applications based on full genome data as the cost of sequencing decreases [42,53].

Geographic & temporal spread of diversity

Phylodynamics [54], or the description of infectious disease behavior that arises from the blending of evolutionary and epidemiological processes, has become a hot subject in virology and epidemiology, especially following recent statistical developments (see Table 1) [55–69]. These new methods often use the coalescent under a maximum likelihood or Bayesian inference framework and have been applied in HIV to study questions related to its global, regional, local and within-host dynamics.

Estimating phylodynamics

Accurate phylodynamic estimation relies on an adequate sampling strategy [70]. Sparse sampling could lead to inappropriate inferences – for example, the east Africa direct transmission of South American HIV-1 subtype C [71]. Since phylodynamic inferences rely on 'time trees' or dated phylogenies, it is also necessary to calibrate the molecular clock model in use. Due to the lack of fossil records in HIV, time-stamped sequence data can be used to produce those inferences. Also, incorporating independent prior knowledge about substitution rates will help any virus dating effort, generally increasing statistical power. Specialized

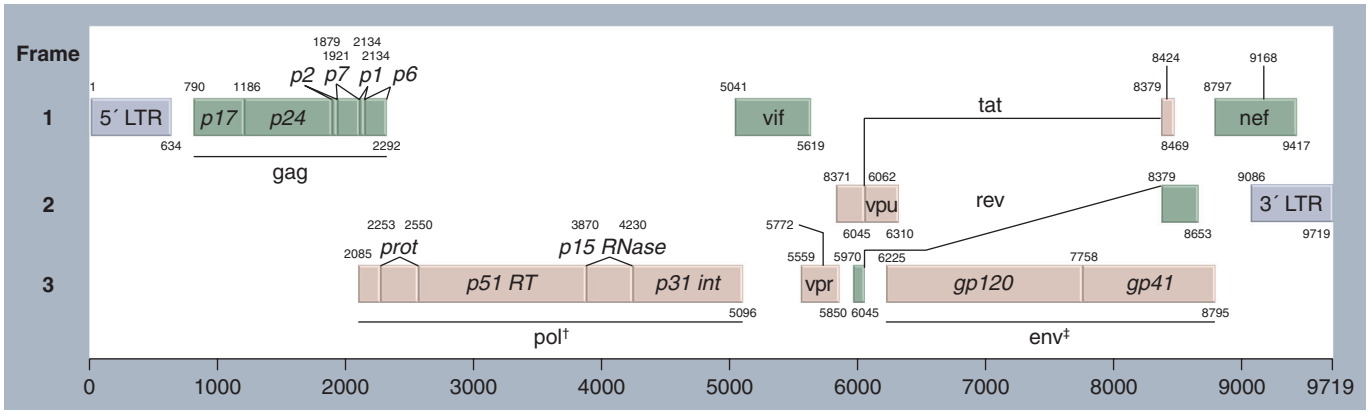

Figure 2. Genome organization. Schematic representation of HIV-1 genome organization. The three coding reading frames are depicted along with their open reading frames (rectangles). Genome position is numbered according to HXB2 reference strain. The small number in the upper left corner of each rectangle indicates the gene start, while the number in the lower right indicates the last position of the stop codon. Trans-spliced *rev* and *tat* forms are represented by black connecting lines between third and second, and second and first open reading frames, respectively.
†Open reading frames used for drug resistance testing.
‡Open reading frames used for subtyping and epidemiological studies.

Table 1. Summary of software used for phylodynamic inferences.

Inference	Implementation
Migration, spatial dispersion	Migrate-n [145], BEAST [146], IMa2, Lamarc [147]
Substitution rates	BEAST
Recombination rates and recombination breakpoints	Lamarc, LDhat [148], RDP3 [149]
Changes in population sizes	Migrate-n, BEAST
Divergence time estimation	BEAST, Multidivtime
Leaf ages	BEAST
Reproductive number	BEAST
Growth rate	BEAST, Lamarc, migrate-n
Population divergence times	IMa2
Haplotype reconstruction from NGS data and genetic diversity estimation	ShoRAH
Detection of selection	HyPhy [150], ADAPTSITE [151], TreeSAAP [152]
Assigning samples to populations, inferring the number of populations	Structurama, Structure, StructHDP
Ancestral state reconstruction	BEAST, MESQUITE [153]

NGS: Next-generation sequencing.

databases – for example, influenza [72] and HIV [201] – may help in this regard, as they can provide more clinically relevant information along with the genetic data (e.g., time of collection data).

In essence, most phylodynamic methods capitalize on analyzing distributions of trees usually obtained by sampling from the posterior distribution of a model, given the data. They work under the theoretical realization that the shape of a tree reflects dynamic processes impacting the data, such as population size changes (constant size, growth and shrinkage), selective processes (intrahost immune selection) or spatial structure (Figure 3). For example, populations with constant size are predicted to give symmetrical phylogenetic patterns, with most of the diversity happening at relative short branch lengths [73]. On the other hand, a population experiencing exponential growth will have longer branches leading to extant (sampled) sequences relative to deeper branches in the phylogeny [73]. In short, the coalescent models used in phylodynamic analyses describe a probability distribution on ancestral genealogies, given a population history. Therefore, if we can estimate the underlying phylogenetic structure of alleles from a collection of sequence data, by extension we can infer their population history and the evolutionary processes impacting that history (Figure 3).

Global spread

The best hypothesis we have regarding the origin and dispersion of HIV indicates that HIV-1 and -2 originated in Africa during the first half of the last century, and that it was the product of several cross-species transmissions between humans and non-human primates [11–13] (see [51] for further reading). Globally, approximately 33 million people were living with HIV worldwide as of 2009. In the same year, 2 million infected people died and the disease grew at a rate of 7400 new infections per day, more than 97% of which occurred in low- and middle-income countries [74]. Global distribution of groups and subtypes has remained rather constant within the last 10 years [75]. Although both HIV lineages spread exponentially at the beginning of the epidemic, HIV-2 occurred mostly in western Africa. In turn, HIV-1 is distributed worldwide, group M being the one that accounts for most infections, while groups O, N and P appear to be concentrated in central Africa. The time to the most recent common ancestors (TMRCAs) of HIV lineages has been dated using different molecular-based methodologies and gene regions with the following inferences: 1905–1942 for HIV-2 group A and 1914–1945 for group HIV-2 B [76,77]. Similarly, HIV-1 TMRCA estimates were as follows: 1894–1931 for group M, 1932–1966 for group N and 1914–1925 for group O [13,76,78].

Regional spread

Founder-effect events are thought to play a major role in the spread of HIV out of Africa, although other factors cannot be ruled out completely, such as viral selective advantages, sociocultural factors and human genetic background. The Democratic Republic of Congo (DRC) is one of the places in which HIV-1 diversity is the greatest, and probably the site where cross-species transmission occurred [79–81]. Two archival samples, DRC60 and ZR59 [82], and the existence in DRC of almost all group M subtypes [75,83] support this statement (Figure 4).

Studies worldwide have attempted to infer regional spatial and temporal spread of HIV-1 in particular. In the USA, HIV-1 B seems to have emerged from a single migration out of Haiti in 1969, the place with the highest subtype B diversity. In turn, Haitian HIV-1 B emerged in 1966 from the DRC [84–86].

South Africa has recently shown an increase in HIV infections. Almost 6 million people are infected, the majority with HIV-1 subtype C, which now accounts for 50% of all infected

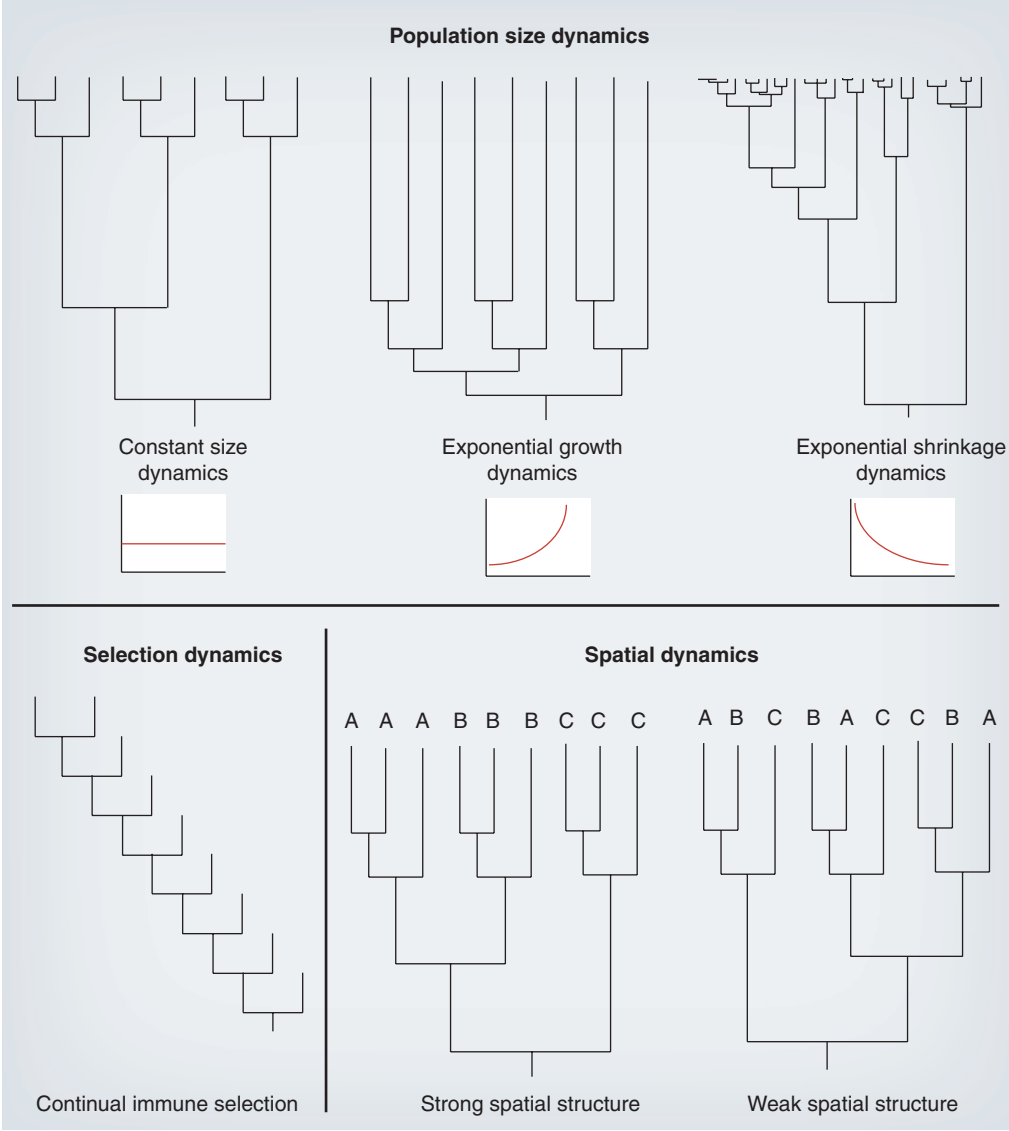

Figure 3. Phylodynamic patterns. Population, selection and spatial dynamic patterns and their respective idealized trees.

individuals worldwide [74,75]. The C subtype was first reported in 1990 and its TMRCA was dated to 1958 [87,88]. The spread of subtype C seems to have occurred eastward from South Africa to India [89,90], and also probably to China, while some founder events have also been identified from east Africa to South America and to Israel (Figure 4) [91,92]. The introduction of HIV-1 subtype C in South America goes back to the 1980s, most likely through Brazil [91,92]. Nonetheless, in a more comprehensive study, South American subtype C appears to be more related to UK subtype C, with these two groups related to east African isolates [71], which stresses the importance of including global isolates in phylogenetic studies of HIV phylogeography.

Besides being present in the western world, HIV-1 subtype B is also present in Asia, where its introduction seems to have occurred through Thailand in 1985 and was termed subtype B′ [93]. From here, HIV-1 subtype B′ expanded into Asia, coexisting with the pandemic subtype B and others, and fueling the development of CRFs across the continent [93–96]. It is worth noting that CRFs represent 20% of all HIV-1 infections, with half of these infections involving CRF02_AG and CRF01_AE [75]. Additionally, there are some indications that certain subtypes are preferentially associated with behavioral factors, such as intravenous drug users and drug-trafficking regions, in particular within ex-USSR countries and southeast Asia (subtype A and subtype CRF01_AE, respectively; Figure 4) [97].

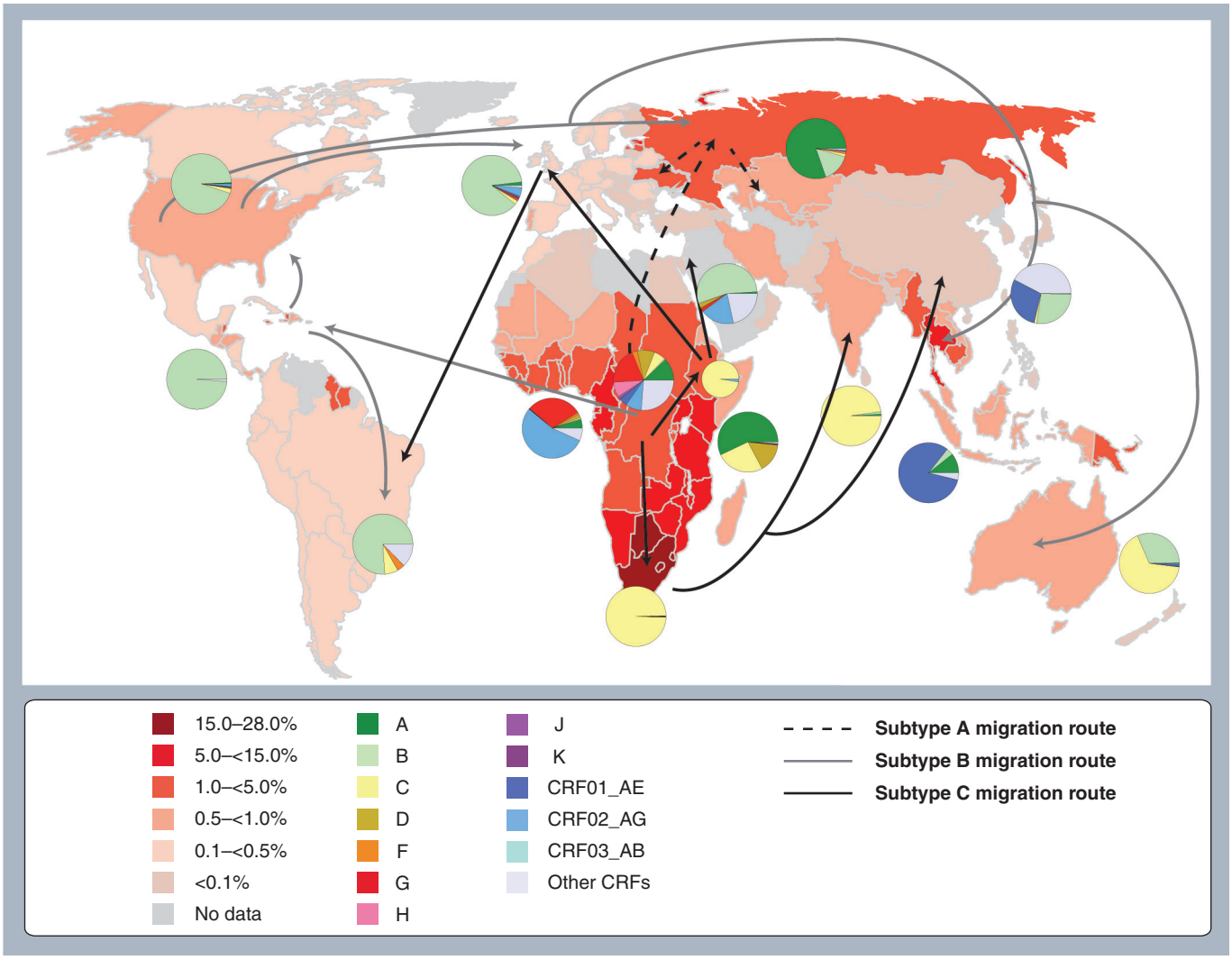

Figure 4. HIV-1 group M global distribution. Countries are color-coded according to their last reported prevalence [70]. Pie charts represent the distribution of subtypes and circulating recombinant forms over the globe. Arrows represent potential migration routes for A, B and C subtypes.
CRF: Circulating recombinant form.

HIV studies implementing phylodynamic methods have been used to address a variety of questions, including epidemics origins [84–90,93–96], correlations between epidemiological data and changes in population size or GD, viral spread over geographic regions, and evolutionary processes associated with certain risk groups, such as MSM and intravenous drug users. Some have shown that currently circulating subtype B has been introduced multiple times into MSM European (UK and Italy) populations in intervals ranging from 14 to 30 months [98–100]. Others have looked at dissemination patterns and transmission routes between risk groups in order to explain the distribution of HIV strains [101], suggesting a spread from heterosexual patients to intravenous drug users.

Local transmissions

Phylogenies are also useful for reconstructing transmission histories. For example, the Swedish transmission chain reconstructed transmissions in a case involving nine infected individuals, for whom the exact transmission history was known [102]. However, caution should be exercised, as convergence evolution might be a common theme in HIV-1 evolution [103]. Phylogenetic inference of transmission histories has also been accepted as evidence in court trials such as the Florida dentist case [104], a Swedish rape case [105], and healthcare-related cases in Baltimore [106] and Louisiana [107]. Bernard *et al.* present a detailed review on the use of HIV phylogenies in forensics [108].

HIV evolution has been thoroughly studied across transmissions because of the opportunity

for treatment due to a reduction in GD [7,109]. Most of the work has focused on monitoring discordant couples (i.e., couples in which one partner is HIV positive and the other is not) and to test whether there is a reduction in GD down to one virus at the transmission event. Using phylodynamic methods, Edwards *et al.* showed that GD reductions (<1%) are no different between horizontal (homo- or heterosexual) and vertical transmissions (mother-to-child) [109]. Understanding GD at transmission events has therapeutic implications, as less diverse populations of small size are strongly influenced by genetic drift, decreasing the chance of transmission of high-fitness variants. Indications that single virus variants were transmitted horizontally came from studies using Sanger sequencing and single-genome amplification coupled with phylogenetic and mathematical modeling [110]. These initial observations were additionally supported by the enhanced capabilities of 'ultra-deep sequencing' revealing that early HIV variants explored an extensive sequence space within epitope regions. Interestingly, as the infection proceeds, reversion to the canonical subtype sequence occurred in positions under immune selection pressure but not in positions that were not under selection pressure even in the earliest samples, suggesting that immune selection is acting earlier than previously thought [7].

Within-host dynamics

Different evolutionary forces are in action when we look at viral evolution within the host. The viral population is targeted by both cellular and humoral immune responses, resulting in relatively strong diversifying selection. As a result, when reconstructing within-host evolutionary histories, we observe a ladder-like pattern (continual immune selection in Figure 3 and [111]), as opposed to the strong spatial structure we observed in HIV population phylogenies (spatial dynamics in Figure 3). However, whether within-host HIV-1 evolution is governed by natural selection or genetic drift (deterministic or stochastic models, respectively) has been the subject of considerable debate. This dispute stems from disparate estimates of effective population sizes [112]; the smaller the population size, the more susceptible the population is to genetic drift. However, signatures of natural selection have been demonstrated using different estimators [113]. Recombination also plays an important role in shaping intrahost viral diversity. Conventionally, it can purge 'bad mutations' out of the gene pool and 'put together' novel combinations of genome regions, increasing allele diversity. Although it is not entirely clear whether recombination is associated with increased fitness, simulations have shown that under strong drug pressures, recombination will favor the appearance of drug resistance variants [114].

Within-host variation has been extensively studied because of the development of drug resistance, viral reservoirs and clinical conditions related to some tissues such as dementia and lymphoma. For example, phylodynamic analyses of *post mortem* brain tissues have revealed that HIV is evolving at different rates in different brain compartments. Apparently, this is not due to selective pressure, but rather to inherent drift associated with macrophage-tropic viral expansion after immune failure [115]. Within-host variation has been also addressed under a phylodynamic framework for coreceptor usage [116]. For example, coreceptor dynamics in tissues and peripheral blood mononuclear cells showed temporal structure between CCR5-tropic (R5) virus and the appearance of CXCR4 (X4) variants, suggesting that the majority of X4 virus found in thymus tissue seemed to come from peripheral blood mononuclear cell viruses [116]. Similarly, a high degree of compartmentalization and differences in population size have been described in HIV viruses from macrophages found in tumor and nontumor tissues and an intermixing of HIV strains obtained from axillary lymph nodes [117].

Population structure

Population structure refers to the degree of subdivision or differentiation a population exhibits. It has evolutionary consequences, as subdivided populations can evolve somewhat independently. Examples of structured HIV populations can be found in within-host samples from different tissues, which usually exhibit a large degree of compartmentalization [117], as well as in among-host samples, for instance founder effects and the global HIV-1 distribution (Figure 4). Traditional methods for estimating population structure include summary statistics that attempt to measure the diversity of randomly chosen sequence markers within the same subpopulation, relative to what is found in the entire population. These include the fixation index (F_{st}), as well as all its relatives (G_{st}, R_{st}, D_{st}). Summary statistical methods have been used to assess compartmentalization in lymphocyte reservoir populations [118,119].

New methodologies can also take advantage of the flexibility of the Bayesian framework to accommodate uncertainties in parameter estimation and to incorporate population history using the coalescent (see [120]). Bayesian approaches to infer population structure have been developed, for example, to estimate the posterior probability of viruses to belong to different populations (K) [121–123]. These methods could also be used to study HIV tissue differentiation, viral reservoirs, and their link to plasma GD.

Implications of the GD of HIV

Estimating the GD of HIV and mapping its distribution across space and time has clinical implications in terms of disease progression, drug resistance and vaccine development.

Disease progression & drug resistance

Disease progression is seemingly related to GD. Several studies have found a correlation between them when estimating relative and absolute substitution rates from time-stamped sequence data [10,124,125]. In general, there are three phases that can be identified on the basis of diversity: diversity increases linearly in association with initial features of infection and the appearance of X5 HIV populations; diversity levels off or even decreases, which could be correlated with the appearance of X4 HIV populations; and decline in $CD4^+$ T cells and failure of T-cell homeostasis with a general reduction in GD. However, Carvajal-Rodriguez *et al.* have reported no relationship between disease progression and substitution rates when analyzed separately as adaptive and neutral categories of variation [9].

Disease progression is also related to transmitted drug resistance (TDR) and the clinical impact of low-frequency variants. TDRs are thought to exhibit diminished fitness compared with 'wild-types', although additional compensatory mutations might restore fitness levels [126]. Low-frequency variants, typically characterized using NGS, have clinical significance, especially when the drug-resistant mutation has a low genetic barrier – that is, few mutations instead of several mutations that confer resistance. Non-nucleoside reverse transcriptase inhibitors and protease inhibitors are examples of low and high genetic barriers, respectively [127,128]. When using NGS, TDRs seem fairly common (~30% for B and non-B subtypes) and subjects with multiple protease TDRs are infrequent [129]. However, Sanger sequencing reveals values of TDR prevalence of only 10% [130]. Knowing what TDR estimate is accurate remains an open question.

Vaccine strategies

Designing an HIV vaccine is an extraordinarily difficult challenge. A successful vaccine is typically thought to be capable of eliciting broadly cross-reactive neutralizing antibodies in order to cope with the extreme GD of HIV [121]. New vaccine designs have considered accounting for HIV geographical distribution and HLA allele frequencies in different countries or regions [120]. This makes sense in the light of the geographic structure of HIV (see above). Thus, in southwest China a vaccine must be capable of neutralizing mainly HIV-1 M C/B′ recombinants, whereas a South African vaccine must consider, among others, HIV-1 M subtype C variants (Figure 4) [131,132]. Although this strategy seems sound, several considerations have led to a search for a global vaccine. Even in countries where the HIV epidemic is dominated by one subtype, for example, India, there are other subtypes that might increase its frequency if the vaccine targets just one variant. Also, and apart from cost concerns, nations with high HIV prevalence have complex epidemics, such as central and southern Africa, which will be less impacted by single subtype vaccines (pie charts in Figure 4). Alternative strategies have tried to account for HIV GD by looking for potential T-cell epitopes within the *env* 3D structure [133].

Novel strategies for vaccine design include, but are not limited to, poly-epitope vaccines, the use of conserved regions on the proteome, central vaccines (ancestral state, consensus and center-of-the-tree) and polyvalent mosaic vaccines. For details, the reader is referred to Korber *et al.* [134]. Central vaccines are appealing from a phylogenetics standpoint and have also proven useful [135]. The idea is to maximize GD coverage by inferring the ancestral sequence of a particular clade or tree region, under the assumption that such a sequence, once translated, will elicit antibodies against any and every member of the clade or group of interest. In effect, central vaccines are designed to minimize the genetic distance to circulating strains. Other strategies include immunization against cell coreceptors such as the CCR5 chemokine receptor [136–138].

Recent research has focused on characterizing broadly neutralizing antibodies that occur naturally. Independently, and using different approaches, Wu *et al.* [139] and Scheid *et al.* [140] converged in identifying CD4 binding sites in antibodies derived from multiple unrelated HIV-infected individuals. These promising results

have obvious vaccine design implications as broadly neutralizing antibodies can protect new cells from infection.

Conclusion

Understanding the geographical and temporal spread of HIV GD is crucial for planning effective intervention strategies to combat infection, whether through educational programs, drug treatments or vaccine strategies. Is there a sequence type that is preferentially transmitted? If yes, is this sequence type similar among different transmission routes? What is the GD that needs to be covered by a vaccine? What proportion of TDR variants is clinically relevant? New sequencing technologies have the potential to improve our understanding of the patterns and processes that lead to the temporal and spatial spread of HIV and to answer the abovementioned questions. Methods such as ultra-deep sequencing have already impacted our understanding of HIV in terms of within-host variation: emergence and transmission of drug resistance, identification of super-infections, characterization of early dynamics and immune escape. At the population level, NGS has provided data to test for coevolving sites and added to our knowledge regarding the clinical impact of low-frequency variants. Then, phylogenetic and population genetic approaches have provided an appropriate probabilistic framework to analyze NGS data and understand better the impact of genetic diversity.

Future perspective

NGS data have already demonstrated their utility in HIV biology; nevertheless, there are sequencing applications and aspects of phylodynamics that are presently beyond the reach of current sequencing technologies, paving the way for additional innovation. It is imperative for the continuous progress of phylodynamics, then, to capitalize and incorporate in a 'biologist-accessible' manner the power of this technology and of the just-arriving 'third-generation sequencing technologies', namely single-molecule sequencing (PacBio, Helicos) and semiconductor sequencing technology (Ion Torrent) [141]. Additionally, an increase in the number of epidemiological models, coupled with population genetic and phylogenetic models, will provide researchers with more flexibility when applying these tools. Examples of this are models borrowed from epidemiology [67] and other recent advances in estimating epidemiological parameters directly from sequence data [68,69]. For instance, models for cophylodynamics among coinfecting viruses that are transmitted similarly, notably HIV and HCV, will be necessary to explore the still controversial consequences of this coinfection [3,142–144]. Over the next few years, this race to generate large amounts of data should be coupled with a similar effort to expand our models of HIV evolution to incorporate these data into the analysis of HIV dynamics and disease progression.

Executive summary

- The high genetic diversity of HIV stems from at least three different sources, namely high mutation and recombination rates, high replication error and multiple introductions into human populations.
- Traditional measures of genetic diversity (summary statistics and coalescent estimators) have helped to understand HIV diversity and distribution. Next-generation sequencing is providing a new dimension to explore within-host and population diversity in depth.
- The integration of epidemiological and evolutionary processes through the use of sequence data under a phylogenetic framework (phylodynamic) has allowed us to investigate changes in population size over time, estimate relevant parameters as the reproductive number and substitution rates, and most notably, the dispersion of lineages over a geographic region.
- Through the use of phylogenetics, we now know that HIV-1 and -2 originated in central and west Africa, respectively. HIV-1, sub-classified in groups and subtypes, is present worldwide and exhibits a large geographic distribution, probably because of several founder-effect events.
- Evolutionary forces within and among hosts differ. Within-host HIV-1 evolution seems to be governed by continual immune selection, whereas among-hosts evolution seems to be driven by founder effects, as reflected in the strong geographic structure observed.
- Increased HIV genetic diversity has implications in almost every aspect of its biology, including vaccine design, drug treatment, disease progression, viral reservoirs, transmissibility and viral load testing. It also allows us to apply ambitious statistic models to interrogate these very same aspects. On the other hand, its high genetic diversity has prompted HIV as a model entity to test new developments in phylogenetics and evolutionary theory.
- Over the next few years, phylodynamic studies will couple with next-generation sequencing and third-generation sequencing to integrate the large amount of data into a robust and informative framework for testing hypotheses of HIV evolution, epidemiology and diversity. In turn, HIV biologists will need to better understand the strengths and weaknesses of these evolutionary and population genetic approaches to gain insights as applied to HIV data.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:
- ▪ of interest
- ▪▪ of considerable interest

1. Laikre L, Allendorf FW, Aroner LC *et al.* Neglect of genetic diversity in implementation of the convention on biological diversity. *Conserv. Biol.* 24(1), 86–88 (2010).
2. Tishkoff SA, Reed FA, Friedlaender FOR *et al.* The genetic structure and history of Africans and African Americans. *Science* 324(5930), 1035–1044 (2009).
3. De Oliveira T, Pybus O, Rambaut A *et al.* Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* 444(7121), 836–837 (2006).
4. Groeneveld L, Lenstra J, Eding H *et al.* Genetic diversity in farm animals – a review. *Anim. Genet.* 41(Suppl. 1), 6–31 (2010).
5. Agashe D, Falk J, Bolnick D. Effects of founding genetic variation on adaptation to a novel resource. *Evolution* 65(9), 2481–2491 (2011).
6. Merlo LM, Maley CC. The role of genetic diversity in cancer. *J. Clin. Invest.* 120(2), 401–403 (2010).
7. Fischer W, Ganusov V, Giorgi E *et al.* Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5(8), e12303 (2010).
- ▪ **Characterization of early HIV dynamics by ultra-deep sequencing, showing the reduction of genetic diversity at the point of transmission and the early diversification of transmitted viruses.**
8. Ross LL, Weinberg WG, Dejesus E *et al.* Impact of low abundance HIV variants on response to ritonavir-boosted atazanavir or fosamprenavir given once daily with tenofovir/emtricitabine in antiretroviral-naive HIV-infected patients. *AIDS Res. Hum. Retrov.* 26(4), 407–417 (2010).
9. Carvajal-Rodriguez A, Posada D, Pérez-Losada M *et al.* Disease progression and evolution of the HIV-1 *env* gene in 24 infected infants. *Infect. Genet. Evol.* 8(2), 110–120 (2008).
10. Lemey P, Pond SLK, Drummond AJ *et al.* Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* 3(2), 282–292 (2007).
11. Salemi M, Strimmer K, Hall WW *et al.* Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB J.* 15(2), 276–278 (2001).
12. Sharp P, Bailes E, Gao F, Beer B, Hirsch V, Hahn B. Origins and evolution of AIDS viruses: estimating the time-scale. *Biochem. Soc. Trans.* 28(2), 275–282 (2000).
13. Korber B, Muldoon M, Theiler J *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* 288(5472), 1789–1796 (2000).
14. Hu W, Temin H. Retroviral recombination and reverse transcription. *Science* 250(4985), 1227–1233 (1990).
15. Shriner D, Rodrigo AG, Nickle DC, Mullins JI. Pervasive genomic recombination of HIV-1 *in vivo*. *Genetics* 167(4), 1573–1583 (2004).
16. Preston B, Poiesz B, Loeb L. Fidelity of HIV-1 reverse transcriptase. *Science* 242(4882), 1168–1171 (1988).
17. Wei X, Ghosh SK, Taylor ME *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 373(6510), 117–122 (1995).
18. Holmes EC. On the origin and evolution of the human immunodeficiency virus (HIV). *Biol. Rev.* 76(2), 239–254 (2001).
19. Plantier J, Leoz M, Dickerson J *et al.* A new human immunodeficiency virus derived from gorillas. *Nat. Med.* 15(8), 871–872 (2009).
20. Robertson DL, Anderson JP, Bradac JA *et al.* HIV-1 nomenclature proposal. *Science* 288(5463), 55 (2000).
21. Ibe S, Yokomaku Y, Shiino T *et al.* HIV-2 CRF01_AB: first circulating recombinant form of HIV-2. *J. Acquir. Immune Defic. Syndr.* 54(3), 241–247 (2010).
- ▪ **Circulating recombinant forms also occur in HIV-2. Using phylogenetic methods, the authors demonstrated the recombinant nature of HIV-2 isolates.**
22. Cheng Z, Ventura M, She X *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437(7055), 88–93 (2005).
23. Peeters M, Sharp PM. Genetic diversity of HIV-1: the moving target. *AIDS* 14, S129–S140 (2000).
24. Barroso H, Borrego P, Bartolo I *et al.* Evolutionary and structural features of the C2, V3 and C3 envelope regions underlying the differences in HIV-1 and HIV-2 biology and infection. *PLoS ONE* 6(1), e14548 (2011).
25. Kanki PJ, Travers KU, Marlink RG *et al.* Slower heterosexual spread of HIV-2 than HIV-1. *Lancet* 343(8903), 943–946 (1994).
26. Leligdowicz A, Yindom L-M, Onyango C *et al.* Robust Gag-specific T cell responses characterize viremia control in HIV-2 infection. *J. Clin. Invest.* 117(10), 3067–3074 (2007).
27. Kanki PJ, Hamel DJ, Sankale JL *et al.* Human immunodeficiency virus type 1 subtypes differ in disease progression. *J. Infect. Dis.* 179(1), 68–73 (1999).
28. Tscherning C, Alaeus A, Fredriksson R *et al.* Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. *Virology* 241(2), 181–188 (1998).
29. Kiwanuka N, Laeyendecker O, Quinn TC *et al.* HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda. *AIDS* 23(18), 2479–2484 (2009).
30. Abecasis AB, Lemey P, Vidal N *et al.* Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. *J. Virol.* 81(16), 8543–8551 (2007).
31. Tajima F. Evolutionary relationship of DNA-sequences in finite populations. *Genetics* 105(2), 437–460 (1983).
32. Nei M. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* 30, 371–403 (1996).
33. Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7(2), 256–276 (1975).
34. Kuhner MK. Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.* 24(2), 86–93 (2009).
- ▪▪ **Summary of diversity and feature of current coalescent-based estimators of phylodynamic-related parameter. Good reference to compare current implementations.**
35. Kuhner MK, Smith LP. Comparing likelihood and Bayesian coalescent estimation of population parameters. *Genetics* 175(1), 155–165 (2007).
36. Kuhner MK. Robustness of coalescent estimators to between-lineage mutation rate variation. *Mol. Biol. Evol.* 23(12), 2355–2360 (2006).

37. Kuhner MK, Yamato J, Felsenstein J. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140, 1421–1430 (1995).
38. Crandall K, Posada D, Vasco D. Effective population sizes: missing measures and missing concepts. *Anim. Conserv.* 2(4), 317–319 (1999).
39. Wang J. Estimation of effective population sizes from data on genetic markers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360(1459), 1395–1409 (2005).
40. Schwartz MK, Tallmon DA, Luikart G. Review of DNA-based census and effective population size estimators. *Anim. Conserv.* 1(4), 293–299 (1998).
41. Metzker ML. Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11(1), 31–46 (2010).

▪▪ **Excellent review covering next-generation sequencing technologies, their pros and cons and applications.**

42. Vrancken B, Lequime S, Theys K, Lemey P. Covering all bases in HIV research: unveiling a hidden world of viral evolution. *AIDS Rev.* 12(2), 89–102 (2010).
43. Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.* 38(21), 7400–7409 (2010).
44. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol.* 17(3), 417– 428 (2010).
45. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12(1), 119 (2011).

▪ **Original description of an algorithm that can take next-generation sequencing data and return genetic diversity measures and haplotype number and frequencies using a Bayesian framework.**

46. Willerth SM, Pedro HAM, Pachter L, Humeau LM, Arkin AP, Schaffer DV. Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS ONE* 5(10), e13564 (2010).
47. Henn M, Boutwell C, Lennon N *et al.* P09–20 LB. Ultra-deep sequencing of full-length HIV-1 genomes identifies rapid viral evolution during acute infection. *Retrovirology* 6(Suppl. 3), P400 (2009).
48. Fischer W, Keele B, Bhattacharya T *et al.* P09–21 LB. Deep sequencing of HIV-1 from acute infection: low initial diversity, and rapid but variable CTL escape. *Retrovirology* 6(3), P401 (2009).
49. Poon AFY, Swenson LC, Dong WWY *et al.* Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the *nef* gene of HIV-1. *Mol. Biol. Evol.* 27(4), 819–832 (2010).
50. Redd AD, Collinson-Streng A, Martens C *et al.* Identification of HIV superinfection in seroconcordant couples in Rakai, Uganda using next generation deep sequencing. *J. Clin. Microbiol.* 49(8), 2859–2867 (2011).
51. Castro-Nallar E, Pérez-Losada M, Burton GF, Crandall KA. The evolution of HIV: inferences using phylogenetics. *Mol. Phylogenet. Evol.* 62(2), 777–792 (2012).
52. Kosakovsky Pond SL, Posada D, Stawiski E *et al.* An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput. Biol.* 5(11), e1000581 (2009).
53. Willerth SM, Pedro HLaM, Pachter L, Humeau LM, Arkin AP, Schaffer DV. Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS ONE* 5(10), e13564 (2010).
54. Grenfell BT, Pybus OG, Gog JR *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656), 327–332 (2004).
55. Drummond A, Ho S, Phillips M, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4(5), 699–710 (2006).
56. Thorne JL, Kishino H. Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* 51(5), 689–702 (2002).
57. Drummond A, Suchard M. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8, 114 (2010).
58. Drummond A, Rambaut A, Shapiro B, Pybus O. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22(5), 1185–1192 (2005).
59. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* 27(8), 1877–1885 (2010).
60. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5(9), e1000520 (2009).
61. Beerli P, Felsenstein J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl Acad. Sci. USA* 98(8), 4563–4568 (2001).
62. Hey J. Isolation with migration models for more than two populations. *Mol. Biol. Evol.* 27(4), 905–920 (2010).
63. Shapiro B, Ho SYW, Drummond AJ, Suchard MA, Pybus OG, Rambaut A. A Bayesian phylogenetic method to estimate unknown sequence ages. *Mol. Biol. Evol.* 28(2), 879–887 (2011).
64. Choi S, Hey J. Joint inference of population assignment and demographic history. *Genetics* 189(2), 561–577 (2011).
65. Huelsenbeck JP, Andolfatto P. Inference of population structure under a Dirichlet process model. *Genetics* 175(4), 1787–1802 (2007).
66. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 155(2), 945–959 (2000).
67. Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SDW. Phylodynamics of infectious disease epidemics. *Genetics* 183(4), 1421–1430 (2009).

▪ **Excellent article that deals with phylodynamic patterns and their associated processess. It also describes algorithms to integrate epidemiological and phylogenetic models.**

68. Stadler T. Sampling-through-time in birth-death trees. *J. Theor. Biol.* 267(3), 396–404 (2010).
69. Stadler T, Kouyos R, Von Wyl V *et al.* Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* 29(1), 347–357 (2012).
70. Stack JC, Welch JD, Ferrari MJ, Shapiro BU, Grenfell BT. Protocols for sampling viral sequences to study epidemic dynamics. *J. R. Soc. Interface* 7(48), 1119–1127 (2010).
71. de Oliveira T, Pillay D, Gifford R; the UK Collaborative Group on HIV Drug Resistance. The HIV-1 subtype C epidemic in South America is linked to the United Kingdom. *PLoS ONE* 5(2), E9311 (2010).
72. Bao Y, Bolotov P, Dernovoy D *et al.* The influenza virus resource at the national center for biotechnology information. *J. Virol.* 82(2), 596–601 (2008).
73. Frost SDW, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365(1548), 1879–1890 (2010).
74. UNAIDS. *Global report: UNAIDS report on the global AIDS epidemic 2010.* UN, Geneva, Switzerland (2010).

75. Hemelaar J, Gouws E, Ghys PD, Osmanov S; WHO–UNAIDS. Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS* 25(5), 679–689 (2011).
76. Wertheim JO, Worobey M. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput. Biol.* 5(5), e1000377 (2009).
77. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, Vandamme A-M. Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl Acad. Sci. USA* 100(11), 6588–6592 (2003).
78. Lemey P, Pybus OG, Rambaut A *et al.* The molecular population genetics of HIV-1 group O. *Genetics* 167(3), 1059–1068 (2004).
79. Gao F, Bailes E, Robertson D *et al.* Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* 397(6718), 436–441 (1999).
80. Santiago ML, Bibollet-Ruche F, Bailes E *et al.* Amplification of a complete simian immunodeficiency virus genome from fecal RNA of a wild chimpanzee. *J. Virol.* 77(3), 2233–2242 (2003).
81. Santiago ML, Rodenburg CM, Kamenya S *et al.* SIVcpz in wild chimpanzees. *Science* 295(5554), 465 (2002).
82. Worobey M, Gemmel M, Teuwen DE *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455(7213), 661–664 (2008).
83. Vidal N, Peeters M, Mulanga-Kabeya C *et al.* Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in central Africa. *J. Virol.* 74(22), 10498–10507 (2000).
84. Gilbert M, Rambaut A, Wlasiuk G, Spira T, Pitchenik A, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl Acad. Sci. USA* 104(47), 18566–18570 (2007).
85. Pérez-Losada M, Jobes DV, Sinangil F, Crandall KA, Posada D, Berman PW. Phylodynamics of HIV-1 from a Phase-III AIDS vaccine trial in North America. *Mol. Biol. Evol.* 27(2), 417–425 (2010).
86. Robbins KE, Lemey P, Pybus OG *et al.* U.S. human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.* 77(11), 6359–6366 (2003).
87. Rousseau CM, Learn GH, Bhattacharya T *et al.* Extensive intrasubtype recombination in South African human immunodeficiency virus type I subtype C infections. *J. Virol.* 81(9), 4492–4500 (2007).
88. Ayehunie S, Johansson B, Sonnerborg A *et al.* New subtype of HIV-1 in Ethiopia. *Lancet* 336(8720), 942 (1990).
89. Sridharan G, Kandathil AJ, Ramalingam S, Kannangai R, David S. Molecular epidemiology of HIV. *Indian J. Med. Res.* 121(4), 333–344 (2005).
90. Shankarappa R, Gupta P, Chatterjee R *et al.* Human immunodeficiency virus type 1 env sequences from Calcutta in eastern India: Identification of features that distinguish subtype C sequences in India from other subtype C sequences. *J. Virol.* 75(21), 10479–10487 (2001).
91. Fontella R, Schrago C, Soares M. On the origin of HIV-1 subtype C in South America. *AIDS* 22(15), 2001–2011 (2008).
92. Bello G, Passaes C, Guimaraes ML *et al.* Origin and evolutionary history of HIV-1 subtype C in Brazil. *AIDS* 22(15), 1993–2000 (2008).
93. Deng X, Yang R, Liu H, Shao Y, Rayner S. The epidemic origin and molecular properties of B′: a founder strain of the HIV-1 transmission in Asia. *AIDS* 22(14), 1851–1858 (2008).
94. Piyasirisilp S, Mccutchan FE, Carr JK *et al.* A recent outbreak of human immunodeficiency virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. *J. Virol.* 74(23), 11286–11295 (2000).
95. De Silva U, Warachit J, Sattagowit N *et al.* Genotypic characterization of HIV type 1 env gp160 sequences from three regions in Thailand. *AIDS Res. Hum. Retrov.* 26(2), 223–227 (2010).
96. Su L, Graf M, Zhang YZ *et al.* Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B′) recombinant strain in China. *J. Virol.* 74(23), 11367–11376 (2000).
97. Skar H, Hedskog C, Albert J. HIV-1 evolution in relation to molecular epidemiology and antiretroviral resistance. *Ann. NY Acad. Sci.* 1230, 108–118 (2011).
98. Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl Acad. Sci. USA* 102(12), 4425–4429 (2005).
- **Phylodynamic study that attempts to untangle the complex structure of HIV in populations of men who have sex with men in the UK. It shows that HIV has been introduced into that population several times.**
99. Zehender G, Ebranati E, Lai A *et al.* Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. *J. Acquir. Immune Defic. Syndr.* 55, 156–160 (2010).
100. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Brown AJL. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 5(3), 392–402 (2008).
101. Liao HA, Tee KK, Hase S *et al.* Phylodynamic analysis of the dissemination of HIV-1 CRF01_AE in Vietnam. *Virology* 391(1), 51–56 (2009).
102. Leitner T, Escanilla D, Franzén C, Uhlénn M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl Acad. Sci. USA* 93(20), 10864–10869 (1996).
103. Lemey P, Derdelinckx I, Rambaut A *et al.* Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J. Virol.* 79(18), 11981–11989 (2005).
104. Ou C-Y, Ciesielski CA, Myers G *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* 256(5060), 1165–1171 (1992).
105. Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlen M. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 *pol* and *gag* genes. *J. Virol.* 68(9), 5918–5924 (1994).
106. Holmes EC, Zhang LQ, Simmonds P, Rogers AS, Brown AJL. Molecular investigation of human-immunodeficiency-virus (HIV) infection in a patient of an HIV-infected surgeon. *J. Infect. Dis.* 167(6), 1411–1414 (1993).
107. Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM. Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl Acad. Sci. USA* 99(22), 14292–14297 (2002).
108. Bernard E, Azad Y, Vandamme A, Weait M, Geretti A. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Med.* 8(6), 382–387 (2007).
109. Edwards C, Holmes E, Wilson D *et al.* Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol. Biol.* 6, 28 (2006).
- **Relevant study showing the loss of GD across vertical and horizontal transmissions. It also illustrates the use of phylodynamic approaches to answer clinically important questions.**

110. Salazar-González JF, Bailes E, Pham KT *et al.* Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J. Virol.* 82(8), 3952–3970 (2008).

111. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453(7195), 615-619 (2008).

112. Brown AJL. Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. *Proc. Natl Acad. Sci. USA* 94(5), 1862–1865 (1997).

113. Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* 16(3), 372–382 (1999).

114. Carvajal-Rodríguez A, Crandall K, Posada D. Recombination favors the evolution of drug resistance in HIV-1 during antiretroviral therapy. *Infect. Genet. Evol.* 7(4), 476–483 (2007).

115. Salemi M, Lamers SL, Yu S, De Oliveira T, Fitch WM, McGrath MS. Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol.* 79(17), 11343–11352 (2005).

116. Salemi M, Burkhardt BR, Gray RR, Ghaffari G, Sleasman JW, Goodenow MM. Phylodynamics of HIV-1 in lymphoid and non-lymphoid tissues reveals a central role for the thymus in emergence of CXCR4-using quasispecies. *PLoS ONE* 2(9), e950 (2007).

117. Salemi M, Lamers SL, Huysentruyt LC *et al.* Distinct patterns of HIV-1 evolution within metastatic tissues in patients with non-hodgkins lymphoma. *PLoS ONE* 4(12), e8153 (2009).

118. Potter SJ, Lemey P, Achaz G *et al.* HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. *J. Leukoc. Biol.* 76(3), 562–570 (2004).

119. Potter SJ, Lemey P, Dyer WB *et al.* Genetic analyses reveal structured HIV-1 populations in serially sampled T lymphocytes of patients receiving HAART. *Virology* 348(1), 35–46 (2006).

120. Pearse DE, Crandall KA. Beyond FST: analysis of population genetic data for conservation. *Conserv. Genet.* 5(5), 585–602 (2004).

121. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9(5), 1322–1332 (2009).

122. Huelsenbeck JP, Andolfatto P, Huelsenbeck ET. Structurama: Bayesian inference of population structure. *Evol. Bioinform.* 7, 55–59 (2011).

123. Shringarpure S, Won D, Xing EP. StructHDP: automatic inference of number of clusters and population structure from admixed genotype data. *Bioinformatics* 27(13), i324–i332 (2011).

124. Shankarappa R, Margolick JB, Gange SJ *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* 73(12), 10489–10502 (1999).

125. Lee HY, Perelson AS, Park SC, Leitner T. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput. Biol.* 4(12), e1000240 (2008).

126. Llibre JM, Schapiro JM, Clotet B. Clinical implications of genotypic resistance to the newer antiretroviral drugs in HIV-1-infected patients with virological failure. *Clin. Infect. Dis.* 50(6), 872–881 (2010).

127. Van Laethem K, De Munter P, Schrooten Y *et al.* No response to first-line tenofovir+lamivudine+efavirenz despite optimization according to baseline resistance testing: impact of resistant minority variants on efficacy of low genetic barrier drugs. *J. Clin. Virol.* 39(1), 43–47 (2007).

128. Pingen M, Nijhuis M, De Bruijn JA, Boucher CAB, Wensing AMJ. Evolutionary pathways of transmitted drug-resistant HIV-1. *J. Antimicrob. Chemother.* 66(7), 1467–1480 (2011).

129. Lataillade M, Chiarella J, Yang R *et al.* Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naive subjects in the CASTLE study. *PLoS ONE* 5(6), e10952 (2010).

130. Wensing AM, Van De Vijver DA, Angarano G *et al.* Prevalence of drug-resistant HIV-1 variants in untreated individuals in Europe: implications for clinical management. *J. Infect. Dis.* 192(6), 958–966 (2005).

131. Burgers WA, Shephard E, Monroe JE *et al.* Construction, characterization, and immunogenicity of a multigene modified vaccinia Ankara (MVA) vaccine based on HIV type 1 subtype C. *AIDS Res. Hum. Retroviruses* 24(2), 195–206 (2008).

132. Chen Z, Huang Y, Zhao X, Ba L, Zhang W, Ho DD. Design, construction, and characterization of a multigenic modified vaccinia Ankara candidate vaccine against human immunodeficiency virus type 1 subtype C/B′. *J. Acquir. Immune Defic. Syndr.* 47(4), 412–421 (2008).

133. Korber BT, Letvin NL, Haynes BF. T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. *J. Virol.* 83(17), 8300–8314 (2009).

▪▪ **Excellent review covering novel approaches to vaccine design and their known performances. It summarizes novel attempts to cope with HIV GD in the design of vaccines.**

134. Korber B, Gnanakaran S. The implications of patterns in HIV diversity for neutralizing antibody induction and susceptibility. *Curr. Opin. HIV AIDS* 4(5), 408–417 (2009).

135. Rolland M, Jensen MA, Nickle DC *et al.* Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. *J. Virol.* 81(16), 8507–8514 (2007).

136. Sodora DL, Allan JS, Apetrei C *et al.* Toward an AIDS vaccine: lessons from natural simian immunodeficiency virus infections of African nonhuman primate hosts. *Nat. Med.* 15(8), 861–865 (2009).

137. Nakayama D, Misumi S, Mukai R *et al.* Suppression of multiclade R5 and X4 human immunodeficiency virus type-1 infections by a coreceptor-based anti-HIV strategy. *J. Biochem.* 138(5), 571–582 (2005).

138. Misumi S, Nakayama D, Kusaba M *et al.* Effects of immunization with CCR5-based cycloimmunogen on simian/HIVSF162P3 challenge. *J. Immunol.* 176(1), 463–471 (2006).

139. Wu X, Zhou T, Zhu J *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333(6049), 1593–1602 (2011).

140. Scheid JF, Mouquet H, Ueberheide B *et al.* Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333(6049), 1633–1637 (2011).

141. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum. Mol. Genet.* 19(R2), R227–R240 (2010).

142. Smit C, Van Den Berg C, Geskus R, Berkhout B, Coutinho R, Prins M. Risk of hepatitis-related mortality increased among hepatitis C virus/HIV-coinfected drug users compared with drug users infected only with hepatitis C virus: a 20-year prospective study. *J. Acquir. Immune Defic. Syndr.* 47(2), 221–225 (2008).

143. Chen T-Y, Ding EL, Seage GR, Kim AY. Meta-analysis: increased mortality associated with hepatitis C in HIV-infected persons is unrelated to HIV disease progression. *Clin. Infect. Dis.* 49(10), 1605–1615 (2009).

144. Van De Laar TJ, Matthews GV, Prins M, Danta M. Acute hepatitis C in HIV-infected men who have sex with men: an emerging sexually transmitted infection. AIDS 24(12), 1799–1812 (2010).

145. Beerli P, Palczewski M. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* 185(1), 313–326 (2010).

146. Drummond A, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7(1), 214 (2007).

147. Kuhner MK. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22(6), 768–770 (2006).

148. Auton A, Mcvean G. Recombination rate estimation in the presence of hotspots. *Genome Res.* 17(8), 1219–1227 (2007).

149. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26(19), 2462–2463 (2010).

150. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21(5), 676–679 (2005).

151. Suzuki Y, Gojobori T, Nei M. ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics* 17(7), 660–661 (2001).

152. Woolley S, Johnson J, Smith MJ, Crandall KA, Mcclellan DA. TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics* 19(5), 671–672 (2003).

153. Maddison W, Maddison D. Mesquite: A modular system for evolutionary analysis. *Evolution* 62, 1103–1118 (2005).

Website

201. HIV Databases. www.hiv.lanl.gov