Molecular evolution methods to study HIV-1 epidemics

Juan Á Patiño-Galindo¹ & Fernando González-Candelas^{*,2,3}

¹Department of Systems Biology, Columbia University, New York, NY 10032, USA

²Joint Research Unit "Infección y Salud Pública" FISABIO-Salud Pública/Universitat de València-Institute for Integrative Systems Biology (I²SysBio, CSIC-UV) Valencia, Spain

³CIBER in Epidemiology & Public Health, Valencia, Spain

* Author for correspondence: fernando.gonzalez@uv.es

Nucleotide sequences of HIV isolates are obtained routinely to evaluate the presence of resistance mutations to antiretroviral drugs. But, beyond their clinical use, these and other viral sequences include a wealth of information that can be used to better understand and characterize the epidemiology of HIV in relevant populations. In this review, we provide a brief overview of the main methods used to analyze HIV sequences, the data bases where reference sequences can be obtained, and some caveats about the possible applications for public health of these analyses, along with some considerations about their limitations and correct usage to derive robust and reliable conclusions.

First draft submitted: 25 December 2017; Accepted for publication: 4 April 2018; Published online: 21 May 2018

Keywords: HIV • molecular epidemiology • phylogenetics • transmission cluster

Status of the HIV pandemic

HIV-1, the causing agent of AIDS, is an enveloped virus of the *Retroviridae* family and genus *Lentivirus* whose single-stranded positive-sense RNA genome is slightly longer than 9700 nucleotides. HIV infections are one of the most important concerns for public health worldwide. Since the introduction of highly active antiviral therapy, the global incidence of new HIV cases and AIDS deaths has slowed down, especially in countries with highly active antiviral therapy coverage. Nonetheless, the number of HIV-infected persons has been recently estimated to be nearly 37 million, with almost 2 million new infections in 2016 [1]. The HIV epidemic has changed through time among the different risk groups and practices. So, in the European Union and European Economic Area the incidence of new HIV infections among men who have sex with men is experiencing a sustained increase since the early 2000s, while the incidence has stabilized or even declined among other risk groups, such as intravenous drug users and people infected through heterosexual sex [2,3].

Molecular approaches to study HIV epidemics

Molecular epidemiology is a very powerful tool to analyze the epidemics of infectious diseases in depth, by adding molecular data from the pathogen to the clinical and epidemiological information from the host. This branch of epidemiology can provide answers to important questions of public health interest, such as the typing of viral populations and the discovery of new recombinant variants, the testing of resistance to antivirals in newly diagnosed patients, or the study of the virus transmission dynamics [4]. In the particular case of HIV, traditional epidemiological analyses based on contact tracing or interviews with newly diagnosed patients can be hampered by the usually long time elapsed from infection to diagnosis, and also by the social stigma associated with HIV, which altogether result in lack of useful epidemiological information. Importantly, molecular epidemiology analyses have been used to detect transmission clusters that had spread unnoticed to nonmolecular epidemiological analyses and tools [5–7]. Molecular epidemiology analyses have also been able to detect the expansion of drug-resistant HIV variants among the drug-naïve population [8–10].

Typing HIV variants circulating in a specific population is clinically relevant for different reasons, such as the fact that some subtype-specific resistance mutations are not always detected by the resistance algorithms [11], or

that some subtypes and circulating recombinant forms (CRFs) have been associated with a faster progression to AIDS and their expansion could hamper the control of the local epidemic [12–14]. In addition, the genetic and antigenic differences between the different HIV-1 types are important concerns for the development of an effective vaccine [15,16]. Although HIV can be typed by means of phylogenetic analyses that include the query sequences and a representative set of the different subtypes and CRFs of HIV (reference sequences), it is usually performed with online tools such as COMET (<http://comet.retrovirology.lu/>) or REGA [8]. These tools allow not only subtype assignment, but also the identification of new recombination patterns along the viral genome (i.e., detection of new unique recombinant forms). Another tool commonly used for subtyping is the Stanford HIV Resistance Database [17,18], which also allows to detect mutations associated with resistance to HIV antivirals. This database uses partial genome sequences of the virus (regions of the *pol* gene). It identifies previously described resistance mutations of each HIV sequence and generates a profile of susceptibility to different HIV antiretroviral drugs. A comparative analysis of the different subtyping tools [10] found that phylogenies including updated reference sets performed best, although the use of a combination of the aforementioned tools also worked well for clinical settings.

Another clinically relevant information from HIV can be obtained by sequencing portions of the *env* gene, where the cell tropism of the virus is determined. These sequences can be analyzed with tools such as Geno2Pheno (<http://correceptor.bioinf.mpi-inf.mpg.de/>) [19].

In order to study the transmission dynamics of a virus, it is important to do a surveillance as complete as possible, ensuring an accurate representation of the population to be studied [20]. For HIV, datasets comprising hundreds or thousands of sequences from different patients have been obtained in the last years from genotypic tests of resistance to antiviral drugs. Based on sequencing the PR/RT region, these tests are performed routinely in many countries for the design of individualized antiretroviral treatments as well as for estimating the frequency of resistance mutations in a population [21]. These large datasets can be analyzed with different evolutionary tools in order to depict the local epidemic, inferring the subtype distribution and detecting transmission clusters.

One important requirement to consider in the analysis of viral sequences is that they should have enough genetic variability (phylogenetic signal) to reconstruct the epidemiological relationships of the population under study. Although the analysis of longer sequences (such as complete viral genomes) will provide the best phylogenetic information (and potentially will better depict the epidemic) [22], their availability (coverage) is limited. Molecular epidemiology analyses usually rely on the PR/RT region, for which there is a large amount of sequences and which has been demonstrated to contain enough information to study HIV transmission dynamics [23]. However, one obvious limitation of using specific gene sections is that they do not allow the detection of recombinant sequences when recombination occurs outside those regions, or resistance mutations located outside PR/RT [24].

Phylogenetic analyses allow testing the genealogical relationships and evolutionary factors acting on viral sequences sampled in a population. Consequently, they are a common procedure in molecular epidemiology analyses of RNA viruses such as HIV [25]. There are many different phylogenetic methods, which can be grouped in two major categories: based on algorithmic procedures, that usually start from a distance matrix, such as neighbor-joining and UPGMA approaches, and based on an optimality criterion, also called statistical methods, such as maximum-likelihood (ML) or Bayesian phylogenetics [26]. Some of the most widely used programs for phylogenetic inference are PhyML [27], RAxML [28], IQTREE [29] (ML), MrBayes [30] and Beast [31,32] (Bayesian).

Statistical phylogenetics include the currently most commonly used methods in molecular epidemiology analyses of HIV, because they tend to resolve the evolutionary relationships with higher accuracy [33], and also allow testing hypotheses, such as evolutionary and demographic models, alternative topologies or the support for specific branches of the trees [34].

One of the main applications of molecular epidemiology analyses of HIV sequences is to establish the existence and to characterize transmission clusters (groups of individuals infected with similar viral sequences that derive from a common, recent ancestor) in the population. To assess the epidemiological relationships from HIV sequences sampled from a population, a representative set of reference sequences epidemiologically unrelated to the query dataset has to be included in the analysis. This allows avoiding the spurious clustering of sequences that, although sampled from the same population, correspond to different viral introductions [35]. This can be done by incorporating representative sequences from Los Alamos HIV Database (LANL; <http://www.hiv.lanl.gov>) or, alternatively, by adding sequences that display the highest similarity in BLAST searches to the query dataset. Consequently, the detection of transmission clusters usually involves a geographic cut-off for the maximum proportion of reference sequences clustering together with query sequences. Usually, this proportion is set to be lower than 20% for a given

cluster [36,37], to ensure that the transmission clusters identified represent transmissions that have mostly occurred in the studied population. An additional threshold on node support (such as bootstrap, approximate likelihood ratio test [aLRT] or posterior probability [PP] values) is commonly set in order to consider only highly supported clusters. These thresholds are usually set at higher than 0.90 [5,10,38–43].

Alternatively, transmission clusters can be detected as groups of viral sequences clustering at short genetic distances (usually, maximum pairwise distances lower than 0.045 substitutions per site [40], but there are many alternatives in the literature). This approach would save computational time, and could also be effective in avoiding spurious clustering caused by a low coverage of reference sequences in phylogenetic analyses [44]. In practice, both approaches are frequently used together, with transmission clusters detected based on criteria that combine well-supported groups of HIV sequences in a phylogenetic tree with a maximum genetic distance. Since there is not a single threshold that would fit all data sets, one recommendation is to perform sensitivity analyses at different genetic distance cutoffs [42,45,46].

The introduction of a viral lineage in a host population can be dated by assuming the molecular clock hypothesis, which is usually used in combination with coalescent theory of population genetics [47,48] to viral sequences. The most popular program for estimating dates of divergence is BEAST [31,32]. By using a Bayesian MCMC coalescent method with asynchronous samples or using an external estimate of the diversification date of a lineage or the molecular clock rate, BEAST can infer dates on phylogenies. There are two main models of molecular clock. Under the strict molecular clock model, the number of differences between sequences is proportional to their divergence times and the same clock rate applies to all the lineages in a phylogeny. With the relaxed molecular clock model, the assumption is still valid but different lineages may evolve with different rates. Usually, a previous, exploratory test of temporal signal has to be performed in order to assess the validity of the assumption of molecular clock in our data set. This test can be performed with TempEst [49] which, considering a phylogenetic tree (which should always be obtained without the molecular clock assumption) and the sampling date of each sequence, performs a correlation analysis between genetic divergence and sampling times.

Since rates and dates estimate can be affected by demographic changes and deviations from the molecular clock assumption, BEAST implements different models that can be tested to select the most appropriate for the data in the analysis. BEAST also allows setting different prior distributions for the parameters to be estimated, such as the time for the most recent common ancestor (tMRCA) or the evolutionary rate. Given that specifying inappropriate priors can bias the results, informative prior distributions should only be used if prior information on the corresponding model parameters is known (e.g., tMRCA or substitution rate of a given HIV subtype [50–53]). In addition, the impact of the used priors in our analyses should always be assessed by running the Bayesian phylogenetic inference programs sampling from prior only (this means, with an empty alignment).

The purpose of this review is to be a general introduction to the use of molecular epidemiology approaches for HIV. We also recommend, for further reading, other comprehensive works covering the methodology that is used to study HIV dynamics [54–57].

Conclusion

As concluding remarks, the use of molecular and evolutionary approaches has a clear importance for epidemiologists working with infectious diseases. Here, we have reviewed the main applications of molecular epidemiology tools to study the HIV epidemics. We have also addressed the most common approaches used for a correct study design. These analyses allow the detection of resistant and/or pathogenic variants that can hamper the control of the epidemic. In addition, they facilitate the detection of HIV transmission clusters, which potentially represent local outbreaks and whose occurrence is often unnoticed by using only nonmolecular epidemiology analyses. This is relevant, for instance, for enhancing detection and prevention campaigns focusing in the most vulnerable populations to HIV infection.

Future perspective

Next-generation sequencing technologies have a very high-throughput, generating enormous amounts of sequences at a very fast pace and a reduced cost [58]. Hence, they have the potential to generate datasets from an epidemic with excellent geographical and temporal coverage, overcoming the limitations of sampling bias. They also allow more realistic analyses of the viral quasispecies. This is clinically relevant for the detection of rare resistant variants [59]. Computational methods to study viral transmission clusters that use next-generation sequencing-derived sequences are currently being developed and improved, and can account for the presence of mixed infections (patients

infected from two or more different sources) and/or nonsampled patients [60]. In this way, the improvements in sequencing technologies and computational tools to study the generated data are certainly promising to bring a better understanding of current and future viral epidemics.

Executive summary

- The HIV pandemic affects millions of people worldwide, and there is an increased trend of infections among vulnerable groups, such as men who have sex with men.
- The combination of epidemiological, molecular and evolutionary tools is relevant to depict HIV epidemics. In this way, molecular epidemiology can help to develop detection and prevention campaigns focusing in the most vulnerable populations to HIV infection.
- This review aims to highlight the most relevant uses of molecular epidemiology, giving an overview of the most common approaches used for understanding HIV epidemics. Consequently, it may be useful for designing correct protocols in HIV epidemiological studies.

Financial & competing interests disclosure

This work was supported by projects BFU2014-58656R and BFU2017-89594-R from MINECO (Spanish Government) and PROMETEO2016/122 from Generalitat Valenciana. JAPG is supported by the National Institute of General Medical Sciences (NIGMS) of the NIH under award number R01GM117591. The authors have no other 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 apart from those disclosed.

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 UNAIDSWHO. Fact sheet – world AIDS day 2017. (2017). http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf
- 2 ECDC. Men who have sex with men. (2013). Monitoring implementation of the Dublin Declaration on Partnership to Fight HIV/AIDS in Europe and Central Asia: 2012 progress report. <https://ecdc.europa.eu/en/infectious-diseases-public-health/hiv-infection-and-aids/prevention-and-control/monitoring>
- 3 Díez M, Bleda MJ, Varela JR *et al.* Trends in HIV testing, prevalence among first-time testers, and incidence in most-at-risk populations in Spain: the EPI-VIH Study, 2000 to 2009. *Euro Surveill.* 19(47), 20971 (2014).
- 4 Moya A, Holmes EC, González-Candelas F. The population genetics and evolutionary epidemiology of RNA viruses. *Nat. Rev. Microbiol.* 2(4), 279–288 (2004).
- 5 Patiño-Galindo JA, Torres-Puente M, Bracho MA *et al.* The molecular epidemiology of HIV-1 in the Comunidad Valenciana (Spain): analysis of transmission clusters. *Sci. Rep.* 7(1), 11584 (2017).
- 6 Fabeni L, Alteri C, Orchi N *et al.* Recent transmission clustering of HIV-1 C and CRF17_BF strains characterized by NNRTI-related mutations among newly diagnosed men in central Italy. *PLoS ONE* 10(8), e0135325 (2015).
- 7 Peters PJ, Pontones P, Hoover KW *et al.* HIV infection linked to injection use of oxycodone in Indiana, 2014–2015. *N. Eng. J. Med.* 375(3), 229–239 (2016).
- 8 Pineda-Peña AC, Faria NR, Imbrechts S *et al.* Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect. Genet. Evol.* 19, 337–348 (2013).
- 9 Viciano I, González-Domenech CM, Palacios R *et al.* Clinical, virological and phylogenetic characterization of a multidrug-resistant HIV-1 strain outbreak in naive patients in southern Spain. *J. Antimicrob. Chemother.* 71(2), 357–361 (2016).
- 10 Fabeni L, Berno G, Fokam J *et al.* Comparative evaluation of subtyping tools for surveillance of newly emerging HIV-1 strains. *J. Clin. Microbiol.* 55(9), 2827–2837 (2017).
- **An important work that can be useful when designing subtyping protocols.**
- 11 Peeters M, Aghokeng AF, Delaporte E. Genetic diversity among human immunodeficiency virus-1 non-B subtypes in viral load and drug resistance assays. *Clin. Microbiol. Infect.* 16(10), 1525–1531 (2010).
- 12 Kaleebu P, French N, Mahe C *et al.* Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. *J. Infect. Dis.* 185(9), 1244–1250 (2002).
- 13 Kouri V, Khouri R, Alemán Y *et al.* CRF19_cpx is an evolutionary fit HIV-1 variant strongly associated with rapid progression to AIDS in Cuba. *EBioMedicine* 2(3), 244–254 (2015).

- 14 Patiño Galindo JA, Torres-Puente M, Gimeno C *et al.* Expansion of the CRF19_cpx variant in Spain. *J. Clin. Virol.* 69(0), 146–149 (2015).
- 15 Nickle DC, Heath L, Jensen MA *et al.* HIV-specific probabilistic models of protein evolution. *PLoS ONE* 2(6), e503 (2007).
- 16 Hemelaar J. Implications of HIV diversity for the HIV-1 pandemic. *J. Infect.* 66(5), 391–400 (2013).
- 17 Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin. Infect. Dis.* 42(11), 1608–1618 (2006).
- 18 Rhee SY, Sankaran K, Varghese V *et al.* HIV-1 protease, reverse transcriptase, and integrase variation. *J. Virol.* 90(13), 6058–6070 (2016).
- 19 Beerenwinkel N, Daumer M, Oette M *et al.* Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.* 31(13), 3850–3855 (2003).
- 20 Chan JM, Rabadan R. Quantifying pathogen surveillance using temporal genomic data. *mBio* 4(1), e00524–12 (2013).
- 21 WHO. The HIV drug resistance report – 2017. <http://www.who.int/hiv/pub/drugresistance/hivdr-report-2017/en/> (2017).
- 22 Charleston MA, Hendy MD, Penny D. The effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. *J. Comput. Biol.* 1(2), 133–151 (1994).
- 23 Hué S, Clewley JP, Cane PA, Pillay D. HIV-1 *pol* gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18(5), 719–728 (2004).
- 24 Wensing AM, Calvez V, Günthard HF *et al.* 2017 Update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* 24(4), 132–141 (2017).
- 25 Ciccozzi M, Santoro MM, Giovanetti M *et al.* HIV-1 non-B subtypes in Italy: a growing trend. *Microbiologica* 35(4), 377 (2012).
- 26 Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* 4, 275–284 (2003).
- 27 Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52(5), 696–704 (2003).
- 28 Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9), 1312–1313 (2014).
- 29 Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30(5), 1188–1195 (2013).
- 30 Ronquist F, Teslenko M, van der Mark P *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61(3), 539–542 (2012).
- 31 Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214 (2007).
- 32 Bouckaert R, Heled J, Kühnert D *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10(4), e1003537 (2014).
- 33 Huelsenbeck JP, Hillis DM. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42, 247–264 (1993).
- 34 Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* 17(5), 262–272 (2001).
- 35 Caro-Pérez N, Martínez-Rebollar M, Gregori J *et al.* Phylogenetic analysis of an epidemic outbreak of acute hepatitis C in HIV-infected patients by ultra-deep pyrosequencing. *J. Clin. Virol.* 92, 42–47 (2017).
- 36 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).
- **A pioneer work on HIV transmission cluster detection.**
- 37 Kouyos RD, von Wyl V, Yerly S *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* 201(10), 1488–1497 (2010).
- 38 Lewis F, Hughes GJ, Rambaut A, Pozniak A *et al.* Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 5(3), e50 (2008).
- 39 Hughes GJ, Fearnhill E, Dunn D *et al.* Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog.* 5(9), e1000590 (2009).
- 40 Leigh Brown AJ, Lycett SJ, Weinert L *et al.* Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* 204(9), 1463–1469 (2011).
- 41 Yebra G, Delgado R, Pulido F *et al.* Different trends of transmitted HIV-1 drug resistance in Madrid, Spain, among risk groups in the last decade. *Arch. Virol.* 159(5), 1079–1087 (2014).
- 42 Bezemer D, Cori A, Ratmann O *et al.* Dispersion of the HIV-1 epidemic in men who have sex with men in The Netherlands: a combined mathematical model and phylogenetic analysis. *PLoS Med.* 12(11), e1001898 (2015).
- **A relevant work on molecular epidemiology of HIV, both regarding the methodology that was used, and the results that were obtained.**

- 43 Patiño-Galindo JA, Torres-Puente M, Bracho MA *et al.* Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain. *PLoS ONE* 12(2), e0171062 (2017).
- 44 Wertheim JO, Leigh Brown AJ, Hepler NL *et al.* The global transmission network of HIV-1. *J. Infect. Dis.* 209, 304–313 (2014).
- 45 Prosperi M, Prosperi L, Bruselles A *et al.* Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics* 12(1), 5 (2011).
- 46 Ragonnet-Cronin M, Hodcroft E, Hue S *et al.* Automated analysis of phylogenetic clusters. *BMC Bioinformatics* 14(1), 317 (2013).
- 47 Kingman JFC. The coalescent. *Stoch. Proc. Appl.* 13, 235–248 (1982).
- 48 Donnelly P, Tavaré S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29, 401–421 (1995).
- 49 Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2(1), vew007 (2016).
- 50 Abecasis AB, Vandamme AM, Lemey P. Quantifying differences in the tempo of Human immunodeficiency virus type 1 subtype evolution. *J. Virol.* 83(24), 12917–12924 (2009).
- 51 Aldrich C, Hemelaar J. Global HIV-1 diversity surveillance. *Trends Mol. Med.* 18(12), 691–694 (2012).
- 52 Worobey M, Watts TD, McKay RA *et al.* 1970s and ‘Patient 0’ HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 539, 98–101 (2016).
- 53 Patiño-Galindo JA, González-Candelas F. The substitution rate of HIV-1 subtypes: a genomic approach. *Virus Evol.* 3(2), vex029 (2017).
- 54 Dennis AM, Herbeck JT, Brown AL *et al.* Phylogenetic studies of transmission dynamics in generalized HIV epidemics: an essential tool where the burden is greatest? *J. Acquir. Immune Defic. Syndr.* 67(2), 181–195 (2014).
- 55 Grabowski MK, Redd AD. Molecular tools for studying HIV transmission in sexual networks. *Curr. Op. HIV AIDS* 9(2), 126–133 (2014).
- 56 Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjornsson J. Defining HIV-1 transmission clusters based on sequence data: a systematic review and perspectives. *AIDS* 31(9), 1211–1222 (2017).
- **A comprehensive review, comparing the different methodologies that are followed for transmission cluster detection.**
- 57 Novitsky V, Moyo S, Essex M. Phylogenetic inference of HIV transmission clusters. *Infect. Dis. Transl. Med.* 3(2), 51–59 (2017).
- 58 Cruz-Rivera M, Forbi JC, Yamasaki LHT *et al.* Molecular epidemiology of viral diseases in the era of next generation sequencing. *J. Clin. Virol.* 57(4), 378–380 (2013).
- 59 Kyeyune F, Gibson RM, Nankya I *et al.* Low-frequency drug resistance in HIV-infected Ugandans on antiretroviral treatment is associated with regimen failure. *Antimicrob. Agents Chemother.* 60(6), 3380–3397 (2016).
- 60 De Maio N, Wu CH, Wilson DJ. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.* 12(9), e1005130 (2016).