Towards a new paradigm linking virus molecular evolution and pathogenesis: experimental design and phylodynamic inference

Melissa M. Norström[1], Annika C. Karlsson[1], Marco Salemi[2]

[1]*Department of Laboratory Medicine, Division of Clinical Microbiology, Karolinska University Hospital, Karolinska Institutet, Stockholm, Sweden;*
[2]*University of Florida, College of Medicine, Department of Pathology, Immunology and Laboratory Medicine & Emerging Pathogens Institute, Gainesville, Florida, USA*

SUMMARY

Phylogenetic analysis has become a powerful tool for the investigation of evolution at a molecular level. During the last three decades, statistical phylogenetics has increasingly been applied to the study of microbial pathogens. The new field of *phylodynamics* was formally introduced in 2004 and encompasses the interaction between evolutionary and ecological processes that shape the spatiotemporal and phylogenetic patterns of infectious disease dynamics. This novel framework has significantly enhanced the study of measurable evolving pathogen populations, in particular RNA viruses and retroviruses. One of the major challenges in phylodynamic studies, however, is the generation of data in the form of dense coverage in sequence sampling coupled with high quality epidemiological and/or accurate clinical information. This review focuses specifically on experimental and data assembling strategies that are required to test multi-level phylodynamic hypotheses, ranging from intra-host viral evolution to population dynamics of infectious disease pandemics. Ultimately, bridging the gap between rational experimental design and phylodynamic inference will prove to be essential to take full advantage of this new exciting area of research.

KEY WORDS: Phylodynamics, Viral evolution, Experimental design, Phylogenetic analysis, Molecular epidemiology, Intra-host evolution, Molecular clock.

Received April 4, 2012 *Accepted* April 4, 2012

INTRODUCTION

Since the development of the first algorithm to compute a phylogenetic tree from amino or nucleic acid sequences (Fitch and Margoliash, 1967), phylogenetic analysis has become a powerful tool for the investigation of evolution at a molecular level (Lemey *et al.*, 2009a). During the last three decades, statistical phylogenetics has increasingly been applied to the study of microbial pathogens, an approach based on two main theoretical frameworks: coalescent theory and the molecular clock hypothesis. Coalescent theory has provided a solid basis to study how ancestral relationships of individuals sampled from a population, which can be inferred from a gene genealogy (phylogenetic tree), are influenced by the population demographic history (Kingman, 1982; Pybus *et al.*, 2000). The molecular clock hypothesis, which assumes that the accumulation of genetic diversity in a given gene occurs at approximately constant rate along different phylogenetic lineages (Zuckerkandl and Pauling, 1962, 1965; Kimura, 1969, 1980), has provided the ability to calibrate the evolutionary time scale of a genealogy. Clock-like genealogies can be used to estimate the time of the most recent common ancestor (TMRCA) of a group of nucleotide or amino acid sequences (Lemey and Posada, 2009).

Corresponding author
Marco Salemi, Ph.D.
University of Florida College of Medicine
Department of Pathology, Immunology
and Laboratory Medicine
Emerging Pathogens Institute
P.O. Box 103633
Gainesville, FL, 32610-3633, USA
E-mail: salemi@pathology.ufl.edu

The new field of *phylodynamics* was formally introduced by Grenfell and colleagues in a seminal paper published in 2004 as "the melding of immunodynamics, epidemiology, and evolutionary biology" required to correlate the epidemiology and evolutionary behavior of pathogens with the immune system of the host (Grenfell *et al.*, 2004). Phylodynamics encompass both phylogeny inference and the interaction between evolutionary (i.e. mutation, genetic drift, selection) and ecological (population dynamics and environmental stochasticity) processes, which shape the spatiotemporal and phylogenetic patterns of infectious disease dynamics, both at the intra- and inter-host level (Raffel *et al.*, 2008; Kuhnert *et al.*, 2011).

This new analytical approach has enhanced the study of the evolution of fast evolving viruses, especially the human immunodeficiency virus (HIV) (Salemi *et al.*, 2005a, 2005b, 2007; Lewis *et al.*, 2008; Gray *et al.*, 2009; Herbeck *et al.*, 2011), hepatitis C virus (HCV) (Magiorkinis *et al.*, 2009; Bull *et al.*, 2011; Ciccozzi *et al.*, 2011; Gray *et al.*, 2011a; Ramachandran *et al.*, 2011), and influenza (Koelle *et al.*, 2006; Lemey *et al.*, 2009b; Lin *et al.*, 2011; Murcia *et al.*, 2011). Phylodynamic inference has become a rapidly expanding field, resulting in 492 publications in peer-reviewed scientific journals between 2004 and 2011 (Figure 1). In the first quarter of 2012, 36 papers have already been published, more than the yearly number of papers printed during 2005-2007. Moreover, a closer look at the literature shows that phylodynamic analysis has been applied, over the years, to investigate an increasingly diverse number of viral (Salemi *et al.*, 2004; Kerr *et al.*, 2009; Bennett *et al.*, 2010; Gray *et al.*, 2010; Hoelzer *et al.*, 2010; Castro-Nallar *et al.*, 2011; Zehender *et al.*, 2011; Patricio *et al.*, 2012) as well as bacterial pathogens (Harris *et al.*, 2010; Nubel *et al.*, 2010; Gray *et al.*, 2011d; Mutreja *et al.*, 2011).

A comprehensive synopsis of current phylodynamic literature is beyond the scope of this review and can be found elsewhere (Holmes and Grenfell, 2009; Kuhnert *et al.*, 2011). Herein, the emphasis will be on discussing specific experimental and data assembling strategies required to test multi-level phylodynamic hypotheses, ranging from intra-host viral evolution to population dynamics of infectious disease pandemics.

The concept of measurably evolving populations

Phylodynamic analysis employs coalescence theory to investigate how the genealogy of a pathogen population, both at the intra- and inter-host level, is influenced by the interaction among pathogen's demographic history, host immunological milieu and environmental/ecological factors. Pathogen populations for which such studies are possible, however, must be measurably evolving populations (MEPs). MEPs are characterized by sufficiently long, or numerous, sampled sequences and fast evolutionary rate relative to the available range of sequence sampling times, such that a statistically significant number of mutations can be detected between sequences obtained at different time points (Drummond *et al.*, 2003). The evolutionary rate of living organisms, i.e. the rate at which mutations

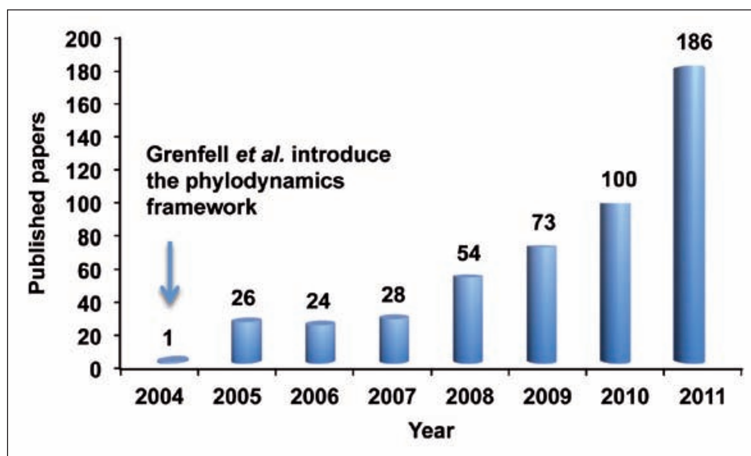

FIGURE 1 - *Publications on phylodynamics of viral pathogens. Grenfell et al. introduced the term "phylodynamics" in 2004; since then, the field has been rapidly growing. PubMed search included keywords such as phylodynamic(s), phylogeography, molecular epidemiology, intra-host evolution, within-patient evolution, viral pathogen(s).*

are fixed during evolution, can vary from 10^{-9} nucleotide substitutions/site/year in most cellular genes of eukaryotes, prokaryotes and *archaea*, to 10^{-5}-10^{-2} nucleotide substitutions/site/year in fast-evolving single strand DNA viruses, RNA viruses and retroviruses (Figure 2). Considering that the RNA genome of viruses such as HIV or HCV contains around 10^4 base pairs, a viral population replicating and spreading among susceptible hosts, or even within a single host, can easily accumulate between 0.1 and 10% genetic diversity in one year. Fast-evolving viruses are MEPs, since sequences longitudinally sampled at intervals of a few months or a few years usually contain significant phylogenetic and molecular clock signal. Phylodynamic analysis of MEPs has been fostered, in particular, by the development of phylogeny inference methods that can estimate trees and divergence dates even in the face of uncertainty in evolutionary rates and calibration times (Drummond *et al.*, 2006). The Bayesian coalescent framework implemented in the BEAST program, which is included in one of the most used software packages in phylodynamic inference (http://evolve.zoo.ox.ac.uk/beast/), was specifically developed to allow the estimation of evolutionary trees from data set of sequences sampled at different time points (Drummond *et al.*, 2005, 2006). By using a statistical model, known as the relaxed molecular clock, the method is able to infer simultaneously the phylogeny and the evolutionary time-scale leading to the sampled sequences. In addition, BEAST implements a Markov Chain Monte Carlo (MCMC) algorithm that samples proposed phylogenies (according to prior distributions) by comparing the ratio of their posterior probability (Drummond and Rambaut, 2009). Therefore, while classic clustering algorithms (e.g. Neighbor-joining or UPGMA) produce one single tree with branch lengths scaled in substitutions per site, the MCMC generates a collection of trees, the posterior tree distribution, with branch lengths proportional to a specific time scale (e.g. years, months, days). It is then possible to compare different tree topologies within the distribution and statistically evaluate the one that has the highest probability to represent the true evolutionary history of the sequences analyzed.

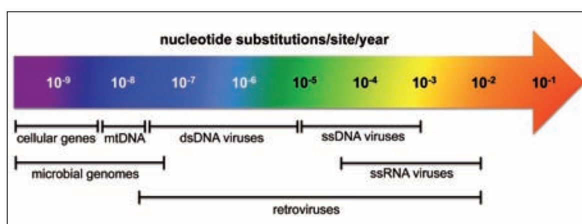

FIGURE 2 - *Evolutionary rates of different organisms. Rates characteristic of measurably evolving populations are indicated in black.*

Multi-scale phylodynamics

Phylodynamics can be applied at multiple levels, from the micro-level of intra-host evolution to the macro-level of pandemic spread (Figure 3). At the bottom level, several studies have investigated, for example, HCV evolutionary patterns in chronically infected patients (Gray *et al.*, 2011a, 2012), as well as the dynamics of HIV-1 sub-populations infecting different tissues such as breast milk (Gray *et al.*, 2011c), brain (Salemi *et al.*, 2005b; Lamers *et al.*, 2011), lymphoid tissues (Salemi *et al.*, 2007), and metastatic tissues from patients with AIDS-related lymphomas (Salemi *et al.*, 2009). The next level of phylodynamic inference has been employed to investigate population dynamics of pathogens spreading within commu-

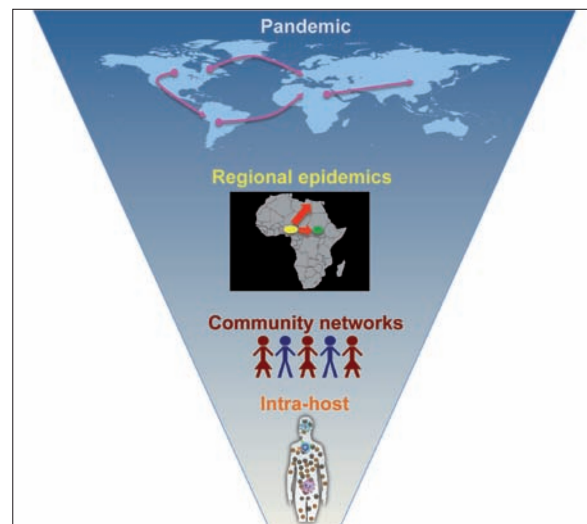

FIGURE 3 - *Multi-scale phylodynamics. The figure illustrates (from bottom to top) different levels of phylodynamic inference that can be employed to analyze the evolution and population dynamics of pathogens within an infected host, a community network, a geographic region or worldwide.*

nity networks. One of the best examples is a “célèbre” paper (de Oliveira *et al.*, 2006) describing the HIV/HCV outbreak in a Libyan hospital in 1998, whose origin was initially attributed to the malicious intervention of foreign medical staff, and eventually proven to be antecedent to the time in which the staff were hired. At a further level, phylodynamics has been used to track epidemic spread within a defined geographic area by inferring patterns of gene flow (migration), demographic history and TMRCA of the viral population (Hue *et al.*, 2005; Salemi *et al.*, 2005a; Gilbert *et al.*, 2007; Salemi *et al.*, 2008). Such an approach, coupled with the analysis of geographic information system (GIS) data, has been used, for example, to explain HIV-1 epidemic emergence in Uganda and Kenya during the 1970s (Gray *et al.*, 2009), or the dissemination of dog rabies virus (RABV) in North Africa (Talbi *et al.*, 2010). Finally, the highest level of phylodynamic inference is concerned with the population dynamics of pathogens spreading worldwide. A representative example was the uncovering of the origin and evolutionary genomics of the 2009 H1N1 influenza A pandemic strain (Smith *et al.*, 2009). Overall, these studies illustrate how phylodynamic analysis of pathogens has been providing a more holistic and evidence-based interpretation of intra-host viral evolution, as well as epidemic spread.

Experimental design and phylodynamic inference

The link between experimental design and phylodynamic inference can ideally be construed as a hypothetical-deductive cycle (Whewell, 1837,

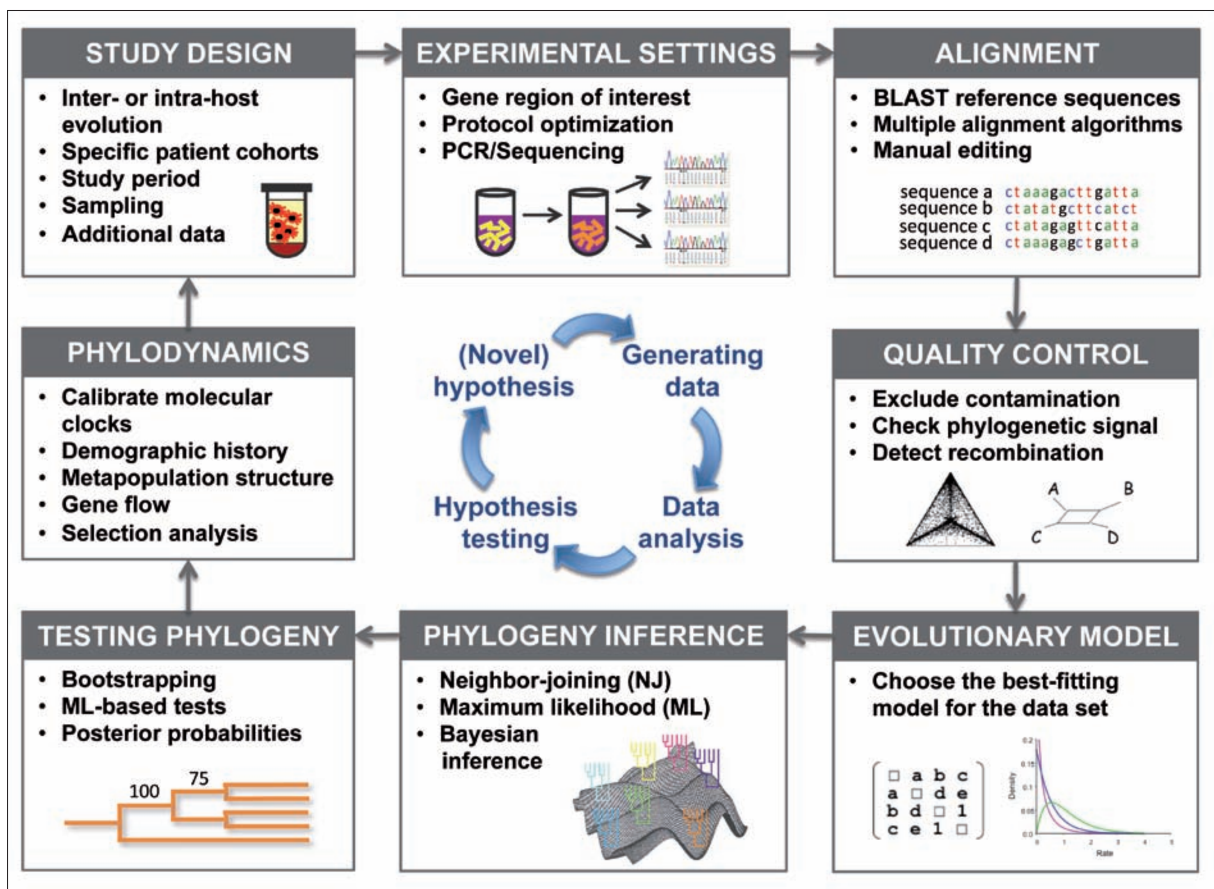

FIGURE 4 - *Flow-chart representing the major steps in phylodynamic inference linking experimental design and data analysis. The inner circle represents the ideal feedback loop in which testing a (novel) scientific hypothesis, through data generation and phylodynamic analysis, eventually results in a new hypothesis and another flow of experimental data generation/analysis.*

1840) consisting of the formulation of a specific evolutionary hypothesis, generation of the data, analysis of the data and subsequent test of the hypothesis, which can lead in turn to the generation of a revised or novel hypothesis and a new cycle (Figure 4). Study design is the first step in the whole process. Depending on the phylodynamic scale being investigated (Figure 3), sequence data sets may include viral strains sampled either from the same host (intra-host phylodynamics) or from different hosts (inter-host phylodynamics). Intra-host studies usually focus on comparisons between phylodynamic patterns of a pathogen infecting a specific cohort and a control group, e.g. slow versus fast disease progression (Lemey *et al.*, 2007; Lee *et al.*, 2008), therapy naïve versus treated subjects (Buzon *et al.*, 2011; Ceccarelli *et al.*, 2012), or mono-infected versus co-infected subjects. Inter-host studies may be focusing on viral data sets including sequences sampled from different hosts in a specific geographic area or from multiple locations, and/or belonging to a specific risk group (Prosperi *et al.*, 2011). Depending on the scientific question, the availability of additional data could also be of value. Examples include HLA-typing of the infected subjects (Karlsson *et al.*, 2007), clinical and immunological information (CD4 counts, viral load, T cell responses, etc.), *in vitro* measurements of viral fitness (replication capacity, entry efficiency), accurate sampling location records (e.g. GPS), and/or GIS data. At this stage, however, one of the most crucial points to consider is how to optimize the longitudinal sampling strategy, i.e. the number of sampled time points, the sampling interval and the number of sequences per time point. Unfortunately, most phylodynamic studies to date have been based on convenience samples or data mining of sequence databases, which are often sparse and/or biased. This approach, albeit unavoidable in some cases, severely limits the power of phylodynamic inference and may result in contradictory findings difficult to disentangle. It has been proven that as few as eight randomly sampled sequences from a panmictic population (i.e. without significant sub-population structure) are sufficient to obtain accurate coalescent estimates (Felsenstein, 2006). Several studies have also shown that consistent inferences on the origin and demographic history of viral epidemics can be obtained by analyzing 30-40 randomly sampled sequences of epidemiologically unlinked individuals (Pybus *et al.*, 2000; Strimmer and Pybus, 2001; Drummond *et al.*, 2005). The calibration of a reliable molecular clock, on the other hand, may require samples collected, at least, at three to five different time points with a sampling interval depending on the specific virus and phylodynamic scale under investigation (Figure 2 and 3). A robust study on HIV intra-host phylodynamics may ideally require 5-10 time points with 20-25 viral sequences per time point, sampled at intervals of 6-22 months, in order to ensure sufficient population turnover and to detect significant temporal structure in the phylogeny (Achaz *et al.*, 2004; Gray *et al.*, 2011b). Finally, the genomic region and the size of the sequenced fragment are also important factors. When possible, longer or multiple regions should be prioritized over the number of clones (Felsenstein, 2006). The experimental protocol could consist of direct, clonal, single genome or ultra-deep sequencing, depending on the specific cohort under investigation. Intra-host phylodynamic studies usually require only one viral sequence from each infected subject, which could easily be obtained by direct sequencing from a blood or plasma sample, while cloning or ultra-deep sequencing may be necessary for intra-host studies concerned with the evolutionary dynamic of minor variants (Henn *et al.*, 2012). For samples with low viral load, single genome sequencing may be used instead of cloning to avoid resampling, as well as reducing the likelihood of PCR mediated recombination (Shriner *et al.*, 2004; Palmer *et al.*, 2005; Lindkvist *et al.*, 2009).

After the generation of a new data set, sequences are aligned and a BLAST search is usually performed to find reference strains (Figure 4). Several algorithms can be employed to obtain multiple sequence alignments, although manual editing is often necessary in regions displaying an elevated proportion of insertion/deletions (indels). In fact, indels are commonly removed from the final alignment, or replaced with a "?", indicating missing data, since most tree-building algorithms are based on evolutionary models that do not take indels into account (Strimmer and von Haeseler, 2009).

The next step of an idealized phylodynamic pipeline must consist of quality control in order

to exclude contamination, investigate the presence of an adequate signal for reliable phylogenetic reconstruction and detect the presence of potential recombinant strains. PCR contamination can easily be excluded when dealing with measurably evolving viral populations from subjects not directly linked by a transmission chain. A simple neighbor-joining tree including sampled, reference and positive control viral strains would allow testing whether sequences from different subjects intermix in the same monophyletic clade, which is a clear indication that contamination has occurred. On the other hand, certain degrees of intermix between samples taken at different time points and/or tissues from the same host is usually expected. Therefore, excluding contamination in intra-host data sets by phylogenetic analysis may be difficult if not impossible. In such cases, improving the experimental procedure (e.g. performing RNA/DNA extractions separately, including appropriate positive and negative controls) is the only option to minimize the risk of contamination.

Several methods have been developed to measure the phylogenetic signal in the data set. Likelihood mapping (Strimmer and von Haeseler, 1997), transition/transversions versus divergence plots (Salemi, 2009), and the Xia test for saturation (Strimmer and von Haeseler, 1997; Xia *et al.*, 2003) are usually employed to assess the reliability of the aligned sequences for phylogeny inference (Salemi and Vandamme, 2002; Gray *et al.*, 2010). If the signal is too low, longer sequences or more variable genes should be considered. The presence of substitution saturation, on the other hand, indicates that sequences have accumulated so many mutations per site that the phylogenetic signal has become randomized (Xia *et al.*, 2003). In this case, amino acid or transversional substitutions, which saturates more slowly during evolution, might be used for phylogenetic reconstruction instead of the full sequence alignment.

Finally, it is important to keep in mind that recombination violates the basic assumption of phylogeny inference (ancestry from a common ancestor). Using algorithms that do not explicitly model recombination (e.g. BEAST) can bias molecular clock and coalescent estimates (Posada, 2001; Posada and Crandall, 2002; Posada *et al.*, 2002). Recombinant strains should be excluded and analyzed separately or with more complex coalescent models (Carvajal-Rodriguez *et al.*, 2006; Kuhner, 2006). A thorough discussion of such topics is beyond the scope of the present review but can be found elsewhere (Posada and Crandall, 2001; Rambaut *et al.*, 2004; Salminen and Martin, 2009; Martin *et al.*, 2011).

Study design, experimental settings, alignment and quality control are essential preconditions for the generation of "clean" data sets. As any other high-end computational method, phylodynamic inference is sensitive to the well-known GIGO (garbage in garbage out) paradigm: bad data will produce unreliable or even nonsensical output (Butler *et al.*, 2010). The next steps of the phylodynamic pipeline represent the "core" of data analysis and hypothesis testing (Figure 4), which usually include: i) testing for the best fitting nucleotide or amino acid substitution model, ii) inferring phylogenies with different algorithms (e.g., distance, maximum likelihood, Bayesian methods), iii) testing the trees reliability (e.g. by bootstrapping), iv) molecular clock, selection and coalescent-based inferences. Several freeware packages exist for these analyses and novel algorithms are continuously being developed. An in-depth survey of currently available phylogeny programs can be found at the website http://evolution.genetics.washington.edu/phylip/software.html, maintained by Joe Felsenstein. Unfortunately, while some of these programs are fairly user-friendly, others can be hard to implement and often lack proper documentation. Contacting the authors, reading the original literature or the main textbooks reviewing the theory and practice of phylogenetic analysis may be necessary to navigate this complex field (Page and Holmes, 1998; Felsenstein, 2004; Gascuel, 2005; Lemey *et al.*, 2009a; Hall, 2011).

CONCLUDING REMARKS

Phylodynamic inference is based on the main observation that viral genealogies display different shapes under different phylodynamic processes (Grenfell *et al.*, 2004). For example, genealogies of viral strains infecting a host exerting continual immune-driven selection, as in intra-host HIV-1, will show strong temporal structure and sequential population bottlenecks (Figure 5A). On the

other hand, viruses where immune selection is absent or weak, as in many RNA viruses or HIV epidemic spread, will be characterized by genealogies depicting population size and spatial dynamics (Figure 5B). Therefore, the analysis of temporally scaled intra- and/or inter-host genealogies is not only the final step of phylodynamic inference, through which specific hypotheses are tested, but also the opportunity to generate novel hypotheses in the hypothetical-deductive cycle (Figure 4).

In order to enrich and develop the field of phylodynamics it is essential to have good comparative data. Rational experimental design and appropriate sampling strategy, both for intra-host and molecular epidemiology studies, is extreme-

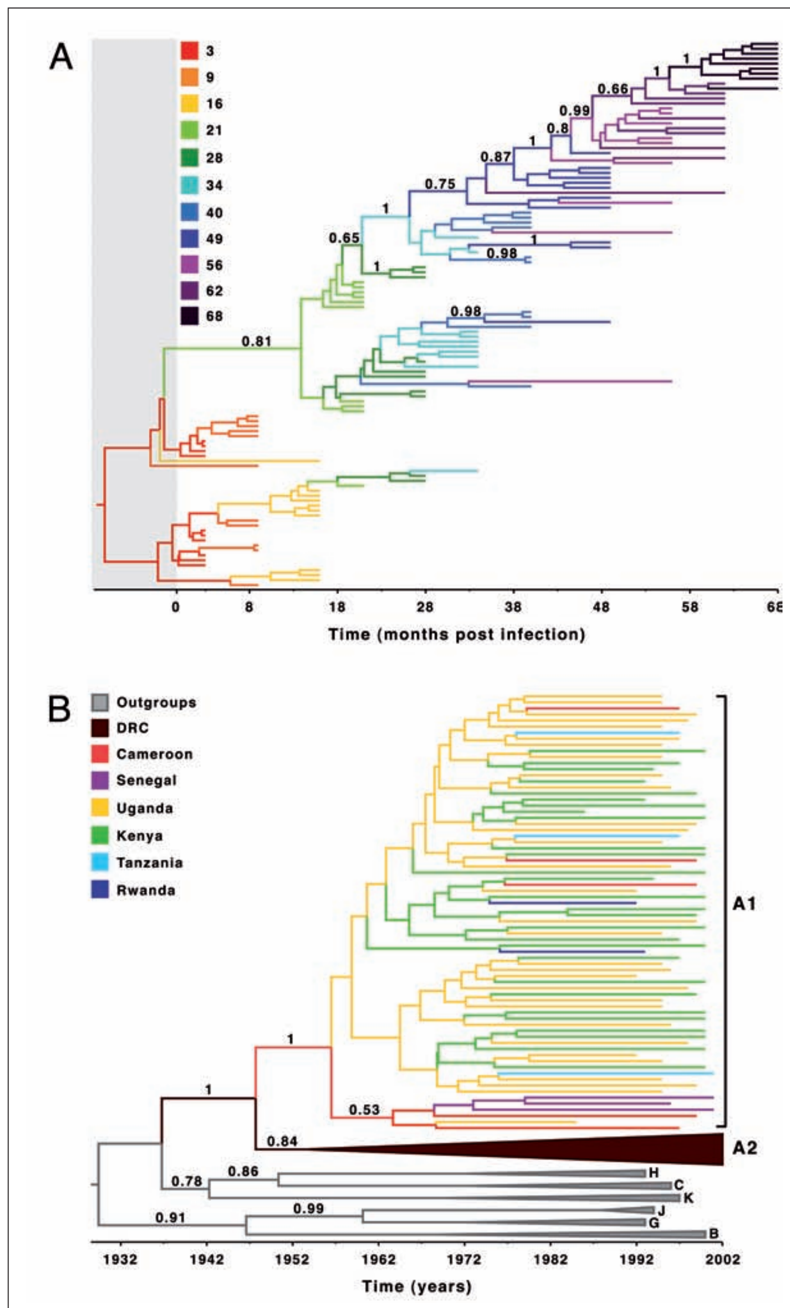

FIGURE 5 - *Example of phylogeny patterns in HIV-1 intra- and inter-host evolution. The figures show Bayesian maximum clade credibility trees inferred by enforcing a relaxed molecular clock and using the HKY+G nucleotide substitution model. Branch lengths are proportional to time according to the scale at the bottom of each tree; tips are colored according to the legend on the top left. Posterior probabilities for internal nodes are given above the branch. Calculations were performed with the BEAST package (http://beast.bio.ed.ac.uk). (A) Intra-host evolution data set using a subset of envelope gp120 C2-V5 (564 base pairs) sequences from patient 5 described by Shankarappa et al. (1999), including 110 sequences sampled at 11 different time points. The gray area corresponds to the pre-transmission interval (Leitner and Albert, 1999). The staircase topology is typical of HIV-1 intra-host evolution characterized by continual immune-driven selection (Grenfell et al., 2004). (B) Inter-host data set with 78 envelope HIV-1A gp41 sequences plus 12 reference sequences of different subtypes downloaded for the HIV Databases (http://www.hiv.lanl.gov/content/index). Internal branches were colored according to maximum parsimony reconstruction of ancestral locations (Slatkin and Maddison, 1989). For display purposes, HIV-1 subtype A2 strains were collapsed. The tree shows a clear spatiotemporal structure with phylogenetic lineages progressively emerging from West to East Africa, which depicts the gene flow (dissemination) of HIV-1 from DRC to Uganda and Kenya during the 1960-1970s (Gray et al., 2009).*

ly important. Intra-host data sets consisting of longitudinally sampled viral sequences are often difficult to obtain for both practical and ethical reasons. Studies on HIV-1 sub-population dynamics in different lymphoid and non- lymphoid tissues, for example, are usually possible only using specimens sampled *post mortem*. Longitudinal sampling from peripheral blood mononuclear cells or plasma may be limited by enrollment protocols of study subjects or patients' compliance. One of the best-characterized HIV-1 intra-host data sets to date includes viral sequences collected from nine patients followed from seroconversion up to 11 years (Shankarappa *et al.*, 1999). This data set has proven to be extremely useful to test phylodynamic models and investigating the relationship between viral evolution and disease progression (Rambaut *et al.*, 2004; Lemey *et al.*, 2007; Lee *et al.*, 2008; Gray *et al.*, 2011b). Yet, it also has significant limitations, e.g. less than eight clones were obtained for some of the sampled time points, only a limited portion of the envelope gp120 gene was sequenced, the HLA-type of the host was not determined, and the subjects underwent different treatment protocols during the study period. This example points out that one of the major challenges for future studies in phylodynamics will be the generation of data in the form of dense coverage in sequence sampling coupled with high quality epidemiological and/or accurate clinical information. Ultimately, bridging the gap between rational experimental design and phylodynamic analysis will prove to be essential to take full advantage of this new exciting area of research.

REFERENCES

Achaz G., Palmer S., Kearney M., Maldarelli F., Mellors J.W., Coffin J.M., ET AL. (2004). A robust measure of HIV-1 population turnover within chronically infected individuals. *Mol. Biol. Evol.* **21**, 1902-1912.

Bennett S.N., Drummond A.J., Kapan D.D., Suchard M.A., Munoz-Jordan J.L., Pybus O.G., et al. (2010). Epidemic dynamics revealed in dengue evolution. *Mol. Biol. Evol.* **27**, 811-818.

Bull R.A., Luciani F., McElroy K., Gaudieri S., Pham S.T., Chopra A., et al. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog.* **7**, e1002243.

Butler J., Lidwell W., Holden K. (2010). Universal principles of design. Gloucester, MA, Rockport Publishers.

Buzon M.J., Codoner F.M., Frost S.D., Pou C., Puertas M.C., Massanella M., et al. (2011). Deep molecular characterization of HIV-1 dynamics under suppressive HAART. *PLoS Pathog.* **7**, e1002314.

Carvajal-Rodriguez A., Crandall K.A., Posada D. (2006). Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. *Mol. Biol. Evol.* **23**, 817-827.

Castro-Nallar E., Cortez-San Martin M., Mascayano C., Molina C., Crandall K.A. (2011). Molecular phylodynamics and protein modeling of infectious salmon anemia virus (ISAV). *BMC. Evol. Biol.* **11**, 349.

Ceccarelli L., Salpini R., Moudourou S., Cento V., Santoro M.M., Fokam J., et al. (2012). Characterization of drug resistance mutations in naive and ART-treated patients infected with HIV-1 in Yaounde, Cameroon. *J. Med. Virol.* **84**, 721-727.

Ciccozzi M., Ciccaglione A.R., Lo Presti A., Yalcinkaya T., Taskan Z.P., Equestre M., et al. (2011). Reconstruction of the evolutionary dynamics of the hepatitis C virus 1b epidemic in Turkey. *Infect. Genet. Evol.* **11**, 863-868.

de Oliveira T., Pybus O.G., Rambaut A., Salemi M., Cassol S., Ciccozzi M., et al. (2006). Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature.* **444**, 836-837.

Drummond A.J., Pybus O.G., Rambaut A., Forsberg R., Rodrigo A.G. (2003). Measurably evolving populations. *Trends. Ecol. Evol.* **18**, 481-488.

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

Drummond A.J., Ho S.Y., Phillips M.J., Rambaut A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88.

Drummond A.J., Rambaut A. (2009). Bayesian evolutionary analysis by sampling trees. *The phylogenetic handbook : a practical approach to phylogenetic analysis and hypothesis testing*. P. Lemey, Salemi M. and Vandamme A.-M. New York, Cambridge University Press: 564-574.

Felsenstein J. (2004). *Inferring phylogenies*. Sunderland, Mass., Sinauer Associates.

Felsenstein J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* **23**, 691-700.

Fitch W.M., Margoliash E. (1967). Construction of phylogenetic trees. *Science.* **155**, 279-284.

Gascuel O. (2005). *Mathematics of evolution and phylogeny*. New York, Oxford University Press.

Gilbert M.T., Rambaut A., Wlasiuk G., Spira T.J., Pitchenik A.E., Worobey M. (2007). The emergence of HIV/AIDS in the Americas and beyond. *Proc. Natl. Acad. Sci USA.* **104**, 18566-18570.

GRAY R.R., TATEM A.J., LAMERS S., HOU W., LAEYENDECKER O., SERWADDA D., ET AL. (2009). Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *Aids.* **23**, F9-F17.

GRAY R.R., VERAS N.M., SANTOS L.A., SALEMI M. (2010). Evolutionary characterization of the West Nile Virus complete genome. *Mol. Phylogenet. Evol.* **56**, 195-200.

GRAY R.R., PARKER J., LEMEY P., SALEMI M., KATZOURAKIS A., PYBUS O.G. (2011a). The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol. Biol.* **11**, 131.

GRAY R.R., PYBUS O.G., SALEMI M. (2011b). Measuring the temporal structure in serially-sampled phylogenies. *Methods Ecol. Evol.* **2**, 437-445.

GRAY R.R., SALEMI M., LOWE A., NAKAMURA K.J., DECKER W.D., SINKALA M., ET AL. (2011c). Multiple independent lineages of HIV-1 persist in breast milk and plasma. *Aids.* **25**, 143-152.

GRAY R.R., TATEM A.J., JOHNSON J.A., ALEKSEYENKO A.V., PYBUS O.G., SUCHARD M.A., ET AL. (2011d). Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a bayesian framework. *Mol. Biol. Evol.* **28**, 1593-1603.

GRAY R.R., SALEMI M., KLENERMAN P., PYBUS O.G. (2012). A new evolutionary model for Hepatitis C virus chronic infection. *PLoS Pathog.* In press.

GRENFELL B.T., PYBUS O.G., GOG J.R., WOOD J.L., DALY J.M., MUMFORD J.A., ET AL. 2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science.* **303**, 327-332.

HALL B.G. (2011). *Phylogenetic trees made easy : a how-to manual*. Sunderland, Mass., Sinauer Associates.

HARRIS S.R., FEIL E.J., HOLDEN M.T., QUAIL M.A., NICKERSON E.K., CHANTRATITA N., ET AL. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science.* **327**, 469-474.

HENN M.R., BOUTWELL C.L., CHARLEBOIS P., LENNON N.J., POWER K.A., MACALALAD A.R., ET AL. (2012). Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathog.* **8**, e1002529.

HERBECK J.T., ROLLAND M., LIU Y., MCLAUGHLIN S., MCNEVIN J., ZHAO H., ET AL. (2011). Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J. Virol.* **85**, 7523-7534.

HOELZER K., MURCIA P.R., BAILLIE G.J., WOOD J.L., METZGER S.M., OSTERRIEDER N., ET AL. (2010). Intrahost evolutionary dynamics of canine influenza virus in naive and partially immune dogs. *J. Virol.* **84**, 5329-5335.

HOLMES E.C., GRENFELL B.T. (2009). Discovering the phylodynamics of RNA viruses. *PLoS Comput. Biol.* **5**, e1000505.

HUE S., PILLAY D., CLEWLEY J.P., PYBUS O.G. (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl. Acad. Sci USA.* **102**, 4425-4429.

KARLSSON A.C., IVERSEN A.K., CHAPMAN J.M., DE OLIVIERA T., SPOTTS G., MCMICHAEL A.J., ET AL. (2007). Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. *PLoS One.* **2**, e225.

KERR P.J., KITCHEN A., HOLMES E.C. (2009). Origin and phylodynamics of rabbit hemorrhagic disease virus. *J. Virol.* **83**, 12129-12138.

KIMURA M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. USA.* **63**, 1181-1188.

KIMURA M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120.

KINGMAN J.F.C. (1982). The coalescent. *Stoch. Proc. Appl.* **13**, 235-248.

KOELLE K., COBEY S., GRENFELL B., PASCUAL M. (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science.* **314**, 1898-1903.

KUHNER M.K. (2006). LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics.* **22**, 768-770.

KUHNERT D., WU C.H., DRUMMOND A.J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infect. Genet. Evol.* **11**, 1825-1841.

LAMERS S.L., GRAY R.R., SALEMI M., HUYSENTRUYT L.C., MCGRATH M.S. (2011). HIV-1 phylogenetic analysis shows HIV-1 transits through the meninges to brain and peripheral tissues. *Infect. Genet. Evol.* **11**, 31-37.

LEE H.Y., PERELSON A.S., PARK S.C., LEITNER T. (2008). Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput. Biol.* **4**, e1000240.

LEITNER T., ALBERT J. (1999). The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl. Acad. Sci. USA.* **96**, 10752-10757.

LEMEY P., KOSAKOVSKY POND S.L., DRUMMOND A.J., PYBUS O.G., SHAPIRO B., BARROSO H., ET AL. (2007). Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol.* **3**, e29.

LEMEY P., POSADA D. (2009). Molecular clock analysis. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. P. Lemey, Salemi M. and Vandamme A.-M. New York, Cambridge University Press: 362-372.

LEMEY P., SALEMI M., VANDAMME A.-M. (2009a). The phylogenetic handbook : a practical approach to phylogenetic analysis and hypothesis testing. New York, Cambridge University Press.

LEMEY P., SUCHARD M., RAMBAUT A. (2009b).

Reconstructing the initial global spread of a human influenza pandemic: a Bayesian spatial-temporal model for the global spread of H1N1pdm. *PLoS Curr.* **1**, RRN1031.

Lewis F., Hughes G.J., Rambaut A., Pozniak A., Leigh Brown A.J. (2008). Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* **5**, e50.

Lin J.H., Chiu S.C., Cheng J.C., Chang H.W., Hsiao K.L., Lin Y.C., et al. (2011). Phylodynamics and molecular evolution of influenza A virus nucleoprotein genes in Taiwan between 1979 and 2009. *PLoS One.* **6**, e23454.

Lindkvist A., Eden A., Norstrom M.M., Gonzalez V.D., Nilsson S., Svennerholm B., et al. (2009). Reduction of the HIV-1 reservoir in resting CD4+ T-lymphocytes by high dosage intravenous immunoglobulin treatment: a proof-of-concept study. *AIDS Res Ther.* **6**, 15.

Magiorkinis G., Magiorkinis E., Paraskevis D., Ho S.Y., Shapiro B., Pybus O.G., et al. (2009). The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Med.* **6**, e1000198.

Martin D.P., Lemey P., Posada D. (2011). Analysing recombination in nucleotide sequences. *Mol. Ecol. Resour.* **11**, 943-955.

Murcia P.R., Wood J.L., Holmes E.C. (2011). Genome-scale evolution and phylodynamics of equine H3N8 influenza A virus. *J. Virol.* **85**, 5312-5322.

Mutreja A., Kim D.W., Thomson N.R., Connor T.R., Lee J.H., Kariuki S., et al. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature.* **477**, 462-465.

Nubel U., Dordel J., Kurt K., Strommenger B., Westh H., Shukla S.K., et al. (2010). A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant Staphylococcus aureus. *PLoS Pathog.* **6**, e1000855.

Page R.D.M., Holmes E.C. (1998). Molecular evolution: a phylogenetic approach. Oxford; Malden, MA, Blackwell Science.

Palmer S., Kearney M., Maldarelli F., Halvas E.K., Bixby C.J., Bazmi H., et al. (2005). Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J. Clin. Microbiol.* **43**, 406-413.

Patricio A.R., Herbst L.H., Duarte A., Velez-Zuazo X., Santos Loureiro N., Pereira N., et al. (2012). Global Phylogeography and Evolution of the Chelonid Fibropapilloma-associated Herpesvirus. *J. Gen. Virol*.

Posada D. (2001). Unveiling the molecular clock in the presence of recombination. *Mol. Biol. Evol.* **18**, 1976-1978.

Posada D., Crandall K.A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA.* **98**, 13757-13762.

Posada D., Crandall K.A. (2002). The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* **54**, 396-402.

Posada D., Crandall K.A., Holmes E.C. (2002). Recombination in evolutionary genomics. *Annu. Rev. Genet.* **36**, 75-97.

Prosperi M.C., Ciccozzi M., Fanti I., Saladini F., Pecorari M., Borghi V., et al. (2011). A novel methodology for large-scale phylogeny partition. *Nat. Commun.* **2**, 321.

Pybus O.G., Rambaut A., Harvey P.H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics.* **155**, 1429-1437.

Raffel T.R., Martin L.B., Rohr J.R. (2008). Parasites as predators: unifying natural enemy ecology. *Trends. Ecol. Evol.* **23**, 610-618.

Ramachandran S., Campo D.S., Dimitrova Z.E., Xia G.L., Purdy M.A., Khudyakov Y.E. (2011). Temporal variations in the hepatitis C virus intrahost population during chronic infection. *J. Virol.* **85**, 6369-6380.

Rambaut A., Posada D., Crandall K.A., Holmes E.C. (2004). The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**, 52-61.

Salemi M., Vandamme A.M. (2002). Hepatitis C virus evolutionary patterns studied through analysis of full-genome sequences. *J. Mol. Evol.* **54**, 62-70.

Salemi M., Fitch W.M., Ciccozzi M., Ruiz-Alvarez M.J., Rezza G., Lewis M.J. (2004). Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. *J. Virol.* **78**, 1602-1603.

Salemi M., de Oliveira T., Soares M.A., Pybus O., Dumans A.T., Vandamme A.M., et al. (2005a). Different epidemic potentials of the HIV-1B and C subtypes. *J. Mol. Evol.* **60**, 598-605.

Salemi M., Lamers S.L., Yu S., de Oliveira T., Fitch W.M., McGrath M.S. (2005b). Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol.* **79**, 11343-11352.

Salemi M., Burkhardt B.R., Gray R.R., Ghaffari G., Sleasman J.W., Goodenow M.M. (2007). 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**, e950.

Salemi M., Goodenow M.M., Montieri S., de Oliveira T., Santoro M.M., Beshkov D., et al. (2008). The HIV type 1 epidemic in Bulgaria involves multiple subtypes and is sustained by continuous viral inflow from West and East European countries. *AIDS Res. Hum. Retroviruses.* **24**, 771-779.

Salemi M. (2009). Genetic distances and nucleotide

substitution models - Practice. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. P. Lemey, Salemi M. and Vandamme A.-M. New York, Cambridge University Press: 126-141.

SALEMI M., LAMERS S.L., HUYSENTRUYT L.C., GALLIGAN D., GRAY R.R., MORRIS A., ET AL. (2009). Distinct patterns of HIV-1 evolution within metastatic tissues in patients with non-Hodgkins lymphoma. *PLoS One*. **4**, e8153.

SALMINEN M., MARTIN D. (2009). Detecting and characterizing individual recombination events. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. P. Lemey, Salemi M. and Vandamme A.-M. New York, Cambridge University Press. 519-548.

SHANKARAPPA R., MARGOLICK J.B., GANGE S.J., RODRIGO A.G., UPCHURCH D., FARZADEGAN H., ET AL. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489-10502.

SHRINER D., RODRIGO A.G., NICKLE D.C., MULLINS J.I. (2004). Pervasive genomic recombination of HIV-1 in vivo. *Genetics*. **167**, 1573-1583.

SLATKIN M., MADDISON W.P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*. **123**, 603-613.

SMITH G.J., VIJAYKRISHNA D., BAHL J., LYCETT S.J., WOROBEY M., PYBUS O.G., ET AL. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*. **459**, 1122-1125.

STRIMMER K., VON HAESELER A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. USA*. **94**, 6815-6819.

STRIMMER K., PYBUS O.G. (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* **18**, 2298-2305.

STRIMMER K., VON HAESELER A. (2009). Genetic distances and nucleotide substitution models - Theory. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. P. Lemey, Salemi M. and Vandamme A.-M. New York, Cambridge University Press: 111-125.

TALBI C., LEMEY P., SUCHARD M.A., ABDELATIF E., ELHARRAK M., NOURLIL J., ET AL. (2010). Phylodynamics and human-mediated dispersal of a zoonotic virus. *PLoS Pathog.* **6**, e1001166.

WHEWELL W. (1837). History of the inductive sciences.

WHEWELL W. (1840). Philosophy of the inductive sciences.

XIA X., XIE Z., SALEMI M., CHEN L., WANG Y. (2003). An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* **26**, 1-7.

ZEHENDER G., EBRANATI E., BERNINI F., LO PRESTI A., REZZA G., DELOGU M., ET AL. (2011). Phylogeography and epidemiological history of West Nile virus genotype 1a in Europe and the Mediterranean basin. *Infect. Genet. Evol.* **11**, 646-653.

ZUCKERKANDL E., PAULING L. (1962). Molecular disease, evolution, and genetic heterogeneity. *Horizonz in Biochemistry*. M. Kasha and Pullman B. New York, Academic Press: 189-225.

ZUCKERKANDL E., PAULING L. (1965). Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*. V. Bryson and Vogel H. J. New York, Academic Press: 97-166.