Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics

Anna Zhukova, Olivier Gascuel, Sebastián Duchene, Daniel Ayres, Philippe Lemey, Guy Baele

Chapter 5.3 Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics

Anna Zhukova
Unité Bioinformatique Evolutive, Hub Bioinformatique et Biostatistique, USR3756 (C3BI/DBC), Institut Pasteur & CNRS, Paris, France
anna.zhukova@pasteur.fr

Olivier Gascuel
Unité Bioinformatique Evolutive, USR3756 (C3BI/DBC), Institut Pasteur & CNRS, Paris, France

Sebastián Duchêne
Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Australia

Daniel L. Ayres
Center for Bioinformatics and Computational Biology, University of Maryland, USA

Philippe Lemey[1]
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium

Guy Baele[2]
Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium
guy.baele@kuleuven.be

Abstract

Viral evolutionary analyses are confronted with increasingly large sequence data sets, both in terms of sequence length and number of sequences. This can result in considerable computational burden, not only to infer phylogenies but also to obtain associated estimates such as their time scales and phylogeographic patterns. Here, we illustrate two frequently-used approaches to obtain phylogenomic estimates of time-measured trees and spatial dispersal patterns for fast-evolving viruses. First, we discuss computationally efficient procedures that employ a fixed tree topology obtained through maximum likelihood inference to estimate molecular clock rates and phylogeographic spread for Dengue virus genomes. Using the same viral example, we also illustrate Bayesian phylodynamic inference that jointly infers time-measured trees and phylogeography, including covariates of spatial dispersal, from sequence and trait data. We highlight state-of-the-art efforts to perform such computations more efficiently. Finally, we compare the estimates obtained by both approaches and discuss their strengths and potential pitfalls.

How to cite: Anna Zhukova, Olivier Gascuel, Sebastián Duchêne, Daniel L. Ayres, Philippe Lemey, and Guy Baele (2020). Efficiently Analysing Large Viral Data Sets in Computational Phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the*

[1] PL acknowledges support by the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G066215N, G0D5117N and G0B9317N).
[2] GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094, and the Research Foundation – Flanders ('Fonds voor Wetenschappelijk Onderzoek – Vlaanderen', G0E1420N).

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 5.3; pp. 5.3:1–5.3:43

A book completely handled by researchers.
No publisher has been paid.

Genomic Era, chapter No. 5.3, pp. 5.3:1–5.3:43. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Acknowledgements This work was supported by the EU-H2020 Virogenesis project (grant number 634650), by the INCEPTION project (PIA/ANR-16-CONV-0005), and by the Reservoir-DOCS ERC grant (agreement no. 725422). The Artic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z.

1 Introduction

According to a "quick guide" to phylogenomics (Telford, 2007), standard phylogenomic approaches leverage the information present in a large number of genes generated in large-scale genome sequencing efforts. In infectious disease research, phylogenomic alignments that comprise single-nucleotide polymorphism (SNP) data of several hundreds or thousands of genes are now also an important focus of modern microbiology studies. However, for rapidly evolving RNA viruses, phylogenetic and evolutionary analyses remain inherently restricted to small genomes that encode for a limited number of genes. In this chapter, we focus on current challenges and opportunities for evolutionary inference from rapidly evolving viral genomes. This goes beyond common phylogenomic approaches and it is perhaps more in line with some of the earliest published mentions of phylogenomics that refer to a mixed bag of gene or genome analyses within a phylogenetic framework. We will primarily focus on the many different types of analyses that are being used in epidemiology, which shares many common interests with the field of phylodynamics.

The mention of "large" viral data sets in our title refers to the increase in two dimensions of the information available for studying molecular epidemiology and virus evolution, which has been brought about by the revolution in sequencing technology. The first dimension concerns the transition from a single gene or a typical PCR amplicon sequenced by Sanger sequencing to complete genomes that are now easily obtained through next generation sequencing, which roughly represents an increase of one order of magnitude for many RNA viruses. This seems less impressive than the increase from a single gene to hundreds of genes that phylogenomic studies of many other organisms have to confront. In addition, due to evolutionary rates that are about a million times faster than our own cellular genes, a limited marker in an RNA virus genome may already offer reasonable resolution for reconstructing evolutionary histories with time-scales of a decade or older. For example, a polymerase gene fragment of about 1 000 bp that is routinely sequenced for drug resistance testing has been extensively and successively used in HIV molecular epidemiology (e.g. Hué et al. 2005) while a fragment of less than half this size – but for the more variable envelope gene – has been used to reconstruct the origin and spread of the virus in Central Africa (Faria et al., 2014). Nevertheless, complete genomes offer an important increase in the resolution of the inferred phylogenies (Yebra et al., 2016), which for short-term outbreak dynamics in particular opens up new opportunities for epidemic reconstructions and tracking transmission. We illustrate this increase of phylogenetic resolution by comparing maximum likelihood trees for complete genome data and the corresponding glycoprotein gene sequences for the 2013 – 2016 West African Ebola virus outbreak in Figure 1. In this case, a single gene does not provide sufficient information about clustering of Ebola virus isolates whereas complete genomes do offer reasonable phylogenetic resolution allowing to identify some degree of structuring by country of sampling.

Complete genome sequencing has in recent years become the standard in outbreak

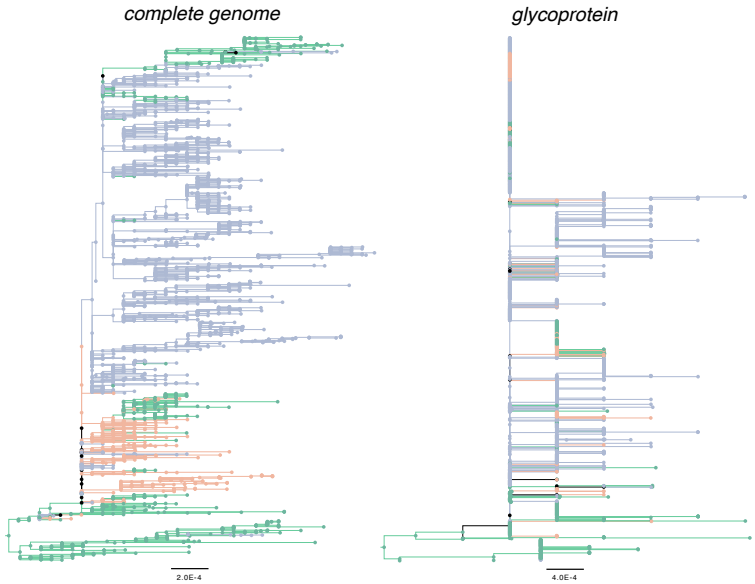

Figure 1 Maximum likelihood phylogenetic trees of Ebola virus inferred using complete genomes (left) and only the glycoprotein gene (right). Branches are coloured according to country (green: Guinea; blue: Sierra Leone; red: Liberia), based on a parsimony reconstruction for internal nodes. Complete genome data allow to infer reasonably resolved phylogenetic trees with a discernible structuring of lineages by country, whereas a single gene produces a multitude of polytomies and essentially prevents from uncovering any relevant (geographic) structure in the resulting phylogeny.

surveillance as illustrated in recent work on Zika (Faria et al., 2017), Chikungunya (Naveca et al., 2019) and Yellow Fever (Faria et al., 2018). Such complete genome data also come with specific challenges, such as the need to take into account recombination in different viruses or the difficulty to combine segments for viruses that undergo reassortment, or with the general challenge of increased computation times in likelihood-based phylogenetics (Chapter 1.2 [Stamatakis and Kozlov 2020]). In this chapter, we devote a great deal of attention to describing the methods that can be employed to address the computational challenges caused by these increases in data set sizes and model complexity. Related to this, Chapter 5.4 (Ayres et al. 2020) describes improvements in the latest version of the BEAGLE library for high-performance likelihood computation (Ayres et al., 2019), which enables parallel calculation of independent data partitions on powerful multi-core computing solutions such as GPUs.

The other dimension we focus on is the increasing sampling intensity leading to the availability of large numbers of sequences for viral evolutionary reconstructions. This is also illustrated by the sequencing efforts during the 2013-2016 West African Ebola virus outbreak that yielded over 1 600 complete genomes (Figure 1), representing over 5% of the known cases and making it the most densely sampled acute viral outbreak to date. This upscaling in sequencing is also impacting the molecular epidemiology of viruses with older transmission histories. For example, the genome sequencing efforts of the "Phylogenetics and Networks for Generalised HIV Epidemics in Africa" consortium (PANGEA-HIV) recently presented almost 4 000 HIV consensus sequences from different cohorts across sub-Saharan Africa (Ratmann et al., 2017). Such initiatives are transforming HIV molecular epidemiology into "big data" science, which hopefully can lead to new insights into HIV-1 transmission dynamics that can be translated to prevention strategies. However, large numbers of sequences also represent

5.3:4 Efficiently Analysing Large Viral Data Sets

a tremendous challenge for computational analyses, even more so than the length of the sequences. Indeed, it is well-known that phylogenetic likelihood computations scale linearly with alignment length, but the number of possible tree topologies grows super-exponentially with the number of taxa, which makes the search for optimal trees or distributions a highly cumbersome task. For this reason, we dedicate a large section in this chapter on phylogenetic approaches aimed at tackling the large data problem in viral sequence analyses.

Viral genomic data analyses are now frequently performed in the context of phylodynamics, a term that was originally introduced to describe "the melding of immunodynamics, epidemiology and evolutionary biology" (Grenfell et al., 2004). A key focus of phylodynamics is how the genetic diversity of viral pathogens is shaped by epidemic processes and natural selection (mostly in the context of immunological processes). Due to high mutation rates, large population sizes and short generation times, RNA viruses evolve at high evolutionary rates ensuring that their genomes can accumulate substitutions even over short-term epidemic time-scales. Sampling viral genomes over the time-scale of such epidemic processes therefore allows us to capture the relationship between time elapsed and sequence divergence. This is illustrated by plotting the root-to-tip divergence for the taxa in Ebola trees (Figure 1) as a function of sampling time in Figure 2, which can easily be done using software packages such as TempEst (Rambaut et al., 2016a). In this case, complete genomes show an increasing divergence over the sampling time range (Figure 2, left) while no such temporal signal is apparent in the corresponding glycoprotein gene sequences (Figure 2, right).

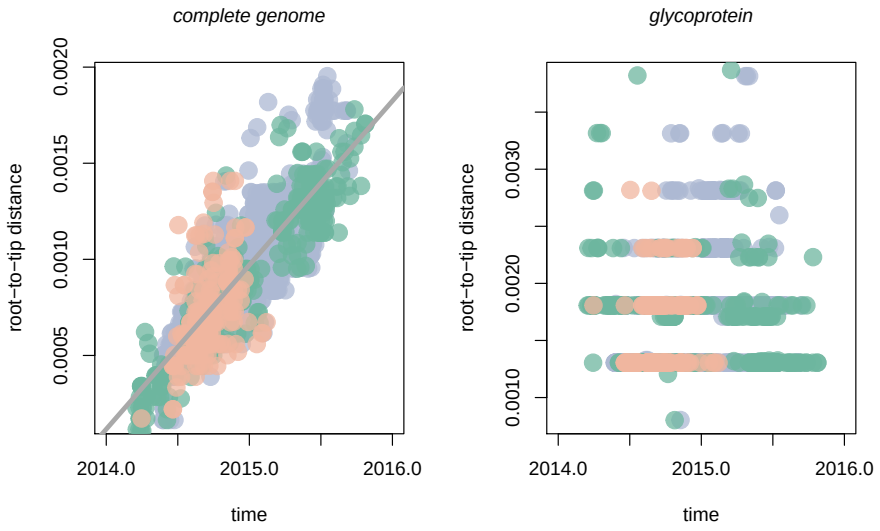

Figure 2 Plotting root-to-tip divergence as a function of sampling for the 2013 – 2016 Ebola virus data set. Data points are coloured according to country of sampling (green: Guinea; blue: Sierra Leone; red: Liberia). A comparison between complete genomes (left) and the glycoprotein gene (right) reveals a clear increase in divergence over the sampling time versus no temporal signal respectively.

This relationship between time and sequence divergence can be used to inform or calibrate molecular clock models (see Chapters 4.4 and 5.1 [Bromham 2020; Pett and Heath 2020]). These models are now routinely applied to phylogenetic trees, or integrated in phylogenetic inference, in order to estimate trees in units of time allowing us to date the epidemic origins or key (transmission) events in epidemic histories. Time-measured trees are also the necessary

prerequisite for the application of coalescent models that aim at estimating changes in epidemic size through time, and of birth-death models that estimate key parameters that determine an epidemic, such as the basic reproductive rate (R_0), i.e. the average number of cases one case generates over the course of their infectious period assuming a fully susceptible population (e.g. Fraser et al. 2009).

More broadly, pathogen genomes have been shown to contain a wealth of information concerning population size dynamics and the process of spatial spread that generated the geographic distribution of an epidemic, which can be recovered using a wide variety of demographic and phylogeographic models within a phylodynamics framework (Minin et al., 2008; Lemey et al., 2009, 2010; Gill et al., 2013, 2016). We illustrate some of the insights these approaches can obtain using specific viral examples.

Data

In this chapter we illustrate computational proaches on a data set of dengue virus (DENV), the most common vector-borne viral disease of humans. The dengue viruses are members of the genus *Flavivirus* in the family *Flaviviridae*, with four serotypes of dengue virus having been discovered to date. The four dengue serotypes are relatively closely related, but even within a single serotype, there is considerable genetic variation (Blok, 1985), with each serotype being further divided into several genotypes. The serotypes diverged approximately $1\,000 - 2\,000$ years ago; the genotypes within each serotype diverged much more recently, about $100 - 200$ years ago (Pollett et al., 2018). Despite these variations, DENV infections result in similar disease and clinical symptoms irrespective of serotype/genotype. Dengue serotypes are increasingly co-circulating in most regions of the world, particularly in Latin America and Asia (Messina et al., 2014), with global phenomena such as urbanisation and international travel acting as key factors in facilitating the spread of dengue.

We here perform phylodynamic inference on a dengue virus data set consisting of 997 genomes spanning the global dengue diversity, with a total of 6 869 unique site patterns across 10 gene-based partitions. This data set was generated in 2014 by downloading from GenBank all 3 289 available dengue genomes with known sampling year and country. Based on a maximum likelihood tree reconstruction, we used Phylogenetic Diversity Analyzer (Chernomor et al., 2015) to select the most diverse subset of 1 000 genomes. Three outliers according to root-to-tip regression explorations were excluded (Rambaut et al., 2016a). The nucleotide partitions correspond to the ten protein-coding genes that make up the dengue genome.

The 997 samples of this data set are annotated with discrete location states, one of the most popular traits associated with virus data sequences. This allows us to perform spatial ancestral state reconstruction in order to determine the origin of specific outbreaks and track viral spread over time. Owing to the widespread nature of dengue viruses, a total of 64 countries across six continents are used as location states to perform such a reconstruction. The number of taxa and the state dimensionality make for a computationally demanding joint inference of nucleotide and trait evolutionary processes.

In this chapter, we show how to analyse such a challenging data set using two popular inference frameworks: an approach oriented towards maximum likelihood inference (Section 2) and a fully Bayesian inference approach through Markov chain Monte Carlo (MCMC) (Section 3).

To illustrate the ability of maximum likelihood methods to handle very large data sets, we used an additional, larger, data set of 5 132 full dengue genomes downloaded from GenBank (Benson et al., 2012). The sequences were serotyped and genotyped with GenomeDetective (Vilsker et al., 2019) and those with type support < 100 removed. When

available, the sequences were annotated with the collection date and country metadata from the Entrez molecular biology database system (Sayers et al., 2009). The larger data set includes all the sequences from the smaller one, represents more (83 vs 64) countries, and is more noisy: no temporal outlier filtering or diversity-based selection was performed; moreover, some of the sequences had no sampling date (4%) or no location (1%) metadata.

2 Analysis of large phylogenies with Maximum Likelihood

Maximum likelihood (ML) inference is a procedure that for a given model finds the parameter values that maximise the observed data likelihood, thereby producing a single (maximum-likelihood) estimate of – for example – a phylogeny. Typically, an ML approach splits the analysis into several steps forming a pipeline, where the phylogeny reconstruction from sequence data is followed by further analysis (e.g. divergence time estimation, ancestral location reconstruction, ...) assuming a fixed phylogenetic tree. Using the dengue data set analysis as an example, we describe an ML pipeline for phylogeographic analysis of virus spread over time, consisting of the following steps detailed below: phylogenetic tree reconstruction from multiple sequence alignment, tree rooting and dating, and ancestral character reconstruction for geographic data.

2.1 Tree reconstruction

Phylogenetic inference using ML aims at finding the tree and model parameters that maximise the likelihood and is known to be NP-hard (Chor and Tuller, 2005). ML tree reconstruction tools generally approach the problem by performing a "hill-climbing" optimisation, i.e. these methods start by generating an initial tree (e.g., a randomly generated tree or a maximum-parsimony tree), and then keep replacing it with a better (in terms of likelihood) neighbouring tree (obtained using certain topological rearrangements), until no better tree can be found. A potential danger in a pure hill-climbing is the possibility to get trapped in a local optimum. To overcome this issue, it is recommended to perform the optimisation starting from multiple initial trees, and then keep the best result, and to use more expansive techniques for searching neighbouring tree space (Zhou et al., 2018). The most common topological rearrangement algorithms for finding a neighbour tree are Nearest-Neighbour-Interchange (NNI), which swaps two non-sibling subtrees adjacent to an internal branch (Robinson, 1971), and Subtree-Pruning-and-Regrafting (SPR), which prunes a subtree from the initial tree and regrafts it onto a different branch (Swofford et al., 1996). SPR can evaluate many more trees (quadratic to the number of tips) from one initial topology than NNI (linear), but as a consequence it is also much slower (Allen and Steel, 2001) and therefore different heuristics have been developed to filter out the unpromising SPR candidates (Stamatakis et al., 2005; Hordijk and Gascuel, 2005; Guindon et al., 2010).

Among multiple ML tools that are available for phylogenetic tree reconstruction, the most popular are FastTree (Price et al., 2010), PhyML (Guindon et al., 2010), RAxML (Stamatakis, 2014) (and its updated version RAxML-NG [Kozlov et al. 2019], see also Chapter 1.3 [Kozlov and Stamatakis 2020]), and IQ-TREE (Nguyen et al., 2015). FastTree works very well to perform a preliminary analysis as it can be orders of magnitude faster than other ML tools, but as a trade-off, generates less accurate tree estimates due to limited tree space exploration and less thorough branch length optimisation. FastTree starts with a distance-based optimisation using both NNI and SPR rearrangements, followed by ML-based NNI rearrangements to search for the final tree, using heuristics at all stages to limit the numbers of tree searches and likelihood optimisations. PhyML performs hill-climbing tree searches using both NNI and

SPR rearrangements (with a parsimony-based filtering of the least promising SPR moves), while RAxML/RAxML-NG implements SPR-based hill-climbing, employing heuristics to reduce the number of unpromising SPR candidates (typically regrafting the pruned subtree in a position remote from the original one). IQ-TREE combines hill-climbing algorithms with random perturbations of the current best trees and broad sampling of initial starting trees, and generates a pool of candidates containing the top 5 trees obtained by NNI hill-climbing from the 20 best parsimony-based starting trees. At each iteration, a randomly chosen candidate tree is perturbed with $0.5(n-3)$ random NNIs, where $n-3$ is the number of inner branches, and optimised with the hill-climbing NNIs. If the resulting tree is better than the best tree in the candidate pool, the iteration is considered successful and the best tree is replaced. Otherwise the worst tree in the candidate pool is replaced if it is worse than the resulting tree. The analysis terminates after a certain number of unsuccessful iterations. RAxML (and especially RAxML-NG) and IQ-TREE are parallelised which makes them faster than PhyML.

In a comparison of these different ML packages in terms of their accuracy, Zhou et al. (2018) recently analysed various middle-sized data sets of about 200 taxa, including genome and transcriptome data of fungi, animals and plants, using both single gene alignments as well as concatenated full genomes. In this comparison, IQ-TREE (version 1.5.5) outperformed RAxML (8.2.11) and PhyML (20160530) in most cases, with the exception of some of the largest data sets. Kozlov et al. (2019) repeated the comparisons on the same data sets using the latest available version of RAxML (RAxML-NG) and found it to be generating the highest tree likelihood while being 1.3 to 4.5 times faster than IQ-TREE. Similar results can be obtained with the most recent (but as of yet unpublished) version of PhyML (`https://github.com/stephaneguindon/phyml`; data not shown), showing that there is still room for improvement in these complex algorithms and programs.

In conclusion, for extremely large trees (dozens of thousands of tips) FastTree is probably the only choice due to its speed, while for the other cases (up to thousands of tips) we recommend to use several programs and compare the results.

2.1.1 Application to dengue data

The four dengue serotypes diverged thousands of years ago while the genotypes within each serotype diverged about 100 to 200 years ago (Pollett et al., 2018). We therefore expect a phylogeny with very long branches connecting the serotypes and much shorter ones within each serotype. This makes the common tree reconstruction challenging: On one hand, for deep phylogenies (like the inter-serotype one) one often uses amino acid alignments due to codon degeneracy and therefore loss of signal for the deep nodes on the nucleotide level (Rota-Stabelli et al., 2013). On the other hand, for the intra-serotype phylogenies, which contain much closer related sequences, amino acid alignments might not have enough resolution, and hence the nucleotide alignments should be preferred. To account for codon degeneracy a partitioning scheme with a distinct group for the third codon positions can be used.

We have reconstructed phylogenies for the smaller (997 sequences) and the larger (5 132 sequences) data sets with RAxML-NG (version 0.9.0, starting from a parsimonious tree) and IQ-TREE (version 1.6.9), both from the amino acid alignment (HIVb+I+G6 evolutionary model) and from the nucleotide one (GTR+I+G6) with a two-group partitioning scheme (Chernomor et al., 2016): for the codon positions $1-2$ and for the position 3. Evolutionary models were selected by the model selection tool SMS (Lefort et al., 2017), we increased the number of GAMMA categories from 4 (as used in SMS) to 6, for a better fit of

the gamma distribution of rates across sites.

On a machine with 12 cores, phylogeny reconstruction for the larger data set took 1 day 8 hours for RAxML-NG on the amino acid alignment, and 7 hours on the nucleotide alignment; for IQ-TREE it took 1 day 2 hours on the amino acid alignment, and 1 day 17 hours on the nucleotide alignment. In terms of likelihood RAxML-NG got a better result on the amino acid data (log likelihood of −174 295 vs −174 473), and IQ-TREE on the nucleotide one (−873 396 vs −873 406).

Once the reconstruction is done it is important to assess the results. It can be done using prior knowledge on the expected tree topology, and by comparing the phylogenies obtained with different tools. The tree topologies of the reconstructed phylogenies were overall close, especially those reconstructed on the same alignment. We formally assessed this using the normalised quartet distance (Estabrook et al., 1985) (calculated with tqDist [Sand et al. 2014]), which takes on values between 0.0 for identical trees and 1.0 for trees that have no quartet in common. The normalised quartet distance was 0.004 [DNA] and 0.008 [AA] for the tree pairs reconstructed with different tools on the same alignment, and varied from 0.015 (RAxML-NG [DNA] vs RAxML-NG [AA]) to 0.022 (IQ-TREE [DNA] vs IQ-TREE [AA]) for the tree pairs reconstructed on different alignments.

However, the reconstructed topologies (for both data sets) had differences for the deep parts of the phylogeny, i.e. serotype subtree rooting. Moreover, none of them corresponded to the the expected tree topology (prior knowledge). We expected to find monophyletic clades for the genotypes within the same serotype with a serotype root that is placed on one of the inter-genotype branches as done by Pollett et al. (2018); we will also confirm this assumption in Section 2.2 with rooting of serotype-specific trees based on dates. For DENV2 and DENV4 this was the case: the root was placed on the branch separating sylvatic genotype from the epidemic ones, all the genotypes were monophyletic, and their relative positions were consistent in all the reconstructed phylogenies. For DENV1 and DENV3 the relative genotype positions were consistent in all the phylogenies, but the root positions varied and the root was often placed within one of the genotypes (see Figure 3). The only exception was the phylogeny reconstructed by IQ-TREE on the nucleotide alignment: it managed to place the DENV1 root correctly (but still had issues with DENV3).

This difficulty with resolving deep parts of the phylogeny is likely due to the fact that all the DENV1 and DENV3 genotypes diverged relatively simultaneously (within dozens of years, while the root age is $1\,000 - 2\,000$ years), and there is a lack of sequences that diverged earlier (e.g. within hundreds of years). Moreover, the branches connecting serotype subtrees were extremely long: $0.56 - 1.1$ mutations per site. Bayesian methods can address this issue by incorporating temporal and phylogeographic information along with the alignment data, which increases the signal. In the maximum clade credibility (MCC) tree reconstructed with Bayesian methods (see Section 3), the rooting of DENV1, DENV2 and DENV4 is correct, however DENV3 subtree root is also misplaced within one of the genotypes (3.II).

To overcome the issues described above, we reconstructed serotype-specific subtrees from the nucleotide alignments using both RAxML-NG and IQ-TREE (as described before) and then kept the most likely tree for each serotype.

2.2 Tree rooting and dating

For data sets in which sufficient genetic change has accumulated during the window of sampling, the divergence in each sequence is expected to correlate with the date of sampling, as illustrated in the introduction (Figure 2). Hence, the sampling times of the sequences can be used to estimate the substitution rate and the divergence dates, transforming a phylogeny

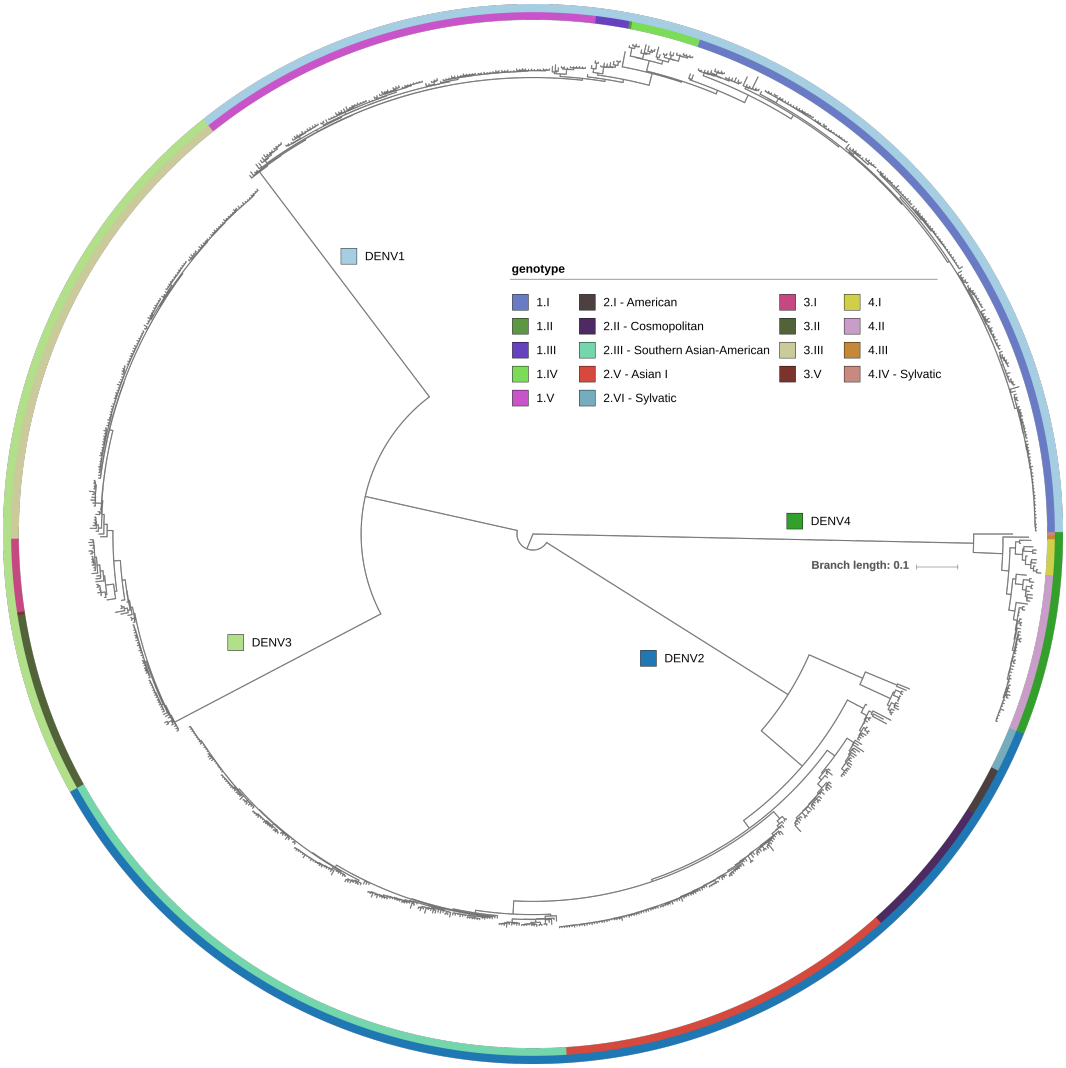

Figure 3 Phylogeny estimated with RAxML-NG on the nucleotide alignment for the small data set, visualised with iTOL (Letunic and Bork, 2007). The rooting of the DENV2 and DENV4 serotype subtrees is correct, with all the genotypes being monophyletic and the roots placed on the branches separating sylvatic genotypes from the epidemic ones. For DENV1 and DENV3 subtrees the roots are misplaced: instead of a branch separating different genotypes, the serotype roots are within one of their genotype trees (3.II for DENV3 and 1.V for DENV1), however the other genotypes are monophyletic.

into a time-scaled tree whose branches are measured in units of time, e.g. years. Various molecular clock models are available to perform such estimations, such as the strict clock (SC) – which assumes substitution rate homogeneity throughout the tree (Zuckerkandl and Pauling, 1965) – and the uncorrelated relaxed clock – which allows a different rate along each branch in the tree, but with rates assumed to follow a chosen statistical distribution (Bromham et al. 2018; Chapter 4.4 [Bromham 2020]). The methods for substitution rate and time-scaled tree estimation can be divided into two groups: those that incorporate sequence dates into the tree reconstruction framework, and the ones that estimate the substitution rate on a fixed phylogeny (with fixed branch lengths measured in substitutions per site).

The methods of the former type evolved from an intermediate approach where only the tree topology was fixed (but not the branch lengths) and were initially implemented in the program TipDate (Rambaut, 2000) (and later as an R package node.dating [Jones and Poon 2017]): given a rooted tree topology and the tip dates, TipDate optimises the substitution rate under the SC model and estimates the times of the internal nodes using ML under a given evolutionary model, e.g. HKY (Hasegawa et al., 1985). Drummond et al. (2001) first extended TipDate by allowing different rates to be estimated for different intervals of time, and later embedded them into a Bayesian framework for joint inference of mutation rate and population size that incorporates the uncertainty in the genealogy by using MCMC integration (Drummond et al., 2002).

Among the methods of the latter type, one of the very first ones was Root-To-Tip (RTT) regression (Shankarappa et al., 1999; Drummond et al., 2003a): assuming a strict clock, the root-to-tip distance in the phylogeny should be proportional to the corresponding elapsed time, and a regression of the root-to-tip distance as a function of tip dates provides estimates of the mean substitution rate (regression slope) and the root date (x-intercept). This constitutes a very fast method that allows estimating the root of the tree, e.g. by searching for a tree branch that minimises the sum of regression residues. However, it does not provide the dates for the internal nodes, and therefore does not output the time-scaled tree. Also, RTT regression violates the assumption of data independence as deep branches contribute to multiple RTT distances, it therefore is not suitable for statistical hypothesis testing and should rather be used as a data exploration tool (Rambaut et al., 2016b).

The LF (Langley and Fitch, 1974) model – implemented in r8s (Sanderson, 2003) – assumes a strict clock with a constant substitution rate, and a Poisson distribution for the number of substitutions along every tree branch. The substitution rate and the internal node dates are estimated by maximising the likelihood of the rooted input tree.

Least-Squares Dating (LSD) (To et al., 2016) uses a normal approximation to the LF model to estimate the substitution rate. LSD assumes the following relationship between the branch lengths in the initial phylogeny and the corresponding time-scaled tree:

$$b_i = y_i \cdot \omega + \epsilon_i,$$

where b_i is the length of the branch i measured in substitutions per site, y_i – its length in years, ω is the substitution rate, and $\epsilon_i \in N(0, \sigma_i^2)$ is a noise (error) term drawn from a normal distribution (independent for different branches). In other words, LSD assumes a strict clock, but the noise term makes it robust to uncorrelated violations. When run in "temporal precedence constrained mode", LSD additionally ensures that all the dated branch lengths are non-negative (which could otherwise be violated due to the noise terms for short branches). LSD minimises the error using the weighted least squares criterion:

$$\sum_i \frac{1}{\sigma_i^2}(b_i - y_i \cdot \omega) \to min,$$

where the variance terms are derived from the Poisson nature of the substitution process as

$$\hat{\sigma_i^2} = \frac{b_i + c/S}{S},$$

S being the sequence length, and c – a constant smoothing factor. Uncertainty can be obtained via parametric bootstrap or by repeating the analyses on a set of non parametric bootstrap trees. LSD can root and date large phylogenies in quadratic time (i.e. proportional to the number of tips squared). The new (but as of yet unpublished) version of LSD – LSD2 (`https://github.com/tothuhien/lsd2`) – extends the tool with several new features, including outlier detection (sequence or date annotation errors or samples that are poorly described by the fitted substitution model), and the local clock model (Yoder and Yang, 2000).

Treedater (Volz and Frost, 2017) extends the concepts of LSD by implementing outlier detection and an uncorrelated relaxed molecular clock (i.e. the global substitution rate ω is replaced with a collection of branch-specific rates ω_i drawn from a common gamma distribution). Treedater implements a heuristic iterative approach to optimise both the parameters of the gamma distribution (shape and scale) and the internal node times. It initialises branch-specific rates to the common rate estimated by RTT, then repeats the optimisation cycle until convergence (defined by a tolerance threshold). At each iteration, first the internal node times are calculated based on the branch-specific rates by solving a (constrained) least-squares problem like in LSD, secondly the gamma distribution parameters are optimised based on the new ancestral dates using gradient-descent. Therefore when compared to LSD, the computation time is multiplied by the number of iterations. Moreover, it is recommended to repeat the optimisation with different starting conditions to avoid local optima.

Treedater also implements a statistical test for selecting the appropriate clock model. In a comparison performed by Volz and Frost (2017), treedater outperformed LSD and BEAST on 9 out of 16 simulated data sets, with all methods showcasing their strengths and weaknesses in at least one scenario.

TreeTime (Sagulenko et al., 2018) is an ML relaxed clock method that allows to either optimise the branch lengths on a given tree topology (along with ancestral sequence reconstruction) or to use the branch lengths provided in an input phylogeny. TreeTime uses dynamic programming for branch length optimisation and its run times scale linearly in the size of the data set.

Duchêne et al. (2016b) have compared the rate estimation by Bayesian, least-squares and RTT regression methods on 81 RNA and DNA virus data sets (9 to 120 sequences of 350 to 10 066 nucleotides sampled over 0.5 to 86 years), and observed that the methods largely produce congruent estimates of substitution rates, provided that the data meet certain criteria, such as the absence of high among-lineage rate variation, congruence between the tree topology and no phylogenetic and temporal clustering. Moreover Duchêne et al. (2016b) pointed out that clock-model testing should be routinely performed, as the use of relaxed molecular clocks can lead to overestimates of the mean substitution rate when the data in fact fit a strict clock.

2.2.1 Application to dengue data

ML tree reconstruction methods estimate an unrooted phylogeny, which one can either root using an outgroup, or using the sampling dates and a molecular clock model. In the case of the dengue data set, no outgroup is present for the full tree (though different serotypes/genotypes

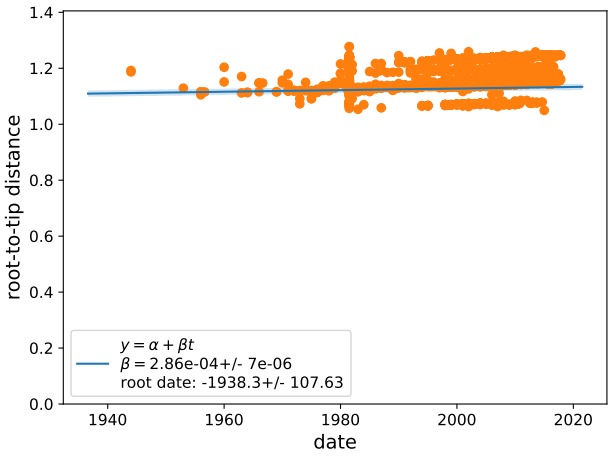

Figure 4 Results of RTT regression (performed using TreeTime) for the large dengue data set (RTT regression on the small data set is similar). The RTT distance is plotted as a function of the tip dates and provides estimates of the mean substitution rate (regression slope) and the root date (x-intercept). The fit line is almost horizontal, therefore an accurate root date estimation is almost impossible: the slightest error in the slope would lead to a large error in the x-intercept.

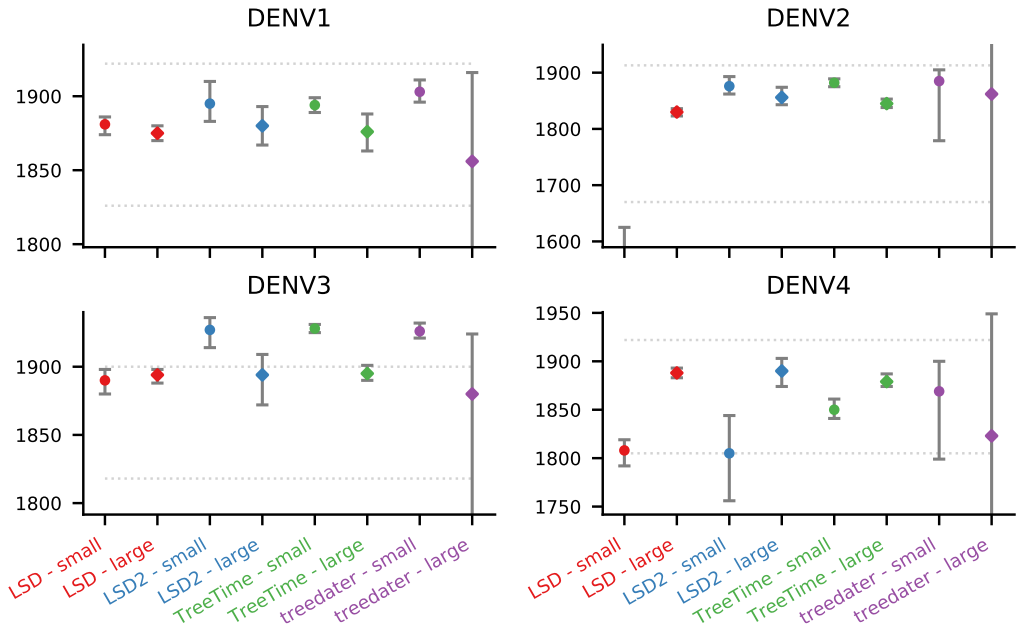

Figure 5 Estimated dates of emergence of each of the four dengue human serotypes [years] (with 95% CIs), obtained with treedater, TreeTime, LSD, and LSD2. The estimates obtained on the smaller data set are shown with circles, on the larger one – with diamonds. Median 95% CIs from the literature that are summarised in Table 3 of (Pollett et al., 2018) are shown by dotted horizontal lines. The estimates provided by different tools are generally within the CIs from the literature (apart from DENV3 small data set and LSD estimate for DENV2 small data set (due to outliers)) but vary for different data sets and tools, which could be due to not enough quality control for the samples included in the data sets.

can serve as outgroups for its subtrees), and we therefore had to resort to using sampling dates for rooting. RTT regression suggests the presence of temporal signal in the data, but potentially too short of a sampling window (with respect to the total time span from the most recent common ancestor up to the present) for the full tree: the x-intercept (root date) of the RTT plot for the full tree (Figure 4) is very old, while the slope (rate) is almost horizontal, therefore the slightest error in the rate estimation (slope) would result in a large error in the root date (x-intercept). Moreover, underestimation of evolutionary distances in deep branches separating serotypes might obfuscate temporal signal (Duchêne et al., 2016a). If we had ancient samples, their dates might have helped to calibrate the slope. Dating, rooting and confidence interval (CI) estimation (the most time consuming part) on the full tree was performed in 2 minutes by LSD/LSD2, in 1 hour 30 minutes by TreeTime, and in 6 days by treedater.

To assess the performance of various ML packages on more recent data, we dated the four serotype phylogenies separately. We used four recently developed applications that allow dating unrooted phylogenies: LSD (version 0.3beta), LSD2 (version 1.4, strict clock), treedater (version 89a0df0) and TreeTime (version 0.5.5, using branch lengths of the input tree). TreeTime and treedater implement both relaxed and SC, and are able to detect outliers, while LSD only implements SC and does not detect outliers. LSD2 extends LSD with outlier detection.

The clock model selection test performed by treedater identified the relaxed clock as the model of choice, therefore we run treedater and TreeTime with the uncorrelated relaxed clock model. For rooting we used the sylvatic outgroup for DENV2 and DENV4, and estimated the root position from dates for DENV1 and DENV3. Note that not all of the outliers detected by different methods were the same. TreeTime generally finds more outliers than LSD2 which in turn finds more than treedater, e.g. for DENV4 subtree TreeTime detected 35 outliers, LSD2 detected 14 (5 of which were among the ones detected by TreeTime), while treedater detected none.

The rates estimated by different tools were close to those reported in the literature (see Table 2 in (Pollett et al., 2018) for a summary) for DENV1-2 and about 2 times lower for DENV3-4. Figure 5 shows the estimated dates of emergence of the four serotypes, where the dotted horizontal lines indicate the median 95% CIs from a literature summary provided in Table 3 of (Pollett et al., 2018). All the estimates are consistent with literature, except for those obtained on DENV3 small data set (potentially due to root position as explained below) and the LSD estimate on DENV2 small data set (due to outliers). Additionally, we observe several phenomena: (1) As expected, the full tree was harder to date than the serotype trees, which led to much larger CIs for the dates, e.g. $[-6392, -1957]$ for treedater (data not shown); (2) The estimates provided by LSD2 (after outlier removal) are often different from those of LSD (with outliers), suggesting an important impact of outliers; (3) CIs calculated by treedater are much larger and generally include CIs from other methods. CIs estimated by different tools on the large data set overlap more, suggesting that adding more data helps to increase the signal; (4) The estimates provided by different tools are quite consistent for DENV1 but much less so for other serotypes, which could be due to the absence of quality control for the samples included in the data sets, such as removal of clone sequences, duplicates, erroneous metadata, etc.

The estimated root position was consistent among all tools and data sets for the full tree (on the branch separating DENV4 from the other serotypes) and DENV1 tree (on the branch separating the common ancestor of genotypes III and V from the other genotypes); and varied for DENV3 tree. For DENV3 three scenarios for the root position were present: (1)

on the branch separating genotype II from the other genotypes (LSD, LSD2 and TreeTime on the small data set); (2) on the branch separating the common ancestor of genotypes II and III from the common ancestor of genotypes I and V (LSD, LSD2 and TreeTime on the large data set, and treedater on the small data set); and (3) unresolved root, with genotype II, genotype III and the ancestor of genotypes I and V as children (treedater on the large data set). Note that (3) represents a consensus between (1) and (2), moreover the branches that needed to be collapsed to reach this consensus with the time-scaled trees obtained by other tools were short (1.5-5 years).

Overall, our analyses indicate that dating an unrooted tree is a complex task, especially when combined with using a relaxed clock and outlier detection and removal. In theory, the use of a relaxed clock should reduce the number of outliers compared to a strict clock but both depend highly on the root position. When applied to noisy (real) data, dating becomes truly challenging.

2.3 Phylogeography

Analyses of epidemic spread through space and time are often performed by defining a finite, non-ordered set of locations (e.g. countries) and using ancestral character reconstruction (ACR) along a phylogenetic tree with the defined locations as possible character states, based on the known tip locations. ACR aims to unravel how the character has changed on the tree from the root to the tips through time, by assigning the most likely ancestral character state to each internal node. Various inference methods can be used to perform ACR, and a range of software packages is available for each type of inference.

Parsimony-based ACR methods infer a scenario with minimum state changes along the tree. These methods are quick and simple, however, due to the over-simplification of evolutionary processes (e.g. not accounting for branch lengths and evolutionary times), parsimony has limited accuracy (Zhang and Nei, 1997; Collins et al., 1994). ML and Bayesian approaches on the other hand are based on probabilistic models of character evolution that adapt standard nucleotide substitution models to s-state (discrete) trait characters. They have been shown to outperform parsimony methods, using both theoretical arguments and simulation studies under a variety of conditions (Zhang and Nei, 1997; Gascuel and Steel, 2014), and are also robust to moderate model violations and phylogenetic uncertainty (Hanson-Smith et al., 2010).

Statistical ACR methods employ continuous time Markov chains (CTMCs) models that emit discrete outcomes as a continuous function of time. This process is assumed to be memoryless, in that the probability of transitioning to a new location only depends on the current location and not the past history. As with nucleotide substitution models, an $s \times s$ infinitesimal rate matrix $\Lambda = \{\lambda_{ij}\}$ completely characterises the CTMC process (Lemey et al., 2009). The rate matrix Λ contains non-negative off-diagonal entries and all rows sum to 0, yielding a stochastic matrix upon exponentiation. To determine the finite-time transition probabilities between states (or locations) over a branch of length t, the following matrix exponentiation is required:

$$P(t) = e^{\Lambda t} \tag{1}$$

In its most general form, such a discrete trait substitution model allows a unique instantaneous rate between each pair of character states to be estimated, which may prove problematic because these rate parameters are only informed by a single discrete trait character observed

at the tips of the phylogeny (Gascuel and Steel, 2019). Hence, simpler models are often used, such as s-state generalisations of 4-state JC and F81 models for DNA (Jukes and Cantor, 1969; Felsenstein, 1981). Under F81-like models, migration rate from a state (location) i to a state j ($i \neq j$) is proportional to the equilibrium frequency of j, π_j; JC-like models are a special case where all equilibrium frequencies are equal: $\forall i\ \pi_i = 1/s$. A big computational advantage of F81-like models is that the probability of changes along a branch of length t can be calculated with a simple formula:

$$P_{i\to j}(t) = \left\{ \begin{array}{ll} (1-e^{-\mu t})\pi_j, & \text{if } j \neq i \\ e^{-\mu t} + (1-e^{-\mu t})\pi_j, & \text{otherwise} \end{array} \right\}, \text{where } \mu = 1/(1-\sum_i \pi_i^2). \quad (2)$$

Dudas et al. (2017) showed that the origin and destination population sizes (π_i and π_j) are two of the main factors explaining Ebola dissemination in West-Africa. This advocates for the use of s-state F81-like models, where the expected number of changes from i to j is proportional to $\pi_i\pi_j$. Moreover, Gascuel and Steel (2014) showed with simulations on DNA-like data generated using an HKY model (Hasegawa et al., 1985) with high transition/transversion rate and heterogeneous nucleotide frequencies that even the simpler JC-like model performs nearly as well as the true one.

In the likelihood framework, to predict ancestral character states based on the selected model of character evolution, one commonly uses the marginal posterior probabilities of the character states (Felsenstein, 1981; Yang, 2007), the joint reconstruction of the most likely scenario (Pupko et al., 2000), or an approach that lies in between these two extremes. Marginal reconstructions provide users with state probabilities, but these are difficult to interpret and visualise, while joint reconstructions select a unique state for every tree node and thus do not reflect the uncertainty of inferences. Intermediate approaches overcome these limitations by predicting a unique state in the regions of the tree that are easy to estimate (typically close to the tips [Gascuel and Steel 2014]), while keeping several likely states in the more difficult regions (typically close to the root), reflecting the uncertainty of the inferences.

PastML (Ishikawa et al., 2019) is a fast ACR tool that implements several parsimony and ML ACR methods: Joint, MAP (maximum a posteriori) that selects the state with the highest marginal probability, and MMPA (marginal posterior probabilities approximation), an intermediate approach that uses decision-theory concepts and the Brier criterion to associate each node in the tree to a set of likely states: a unique state is predicted in the tree regions with low uncertainty, while several states are predicted in the uncertain regions.

An important choice to take before performing a phylogeographic ACR is that of the precision level and whether the geographic sampling captures the range of the pathogen. Choosing a character with many possible states on a small data set can be limiting in terms of phylogeographic signal and can leave many nodes unresolved between several states for methods like MPPA, or predict a poorly supported state (e.g. low marginal probability even for the most likely state) for unique-state methods like Joint or MAP. Moreover, if the states present in the data (tree tips) do not cover all the possibilities, it can bias the predictions, as for instance it is impossible to predict missing countries. Biases in sampling for the states that are represented can also affect the reconstructions. Such biases are prominent in most real data sets, including the examples we study here.

To strengthen the signal extracted from the data, one needs to increase the number of sequences for each character state, sampled over a larger time range. This can be achieved

either by adding more sequences annotated with each state to the data set, or by generalising the annotations to decrease the number of character states, by grouping countries into broader geographic regions.

2.3.1 Application to dengue data

We reconstructed ancestral geographic characters with PastML (version 1.9.20) using the MPPA method and the F81-like model. Performing the country ACR on the full tree of the large data set (4 767 tips after outlier removal by LSD2) took 3 hours 30 minutes. The large data set includes sequences from 83 different countries, of which 65 are present in the smaller data set. Moreover 23 countries are present only once and hence do not provide sufficient information on the migration to and from them. As a result, 3.7% of internal nodes remain unresolved between several countries (10.3% for the small data set). As expected, the majority of unresolved nodes are found deeper in the tree, corresponding to the root and the common ancestors of different serotypes and genotypes.

The difference in percentage of unresolved nodes between the small and large data sets shows that adding more data increases the signal. The ACRs for the two data sets are generally compatible, while some of the nodes that remain unresolved in the small data set are resolved in the large one (see Figure 7), suggesting that the method is to some extent robust against sampling variation.

To further increase the signal we generalised the countries into 11 geographic regions: South America, Caribbean, Central America, Northern America, Africa, Europe, Western Asia, Eastern Asia, South-eastern Asia, Southern Asia, and Oceania. This allowed to reconstruct states closer to the root, and reduce the percentage of unresolved internal nodes to 1.2%. Finally, when we generalised the locations even further, into five continents (Americas, Africa, Europe, Asia, Oceania), the number of unresolved internal nodes was reduced to only 0.5%.

Summarised ancestral scenarios for location reconstruction on the full tree and for country on the DENV3 subtree are shown in Figures 6 and 7. PastML visualises these scenarios by (1) clustering the parts of the tree where no state change happens into meta-nodes (whose size corresponds to the number of samples (tips) they contain); (2) clustering independent events of the same kind into meta-edges (whose size corresponds to the number of such events). For example, the large "South-eastern Asia 1 503" node in Figure 6 represents the reconstructed cluster of DENV1 spread in South-eastern Asia, which includes 1 503 sequences in our data set; while its "Eastern Asia $1-7$" child node connected by a meta-edge of size 36 represents 36 independent DENV1 transmissions from South-Eastern Asia to Eastern Asia.

The predicted ancestral locations generally agree with previous studies, performed on different dengue data sets. For instance, in the study of global DENV2 phylogeography by Walimbe et al. (2014) performed using Bayesian inference on 307 DENV2 E-gene sequences from GenBank (sampled between 1944 and 2011), the authors also could not pinpoint the ancestral location for sylvatic and epidemic strains, and detected Southeast Asia as the ancestral region for the Asian/Asian-American and Cosmopolitan genotypes. The authors also found multiple migrations from the Caribbean to the American mainland (see Figure 6 for our predictions).

In the study of spatio-temporal dynamics of dengue in Colombia, performed on 143 newly sequenced samples from Colombia (sampled between 1998 and 2015) combined with full-length E-gene sequences retrieved from GenBank, Jiménez-Silva et al. (2018) performed a Bayesian analysis of spatial spread and detected significant viral diffusion between Venezuela

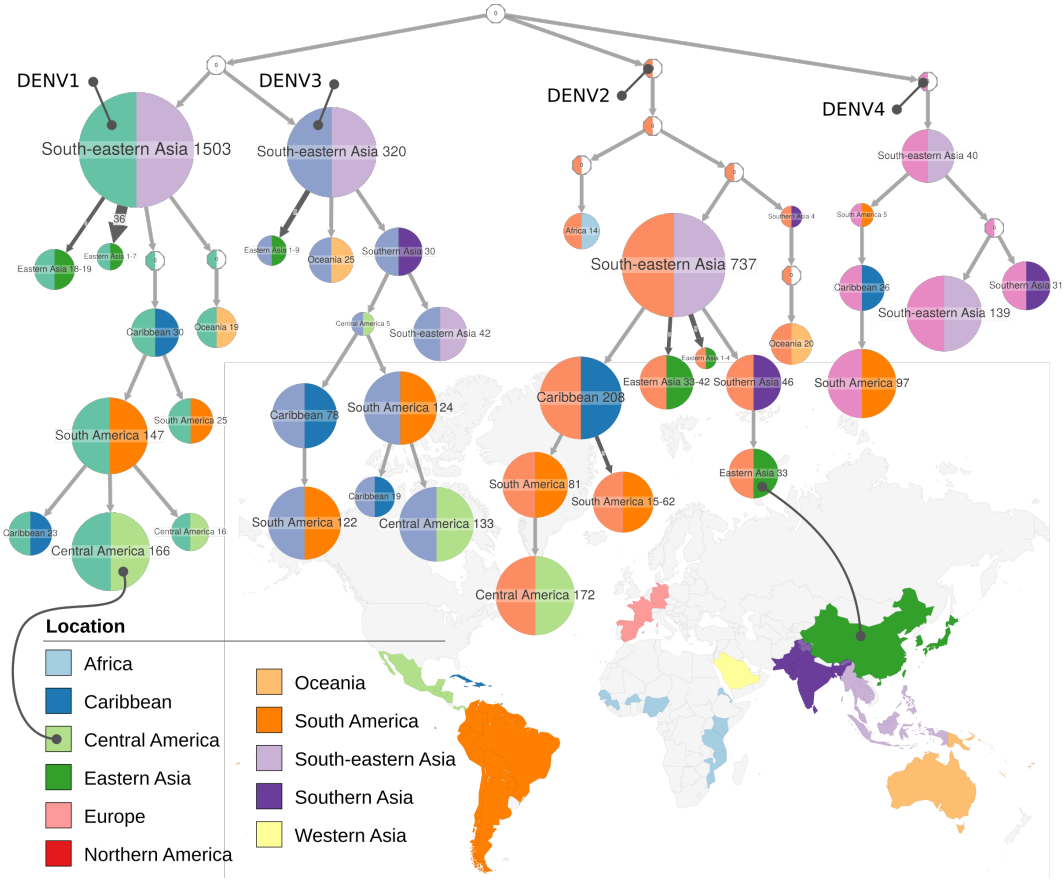

Figure 6 Ancestral location scenario reconstructed on the full DENV tree (large data set) using PastML (MPPA+F81). Locations are colour-coded in the right halves of the nodes and shown as labels, serotypes are colour-coded in the left halves of the nodes, as follows (from left to right): DENV1 (green), DENV3 (blue), DENV2 (orange), DENV4 (pink). Tips representing less than 14 sequences are not shown.

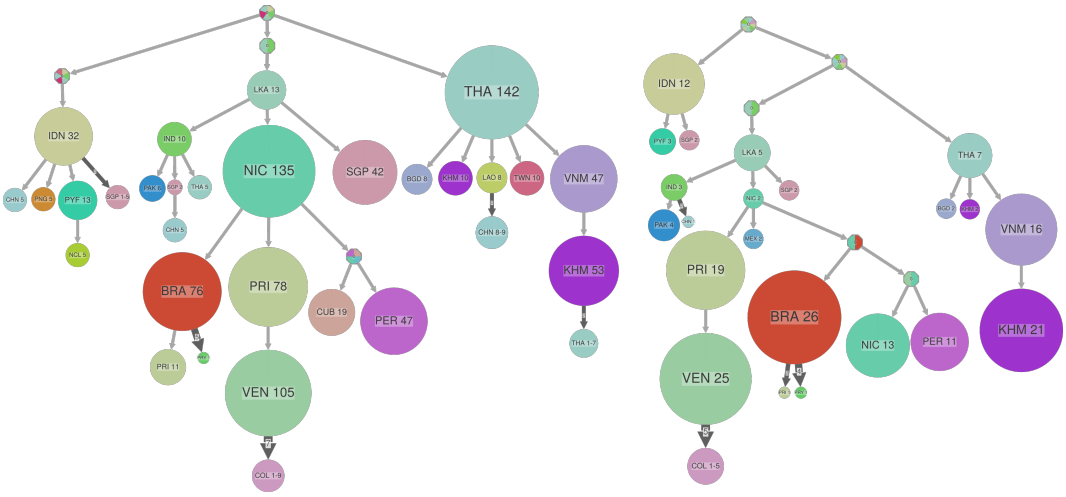

Figure 7 Ancestral country reconstructed on the DENV3 subtree with PastML (MPPA+F81) for the large data set (left) and the small one (right). Countries are colour-coded and shown as labels. Tips representing less than 5 (2) sequences are not shown for the large (small) data set. ACR for the two trees are compatible, additional sequences present in the larger data set allowed to resolve some of the nodes unresolved for the smaller one.

and Colombia for all serotypes. In our predictions using PastML we also observe this pattern, e.g. for DENV3 we inferred 7 introductions from Venezuela to Colombia (meta-edge of size 7 connecting "Ven 105" cluster to "COL $1-9$" one in Figure 7).

Tan et al. (2018) performed a Bayesian analysis of DENV3 genotype III spread to Malaysia using 602 complete coding sequences and 972 E-gene sequences from 56 countries between 1966 and 2014, including the complete genome sequences of 21 newly sequenced Malaysian DENV3 genotype III isolates. They detected an introduction from Sri Lanka to Singapore and from there to Malaysia. Our data set does not contain Malaysian sequences, but the spread of DENV3 genotype III from Sri Lanka to Singapore is also inferred, as indicated by the arrow from cluster "LKA 13" to "SGP 42" in Figure 7.

2.4 Discussion

ML methods are fast and permit the analysis of a large number of sequences in a relatively short time (e.g. 1.5 days for a phylogeny reconstruction on a 5 132 full genome data set with IQ-TREE, 2 minutes of dating with LSD2, and 3.5 hours of ancestral country reconstruction with PastML, on a 12-core machine). Adding more sequence data is desirable as it generally helps to reduce bias and uncertainty (e.g. for phylogeny rooting or ACR).

Phylogeny reconstruction is robust when applied to data with homogeneous time scale (e.g. serotype-specific trees): the results obtained by different tools were highly similar. However care needs to be taken when reconstructing phylogenies with mixed time scale (e.g. deep intra-serotype branches versus much shorter inter-serotype ones in the case of dengue): one needs to verify the reconstruction results, for example, based on prior knowledge on expected tree topology (e.g. monophyletic genotypes within the dengue serotypes), or by comparing the results obtained with different tree reconstruction tools. Rooting and dating with dengue data proved to be a difficult task. Popular ML/distance dating methods (e.g. LSD2, TreeTime) are fast and able to deal with large phylogenies, but particular problems remained difficult to solve (rooting, relaxed molecular clock, outliers, noisy dates). The full

dengue data set did not have enough signal for confident dating of deep nodes, and even for some of the more recent, serotype-specific phylogenies, different results were obtained depending on the method used. Any additional input (such as an outgroup) could be of great help. Preliminary exploratory analyses are very important, e.g. checking temporal structure and using regression to select an appropriate model and to determine whether molecular clock-based dating is valid at all. There is still room for improvement of these fast dating approaches.

ACR and phylogeography on the other hand, showed a strong and robust signal. A fast ML method like PastML was able to analyse large data sets, and provide results that were robust against sampling variations and phylogenetic uncertainty (see results for DENV3 [Figure 7] and [Ishikawa et al. 2019]).

3 Bayesian phylodynamic inference

Within the field of pathogen phylodynamics, Bayesian inference through Markov chain Monte Carlo (MCMC) is a widely used framework owing its popularity to a wide range of available models and its accommodation of phylogenetic uncertainty in generating posterior distributions for all parameters (including the tree topology). In practice however, data set sizes limit the application of Bayesian phylodynamic inference much more than the approaches highlighted earlier in this chapter. Recent applications have involved over a thousand genomes (e.g. for Ebola virus, Dudas et al. (2017)), and one of the largest studies included about 4,000 influenza gene sequences (Bedford et al., 2015). For the latter, strong temporal structure and phylogenetic resolution aided integrating over all plausible evolutionary histories. Many different approaches to confront the computational limitations are currently in development, focusing on different aspects of Bayesian inference.

Typically, MCMC algorithms may suffer from two problems that hamper performance: slow convergence and poor mixing (Nascimento et al., 2017). To remedy this, Bayesian inference often tries to employ a(n approximate) maximum-likelihood tree – which can be quickly estimated (see previous sections) – to yield a better-than-random starting location in tree space to initiate its search. While this aids the search in discrete tree space, it is important to note that other parameters involved in the phylodynamic model are not subject to such a pre-optimisation step and convergence may still take non-negligible time. The continuous need to further optimize MCMC integration is the focus of many current developments in the field.

A related interesting avenue of research is trying to deal with the problem of continuously accumulating data during an ongoing epidemic. The generation of additional sequence data requires an update of previously obtained results, which is typically done with a complete re-evaluation of the integrated Bayesian inference estimation procedure. Such a procedure renders Bayesian approaches costly to maintain an up-to-date estimate of the phylogenetic posterior distribution and this has motivated initial work on "online" Bayesian phylogenetic inference methodology, which can update an existing posterior with new sequences (Dinh et al., 2018; Gill et al., 2020).

While efficiently achieving convergence is one important aspect of the Bayesian challenge, mixing efficiency – which refers to how efficiently the chain samples from the posterior after it has converged on the posterior distribution (Nascimento et al., 2017) – is also of critical importance. In practice, mixing efficiency is frequently assessed by determining the degree of autocorrelation in the MCMC sample, with high autocorrelation reflecting a poor sample to characterise the posterior. If the Markov chain can be made more efficient in

sampling from the posterior, a relatively shorter chain may provide an acceptable estimate of the parameters of interest. Both the model parameterization (and prior specification) and the transition kernels acting upon the model's parameters can have a great effect on mixing efficiency. To deal with the issue of mixing among the potentially vast collection of parameters stemming from different models, adaptive MCMC approaches can potentially increase sampling efficiency for many continuous parameters simultaneously (see Section 3.2).

While convergence and mixing issues will impact the number of MCMC iterations needed to appropriately sample from the posterior, overall computation time will also be determined by the time it takes to evaluate the joint posterior density at each step. The same densities need to be estimated repeatedly and there are no shortcuts to evaluate each observed data likelihood and prior density, no matter which software framework or computational library is being employed. However, developments in multi-core computational hardware – both in the area of traditional central processing units (CPUs) but also of graphical processing units (GPUs) – offer increasing capabilities to compute those densities more efficiently by leveraging high-performance computational libraries. Such libraries provide access for multiple inference software packages to the underlying state-of-the-art hardware, allowing each inference program to avoid implementing low-level access to such hardware. In Chapter 5.4 (Ayres et al. 2020), we describe the BEAGLE high-performance computational library – which is used by multiple phylogenetic inference software packages – to illustrate how such performance increases in computing observed data likelihoods are brought about. All of the approaches described here that deal with Bayesian phylodynamic inference are available through BEAST v1.10 (Suchard et al., 2018), which now by default requires the BEAGLE library to run (Ayres et al., 2019).

3.1 Bayesian phylodynamic inference using BEAST 1.10

While many software packages are available today that focus on phylogenetic and even phylogenomic inference, BEAST (Bayesian Evolutionary Analysis by Sampling Trees; Suchard et al. 2018) unifies molecular phylogenetic reconstruction with complex discrete and continuous trait evolution and allows modelling parameters of interest as a function of external covariate data. In particular, the use of location data associated with genetic sequences has been popularised through Bayesian inference approaches for ancestral location reconstruction, allowing to track the spread of an organism through geographic space.

Since its inception, BEAST has focused on estimating time-scaled (rooted) trees from genetic data. This can be done through the classical use of external calibration information, such as information from the fossil record or through studies on plate tectonics as well as the use of the sampling times of sequences from 'measurably evolving populations' (MEPs, Drummond et al. 2003b). MEPs are characterized by either fast substitution rates and sample availability over a limited time-scale (e.g. for rapidly evolving RNA viruses) or slower substitution rates but with much longer sampling time scales (e.g. ancient DNA, Drummond et al. 2003b). Prior distributions over time-measured genealogies such as the coalescent allow inferring temporal changes in population size. To this end, BEAST provides a wide range of molecular clock models and coalescent models, alongside the traditional range of nucleotide, codon and amino acid substitution models.

In recent years, BEAST has increasingly focused on the analysis of rapidly evolving pathogens and their evolutionary and epidemiological dynamics. Given the rapid growth of pathogen genome sequencing as part of public health responses to infectious diseases, recent areas of interest for further development of BEAST are the exploitation of increasingly parallel computing architectures to decrease time to results, such as multi-core CPU and GPU

hardware, both through the development of novel estimation procedures (such as adaptive MCMC; see Section 3.2) and a much closer integration with the BEAGLE high-performance computational library (see Chapter 5.4 [Ayres et al. 2020] for more information).

3.2 Adaptive MCMC

Novel sequencing technologies are delivering increasingly larger numbers of genome sequences, which may amount to thousands of sequences containing hundreds or even thousands of genes. Viruses with limited genome sizes are typically characterised by a restricted number of genes. In a viral outbreak setting, recent years have witnessed the deployment of portably sequencing technologies (Quick et al., 2016), for example to analyse a large-scale Ebola virus outbreak in West Africa (Dudas et al., 2017) and the Zika epidemic in the Americas (Faria et al., 2017). The increasing reliability and accuracy of portable genome sequencing technologies have now turned them into an important instrument in shedding light on unfolding epidemics.

Large viral data sets can typically be analysed using evolutionary models that take into account the structural properties of those alignments by employing gene-specific and/or codon position-specific partitioning schemes. Such modelling approaches combine the benefits of more accurately modelling the underlying evolutionary processes with increased computational performance, for example by using different nucleotide substitution models per codon position rather than computationally demanding full codon models (Shapiro et al., 2006). However, as partitioning strategies involve estimating conditionally independent models of molecular evolution for different genes and different positions within those genes, they require a large number of evolutionary parameters to be estimated, which may pose difficulties for traditional Bayesian inference approaches. Given the predominance of single-component Metropolis-Hastings approaches, a parameter estimation strategy which proposes a new value for one single parameter at a time (Gilks et al., 1996), estimating large numbers of parameters that are spread across multiple data partitions is associated with a considerable computational burden in Bayesian phylogenetic inference (see Figure 8).

Baele et al. (2017a) have introduced a transition kernel into BEAST (Suchard et al., 2018) based on the adaptive Metropolis (AM) algorithm of Haario et al. (2001) and Roberts and Rosenthal (2009) that continuously adapts its d-dimensional proposal distribution to better match the target distribution and hence learn better parameter values as the analysis progresses. Computing the covariance of the proposal distribution using all of the previous states, the AM algorithm is in turn based on the classical random walk Metropolis algorithm (Metropolis et al., 1953) and earlier work by Haario et al. (1999) that entertains a Gaussian proposal distribution centered on the chain's current state but with the covariance being calculated from a fixed number of previous states. The use of simple recursion formulae to update the covariances ensures that the computational cost associated with the AM algorithm remains constant as the analysis progresses (Haario et al., 2001).

Apart from updating the proposal distribution by using currently available knowledge about the target distribution, the construction of the AM algorithm is identical to the usual random walk Metropolis-based chain. Suppose that at iteration n we have previously sampled the states $X_0, X_1, \ldots, X_{n-1}$, where X_0 is the initial state (typically sampled at random from its prior distribution). A candidate point Y is then sampled from the (asymptotically symmetric) normal proposal distribution, given at iteration n by $Q_n(x, \cdot) = N(x, (C_d)^2 I_d/d)$ for $n \leq C_0$, while for $n > C_0$:

$$Q_n(x, \cdot) = (1 - \beta)N(x, \Sigma_n/d) + \beta N(x, (C_d)^2 I_d/d), \tag{3}$$

where Σ_n is the current empirical estimate of the covariance structure of the target distribution

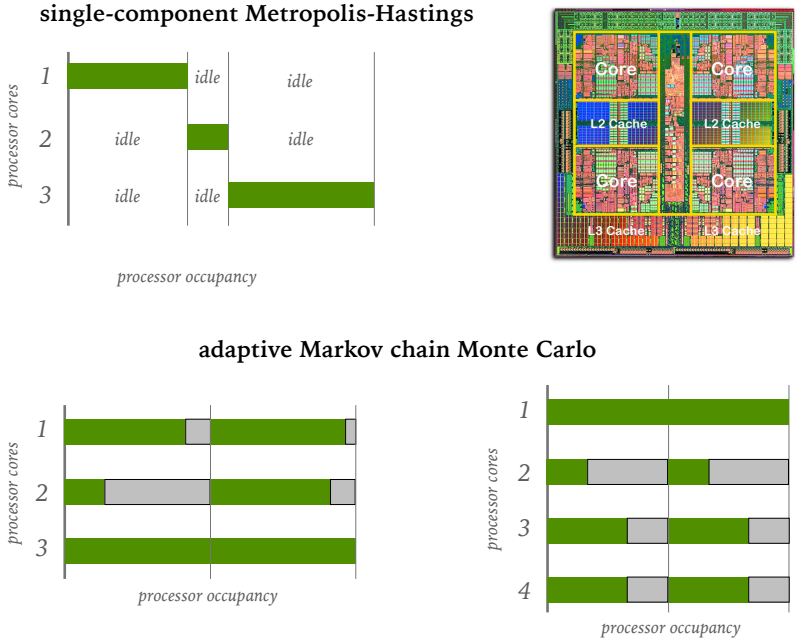

Figure 8 Conceptual visualisation of the potential benefits of an adaptive MCMC algorithm over single-component Metropolis-Hastings when assuming a codon partitioned substitution model (green bars indicate a processor's occupancy while computing a specific likelihood). In Bayesian phylogenetics, the common practice of updating a single parameter at a time typically only requires a single CPU core in order to recompute the observed data likelihood, leaving many CPU cores idle and hence underusing the computational capacity of multi-core CPU architectures. In many cases, the likelihood for the second codon position is quickly computed given the low number of unique site patterns at this position, whereas the third codon position typically accumulates more substitutions resulting in a more demanding likelihood computation. Adaptive MCMC allows updating a collection of continuous parameters simultaneously, putting many cores to work in a parallel fashion to compute the various (codon) partitions. Note that different computational demands between data partitions will lead to waiting times (shown in grey) which hamper performance, a problem that can be tackled by splitting the computation of a particular partition over multiple processor cores. In this example, a fourth processor core is not being used and hence partially offloading the computation of the third codon position likelihood to the remaining free processor core reduces waiting time and increases overall throughput. Quad-Core AMD Opteron processor silicon die shown courtesy of Advanced Micro Devices, Inc. (AMD), obtained from Wikimedia Commons.

based on the run so far, I_d is the d-dimensional identity matrix and β is a small positive constant. The (order of) magnitude for the parameters C_0 and C_d is not easily determined however and, given the only recent introduction of such adaptive transition kernels in Bayesian phylogenetics, is subject to further research. The candidate point Y proposed by the transition kernel is accepted with probability

$$\alpha(X_{n-1}, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X_{n-1})}\right), \tag{4}$$

in which case we set $X_n = Y$, and otherwise $X_n = X_{n-1}$, where $\pi(\cdot)$ is the target distribution.

Note that the chosen probability for the acceptance resembles the familiar acceptance probability of the Metropolis algorithm, but the corresponding stochastic chain is no longer Markovian (Haario et al., 2001). It is known that adaptive MCMC algorithms will not always preserve stationarity of $\pi(\cdot)$ (Roberts and Rosenthal, 2009). However, having proven the ergodicity of adaptive MCMC under certain conditions (Roberts and Rosenthal, 2007), the AM algorithm above will indeed converge to $\pi(\cdot)$ and satisfy the Weak Law of Large Numbers (WLLN), even though it is not Markovian.

The use of a d-dimensional proposal distribution through such an adaptive transition kernel can correspond in its simplest (or standard) use case to updating one parameter in each of d sequence data partitions (e.g. when d genes are present in the multiple sequence alignment). The simultaneous proposal of updated parameter values for all d parameters will in such a case trigger d likelihood calculations, which can be performed in parallel as there is no dependency of the sequence data likelihoods on one another. The current surge in development and availability of multi-core processors, both in the central processing unit (CPU) and graphics processing unit (GPU) processor markets, provides an excellent opportunity to perform such massively parallel computations. Multi-core CPU systems allow evaluating multiple data likelihoods simultaneously on a single processor, employing each processor core to compute the likelihood of a given data partition. GPU cards aimed at the scientific computing market benefit from a different approach towards likelihood computation, with an implementation that is agnostic of the concept of tree topologies (Suchard and Rambaut, 2009). High-performance computational libraries such as BEAGLE (Ayres et al., 2019) provide an extended API and library to support concurrent computation, not only on CPU but also on GPU, of independent partial likelihoods, for increased performance of analyses with greater flexibility of data partitioning (see the separate BEAGLE chapter for more information).

3.3 Discrete phylogeographic inference

Apart from the availability of many nucleotide data partitions in modern-day data sets that need to be analysed using phylogenetic approaches, an increasing number of trait data partitions are being included into phylodynamic analyses (for an overview, see Baele et al. 2017b). Given the specific properties of the data set we analyse in this chapter, we focus here on a discussion of discrete trait analysis in combination with sequence data. As in the previous sections in this chapter, we consider sampling location as a discrete trait to perform phylogeographic analyses in conjunction with approaches to visualize the reconstructed viral spread over time and space.

Discrete phylogeographic inference has witnessed a surge in popularity after its development and inclusion into the BEAST software package (Lemey et al., 2009; Suchard et al., 2018). A popular approach for discrete trait modeling is to take guidance from standard phylogenetics and borrow the process of exchange between sequence character states as a

generic model for how (discrete) traits evolve over the branches of a phylogeny, as described in the previous section.

With most general models, it requires estimating large migration rate matrices and therefore significant computer power and typically makes the resulting analysis no longer suitable even for multi-core CPU systems and requires the use of state-of-the-art GPUs (Suchard and Rambaut, 2009).

Informing the instantaneous rate parameters of a discrete trait model can be aided by providing predictors or covariates that inform the transition rates between discrete states (see Section 3.4). Other approaches to protect against over-parameterization include prior specification, which can for example consist of proposing higher rates of diffusion between nearby locations a priori. In addition, Bayesian stochastic search for variable selection (BSSVS) can be adopted to reduce the number of rate parameters to a restricted set that provides the most adequate parsimonious description of the diffusion process (Lemey et al., 2009). Variable selection and informative prior specification both increase statistical efficiency, which becomes even more important when drawing inference from sparse data – such as a single column of discrete traits – under more complex models, for example, assuming asymmetric transition rates between each pair of discretized trait values (Edwards et al., 2011).

3.4 Incorporating potential predictors of spatial spread

As mentioned in the previous section, the estimation of a potentially large number of transition rates in a discrete trait model can be informed by incorporating predictors that may play an important role in the underlying transition process between trait states. Such predictors can be integrated using a generalized linear model (GLM) formulation for the transition rates, a common approach in statistics that allows to model the linear relationship between a dependent variable (in this case the transition rates) and a collection of independent variables. To identify the relevant subset of predictors out of a number of explanatory variables, the GLM model can be extended with a BSSVS procedure. To this end, the $(K-1) \times K$ parameters that model the instantaneous rates of spread between the K discrete locations $\Lambda_{ij} (\forall i \neq j)$ is modelled as a log linear function of the set of P predictors $(x_1, \ldots, x_P)$ so that

$$\log\Lambda_{ij} = \beta_1\delta_1 x_{i,j,1} + \beta_2\delta_2 x_{i,j,2} + \ldots + \beta_P\delta_P x_{i,j,P}, \tag{5}$$

where $(\beta_1, \ldots, \beta_P)$' represent the effective sizes (or coefficients) for the predictors, quantifying their contribution to Λ, and $(\delta_1, \ldots, \delta_P)$ are (0,1)-indicator variables that govern the inclusion or exclusion of the P predictors in the model. The incorporation of indicator variables allows to perform the BSSVS procedure, which involves letting the data decide whether or not the corresponding predictor provides a significant contribution to the model and hence should be kept as part of the model. In other words, when an indicator δ_p equals 1, then predictor x_p is included in the model, and assessing its effect size through β_p allows interpreting the direction and magnitude of that predictor's contribution. The average value of each indicator across iterations provides an estimate of the inclusion probability of a predictor and can be used to compute a Bayes factor expressing how much the data change our prior opinion about the inclusion of each predictor. Completing the model's specification is done by assuming a small prior probability on each predictor's inclusion that reflects a 50% prior probability on no predictors being included, but specifying equal prior probability on each predictor's inclusion and exclusion yields highly similar results. The Bayes factor

for a predictor BF_p is then calculated by dividing the posterior odds for the inclusion of a predictor with the corresponding prior odds

$$\mathrm{BF}_p = \frac{\mathrm{pp}_p}{1-\mathrm{pp}_p} / \frac{\mathrm{qp}_p}{1-\mathrm{qp}_p}, \tag{6}$$

where pp_p is the posterior probability that predictor p is included, in this case the posterior expectation of indicator δ_p, and qp_p is the prior probability that $\delta_p = 1$.

3.4.1 Predictor data

In order to offer an explanation for the geographic dispersal of emerging pathogens, different sources of information can be incorporated in the phylogeographic testing procedures. To illustrate the principle behind this approach, we have collected two data matrices that we use here as predictors for the geographic spread patterns of dengue. First, we consider air transportation data between the 64 countries from which the dengue sequences were obtained, thereby aggregating data from a passenger flux matrix that quantifies the number of passengers traveling between each pair of airports on a daily basis. We use a dataset provided by OAG (Official Airline Guide) Ltd. (`http://www.oag.com`), containing 4.092 airports and the number of seats on scheduled commercial flights between pairs of airports during the years 2004-2006. We take the number of seats on scheduled commercial flights from airport i to j to be proportional to the number of passengers traveling. Because passenger flux does not differ in a statistically significant manner from symmetry in the global air transportation network (Woolley-Meza et al., 2011), we consider flows that were symmetrized. These air transportation data were converted to "effective distances", which aim to reflect the idea that a small fraction of traffic is effectively equivalent to a large distance, and vice versa (Brockmann and Helbing, 2013). A second predictor consists of the average distances between the locations in our data set. Specifically, we considered the average great-circle distance between two locations based on the pairwise distances between all pairs of airports from the two locations. While further predictors can be added into the GLM, we use these two predictors described here to showcase their use and interpretation.

3.4.2 Results

In our BEAST analysis – comprising 200 million iterations – both predictors were consistently included in the GLM model (the associated indicator variables remained 1, so $E[\delta] = 1$), indicating that our data support an important contribution of these two predictors to the geographic spread of dengue. We also find effect sizes that are centered around negative values, indicating an inverse relationship between the predictors and the instantaneous rates of dengue spread between locations, and credible intervals that exclude 0 ($\beta = -0.71[-0.83, -0.59]$ for air flux and $\beta = -0.49[-0.60, -0.37]$ for geographic distance). In other words, the closer two locations are to one another and the smaller their 'effective distances' are (or the more frequent air travel between them), the more intense the spread of dengue between those two locations becomes. In summary, using a framework that estimates the migration history of dengue while simultaneously testing and quantifying potential predictive variables of spatial spread, we show that the global dynamics of dengue are potentially driven by a combination of air passenger flow and geographic distance between the locations from which these dengue samples originated. However, we caution against drawing strong conclusions from this analysis as only two predictors were tested and considerable bias may exist in

the sampling by country. We note that in addition to offering a phylogeographic testing approach, the GLM parameterisation of discrete trait diffusion also considerably reduces the number of parameters to be estimated. While the standard CTMC model has transition rate parameters that scale quadratically with the number of states, the GLM-diffusion parameters scale linearly with the number of predictors. Although the parameters remain restricted in this way, the likelihood calculation for high state spaces remains associated with a large computational burden. However, this can to a large extent be mitigated by parallelisation as discussed in the BEAGLE chapter.

3.5 Tree visualisation using FigTree

We first focus on presenting the inferred time-stamped phylogenetic tree as one of the key outcomes of our joint inference of sequence and trait data, which includes the parameterization of the geographic spread between location as a function of our predictor data. To this end, we employ FigTree, a popular cross-platform graphical tree display software package. Although it can be used as a general tree visualisation tool, it is particularly powerful to display annotated trees produced by BEAST. In order to construct a maximum clade credibility (MCC) tree, i.e. the single tree in the posterior sample with the largest sum of posterior probabilities across its constituent bifurcations, we first use TreeAnnotator (which is part of the BEAST software package) to summarize the trees from the posterior distribution collected during an analysis that comprised 200 million iterations (after removing the necessary burn-in). The resulting time-stamped MCC tree can then be visualised in different manners and with different annotations in FigTree, as illustrated in Figure 9, where different colours have been used for the clusters of sequences corresponding to the different dengue subtypes.

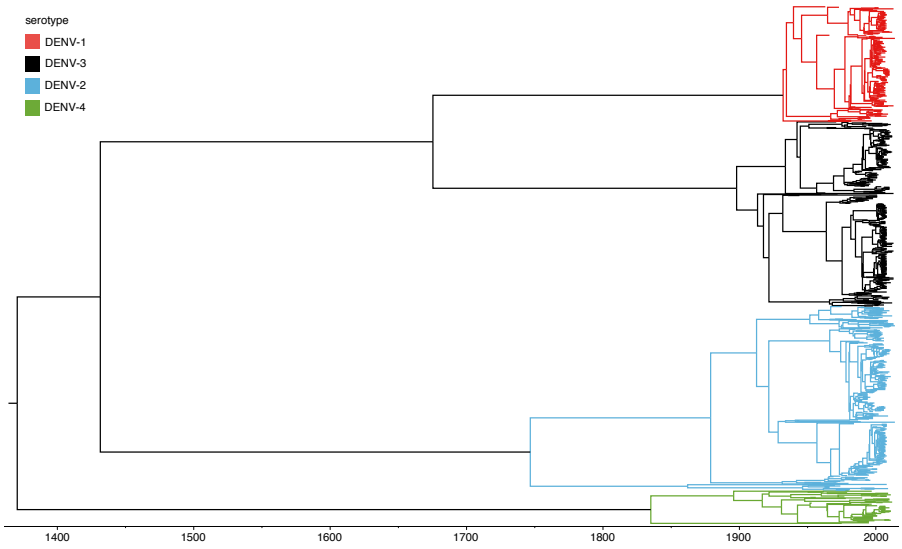

Figure 9 Time-stamped tree visualisation according to the default view in FigTree, with clusters coloured according to the corresponding dengue serotype.

Oftentimes, discrete ancestral trait reconstructions on a tree are represented by branch colour annotations. Such a visualisation allows interpreting the time and location of origin as well as the introduction into different locations at different points in history of dengue. Given that the different dengue serotypes diverged centuries ago (see Figure 9), we here focus on showing the phylogeographic reconstructions of the different serotypes in Figures

10, 11, 12 and 13 (all trees are MCC trees). In doing so, we have "ladderized" the trees, meaning that the nodes have been sorted on one level by the count of their subnodes (on all levels under the node), for easier visualisation and interpretation of the trees.

Dengue virus comprises four serotypes that co-circulate in tropical regions and the relationship between the various dengue serotypes depicted in Figures 9 is well known. It has been hypothesised that antibody-dependent enhancement (ADE) may explain this particular shape of the dengue virus phylogeny, in which the four serotypes are phylogenetically equidistant (see e.g. Grenfell et al. 2004). Natural selection may favour this level of antigenic dissimilarity, as cross-protective antibodies would neutralize more similar strains, whereas more divergent strains would not stimulate ADE. It has also been suggested that independent cross-species (monkey-human) transmission might be able to explain the dengue phylogeny if the serotypes predominate in different geographic areas, followed by later mixing (Holmes and Twiddy, 2003). However, this is difficult to determine based on the visualisation of the trees in Figures 10-13, and a projection onto geographic space would be more useful in such a case, which we discuss in the following section.

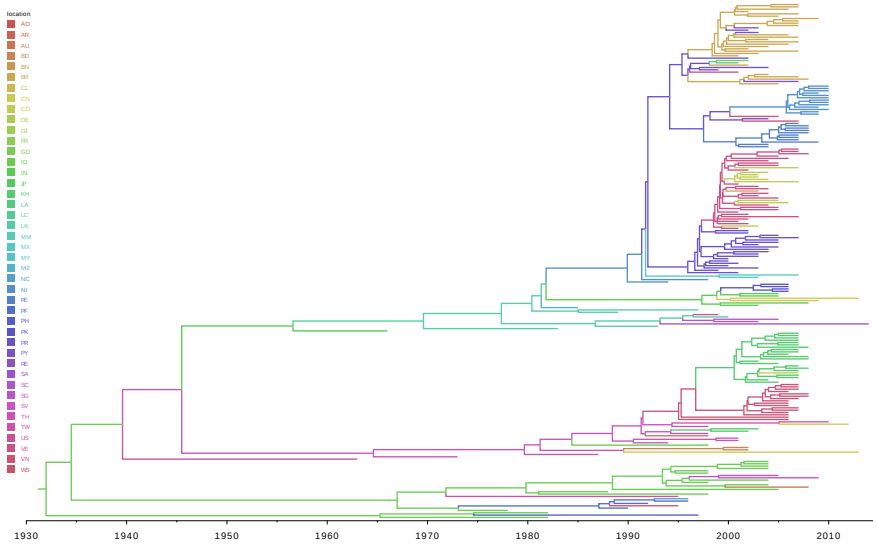

Figure 10 Time-stamped tree visualisation of serotype 1 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 1 is estimated to have originated in Indonesia in the 1930s.

3.6 Visualisation and animation using spreaD3

In order to visualise the estimated spread over time in a geographically explicit way, each discrete location considered in the ancestral reconstruction needs to be complemented with a set of GPS coordinates. We here resort to the centroid of each country, which can be easily retrieved online (an alternative option could be to use the GPS coordinates for the capital of the country). These latitude and longitude data can be readily loaded in spreaD3 (Bielejec et al., 2016), along with a geographic map in GeoJSON format containing the regions of interest. SpreaD3 will then generate, by using the sampling times and locations, and the estimated divergence times and ancestral location reconstruction, an animated visualisation of the viral spread over time, starting from the estimated time to most recent common

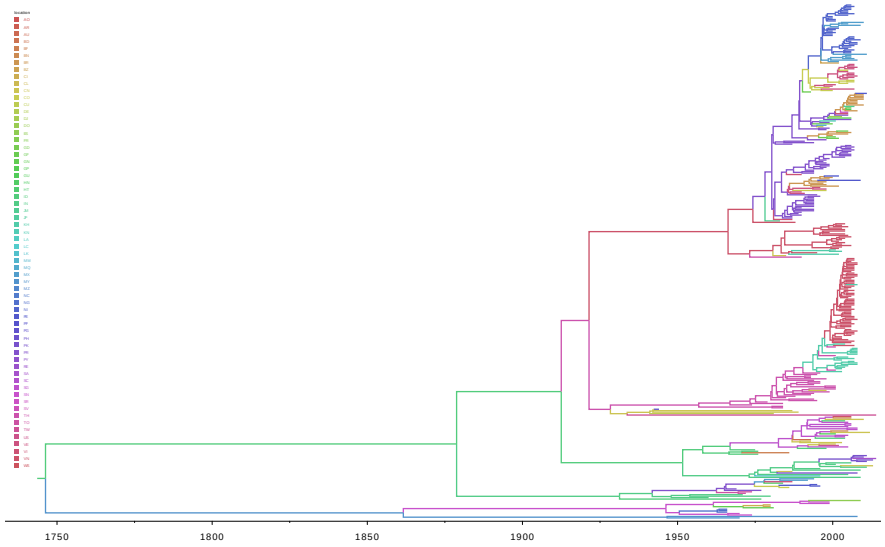

Figure 11 Time-stamped tree visualisation of serotype 2 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 2 is estimated to have originated in Indonesia in the first half of the 18th century.

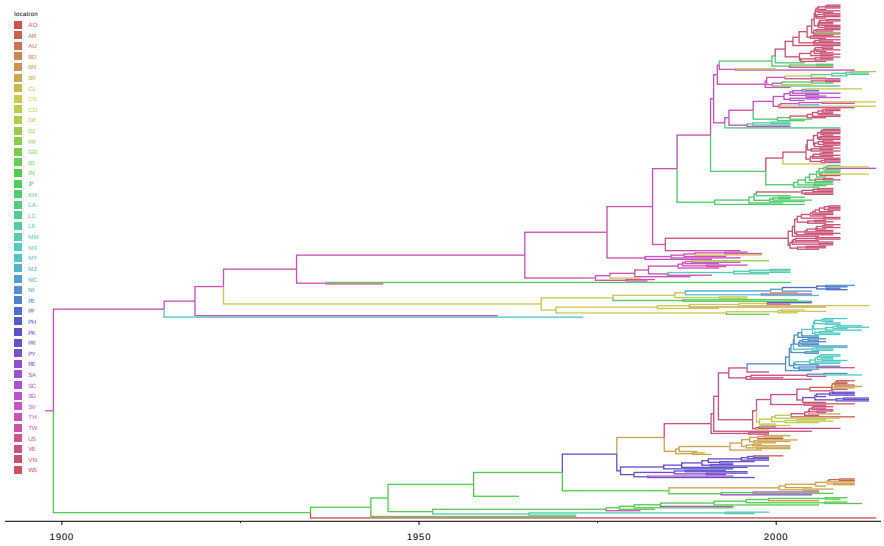

Figure 12 Time-stamped tree visualisation of serotype 3 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 3 is estimated to have originated in Thailand at the end of the 19th century.

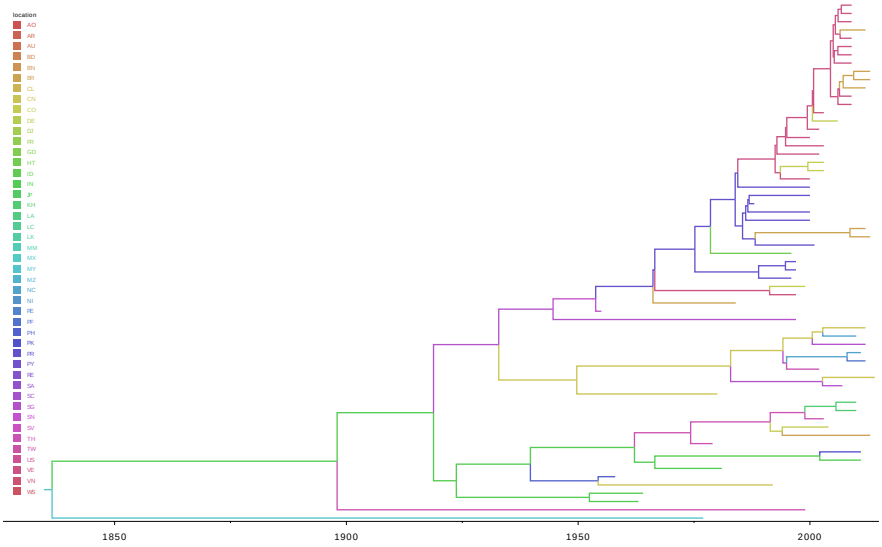

Figure 13 Time-stamped tree visualisation of serotype 4 – based on the full MCC tree (see Figure 9) – with each branch being coloured according to the location state at its descendant node. Based on our Bayesian inference, serotype 4 is estimated to have originated in Myanmar in the first half of the 19th century.

ancestor of the data being analysed, until the most recent sampling date. This is shown in Figures 14-17, where – for each serotype – we present two snapshots of the animation in progress: one taken early on in the epidemic, when dengue still seems to be restricted to South East Asia, and another taken at the start of 2014 (i.e. the most recent sampling date in our dengue data set), where the different dengue serotypes have spread throughout most tropical regions on earth.

4 Comparison between ML and Bayesian estimates

In the previous sections, we have described large whole-genome virus data set analyses using ML and Bayesian approaches, and discussed their inference results on an example of Dengue virus. To assess how their results compare, we looked at the three aspects that correspond to the three ML pipeline steps: (1) the reconstructed tree topologies; (2) the node dates of the time-scaled trees; (3) and the geographic predictions. To make the ML analysis performed on a larger 5 132 complete genome data set comparable with the Bayesian and ML analyses on the small dataset, we pruned the resulting phylogenies to keep only the common tips. We also calculated a consensus topology by collapsing the branches corresponding to internal nodes that were not common to all of the three (pruned) trees (ML (small), ML (large) and Bayesian): each internal node was uniquely identified by the set of tips in its subtree. For the Bayesian case we used the MCC tree.

Tree topologies.

To compare the tree topologies we calculated pairwise normalised quartet distances (Estabrook et al., 1985) (with tqDist [Sand et al. 2014]), where 0.0 indicates identical trees and 1.0 corresponds to trees that have no quartet in common. The topologies obtained using the various inference methods were very similar: the closest ones were reconstructed on the same

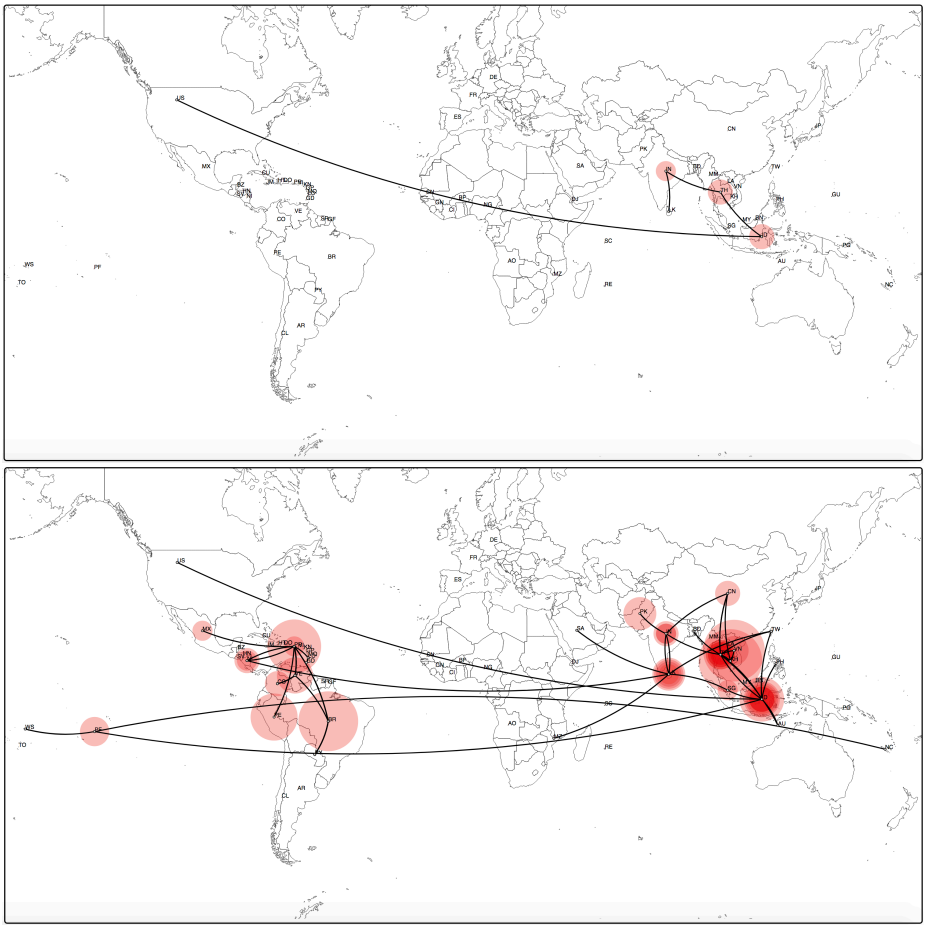

Figure 14 The geographic dispersal over time of dengue serotype 1, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 1 at the start of 1970, when the virus is estimated to have originated in Indonesia, from where it first spread to Thailand and the United States. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

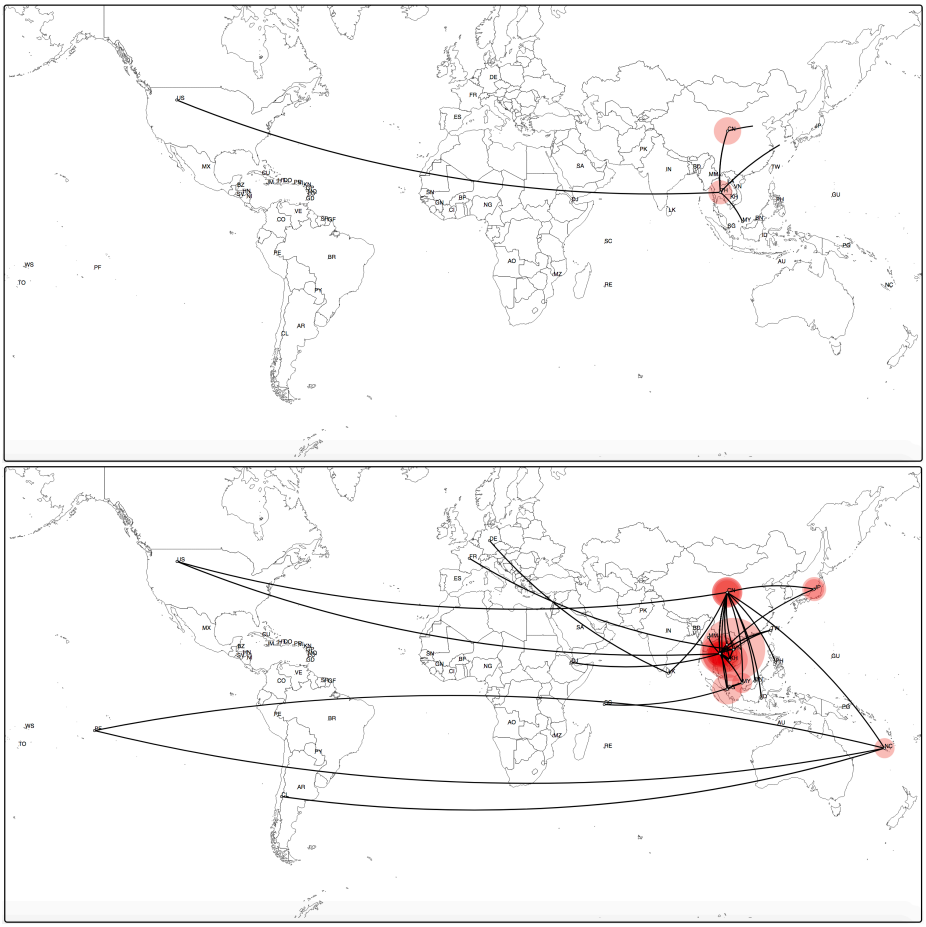

Figure 15 The geographic dispersal over time of dengue serotype 2, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 2 at the start of 1980, when the virus is estimated to have originated in Thailand, from where it first spread to China, Myanmar and the United States. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

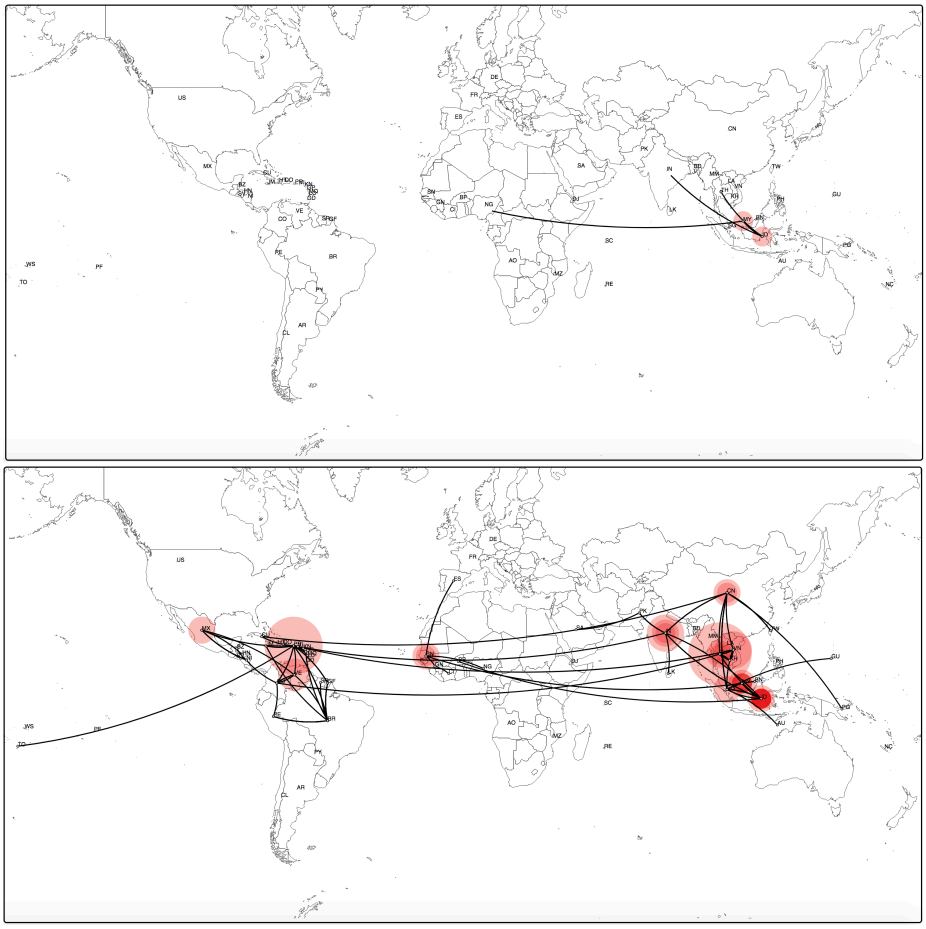

Figure 16 The geographic dispersal over time of dengue serotype 3, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 3 at the end of the 1920s, when the virus is estimated to have originated in Indonesia, from where it spread to Thailand, Myanmar, India and the African continent. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

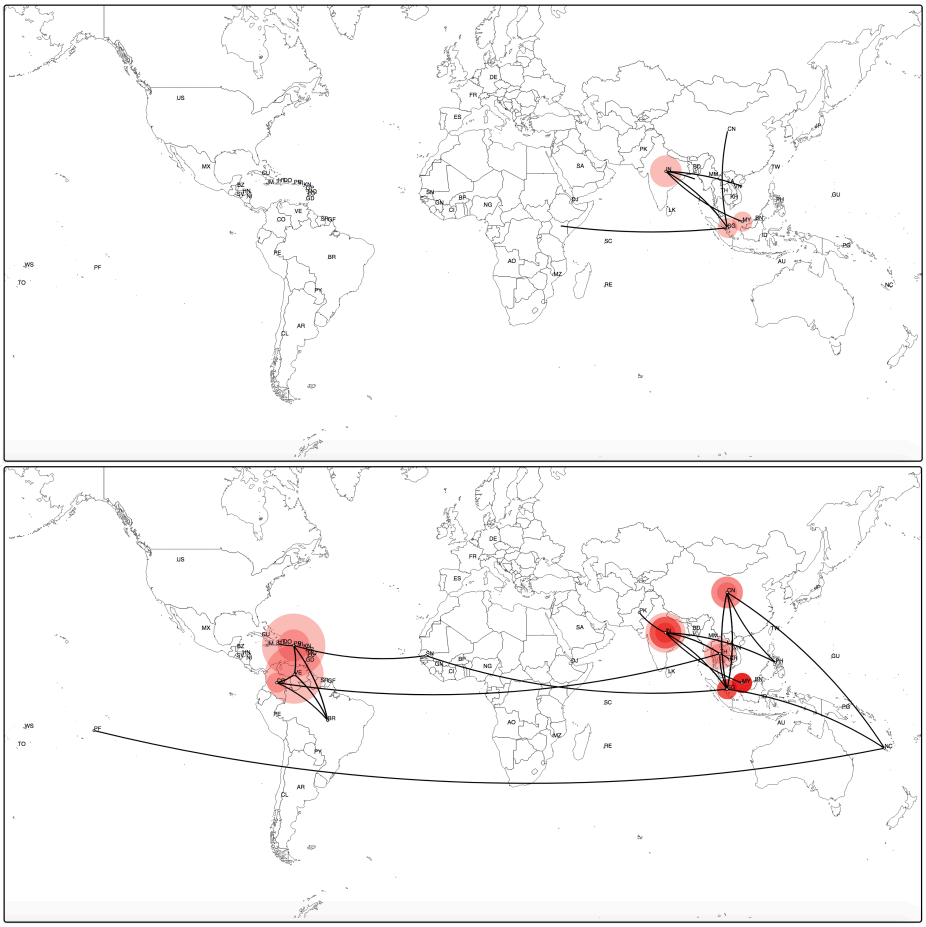

Figure 17 The geographic dispersal over time of dengue serotype 4, visualised using spreaD3 (Bielejec et al., 2016). The top figure shows the spread of dengue serotype 4 at the start of 1950, when the virus is estimated to have originated in Myanmar, from where it spread to India, Myanmar, Singapore and the African continent. The bottom figure shows the "current" spread (i.e. when the final sample in our data set was taken, at the start of 2014).

(small) data set (normalised quartet distance of 0.003), the distance between the two ML trees was 0.009, and 0.011 between the ML tree reconstructed on the large data set and the tree obtained through Bayesian inference.

Time-scaled trees.

We compared the predicted dates of the internal nodes present in the consensus topology. The root (common ancestor of four dengue serotypes) dates were very different and hence strongly depended on the inference used. A root date was estimated to be -2377 $[-2885; -1642]$ for the small ML tree (LSD2) and -54 $[-132; 56]$ for the large one; the Bayesian analysis on the other hand yielded 1370 as the median root date, with a 95% HPD (i.e. $[1215; 1475]$) that can be considered relatively narrow compared to the CIs obtained by various ML tools. These results show that dating relatively deep divergence in a phylogeny can yield drastically different results depending on the methodology used, when all methods rely solely on the sequences and their sampling times, which are all very recent (and given the oldest estimated ages of the full tree may almost be considered to be contemporaneous). This could be predicted from the RTT plot in Figure 4.

We observed considerably smaller root date differences for the serotype subtrees, for which the results are plotted in Figure 18. The time-scaled trees obtained on the same, small, data set are closer to each other than those estimated on different data sets for all the serotypes except for DENV4. In the case of DENV4, the ML time-scaled tree reconstructed on the large data set and the tree from Bayesian inference are very similar. As was already apparent from the full tree root date comparison between the different frameworks, node date estimates obtained through a fully Bayesian inference are consistently more recent than those generated by ML, regardless of the data set size used for the latter. Irrespective of differences in methodology, we stress that estimating relatively deep divergence times based on the divergence accumulating between recent tips can be misleading. Specifically, purifying selection has been shown to lead to severe underestimates of the ancient age of viral lineages (Wertheim and Kosakovsky Pond, 2011). Recent advances in molecular clock modelling attempt to address this (Membrebe et al., 2019).

Geographic reconstruction.

For each internal node in the consensus tree we compared its predicted country and probabilities. As the MPPA method used for ML phylogeographic reconstruction might predict multiple countries per node, we checked whether the modal Bayesian state was among them. For the majority of the nodes it was the case: 98% (91%) for ML ACR on the small (large) data set vs the Bayesian ACR. Less intersection with the large data set could be explained by the fact that it included more countries (83 versus 67): for the ACRs performed on the small data set the 16 unseen countries could never be reconstructed.

5 Conclusions

The combination of high-throughput experimental techniques and advanced methods that stem from physics, statistics and computer science allow to analyse increasingly large quantities of genomic data. ML and Bayesian phylogenomic tools can perform divergence time estimation and phylogeographic reconstruction of thousands of genome-scale virus sequences. However, care should be taken in choosing the data (e.g. removing erroneously annotated and poorly sequenced data that can bias the predictions), choosing correct tools

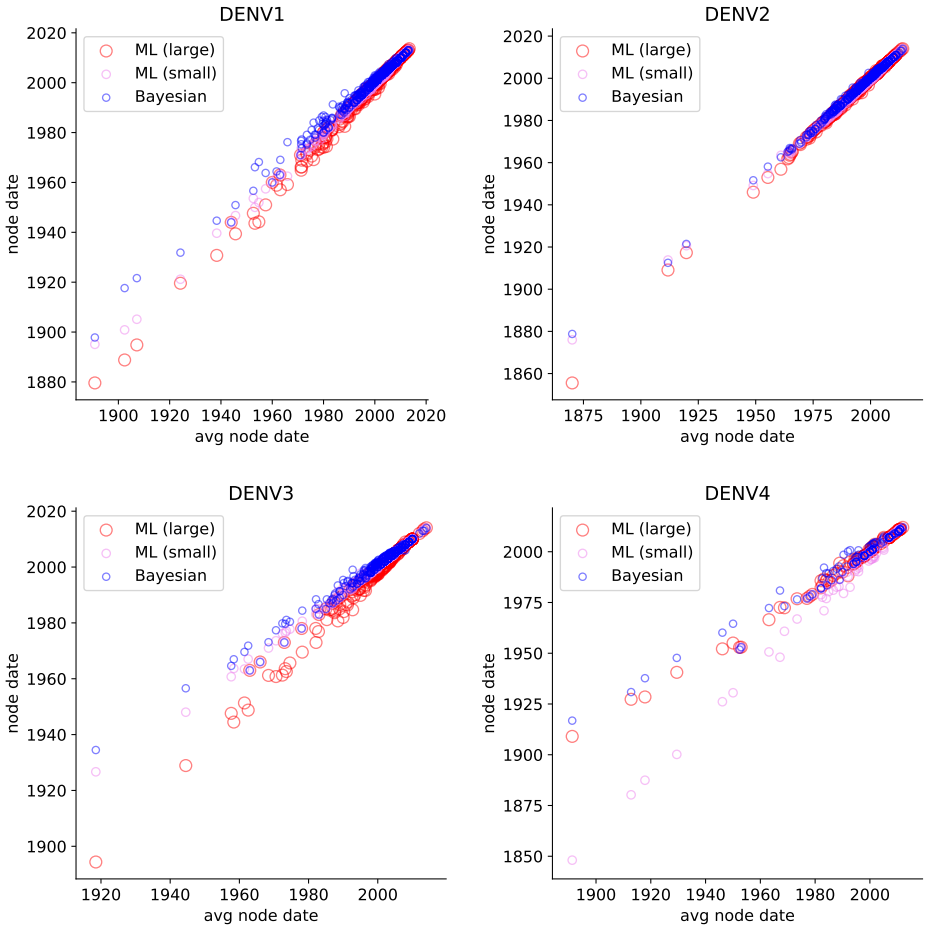

Figure 18 Common node dates for ML (small, dated with LSD2, pink), ML (large, dated with LSD2, red) and Bayesian (blue) trees by serotype (y axes) plotted against the node date averaged over the three trees (x axis): DENV1 (top left), DENV2 (top right), DENV3 (bottom left), and DENV4 (bottom right). Root nodes correspond to the left-most points.

(e.g. FastTree is a good tool for a quick preliminary phylogeny reconstruction but it is not very accurate) and correct configurations (e.g. priors for Bayesian analysis, relaxed vs strict molecular clock), checking the results at all stages of the analysis (e.g. correct tree root position), and comparing the predictions obtained by different methods.

We compared Bayesian and ML analyses on the example of Dengue virus data, and obtained results that are similar overall, but also showed many differences (especially in terms of time predictions). It is however difficult to assess which result is closer to biological truth. ML analysis is much faster to perform ($\sim$ 2 days on a 12-core machine for the large data set and $\sim 2,5$ hours for the small one) and can be therefore applied to larger data sets. However by performing the analysis step by step, it might accumulate error, so each intermediate result needs to be checked. When possible, it may prove useful to perform different analyses and compare the results. The DENV data sets we have used in our comparisons serve to illustrate the computational approaches. We acknowledge that considerable sampling bias may exist between countries, which complicates drawing reliable conclusions about patterns of spatial spread. Furthermore, we have not performed any recombination analyses. As pointed out in the introduction, such analyses should be part of phylogenomic studies of pathogens that may recombine, which is the case for DENV (Worobey et al., 1999).

We note an interesting convergence in the development of Bayesian and ML analysis frameworks. The ML tools for dating trees have evolved towards the inclusion of (1) relaxed molecular clocks, which have to a large extent been advanced in Bayesian frameworks, and (2) statistical tests to decide between strict and relaxed clock models, which also have received much attention in Bayesian frameworks. On the other hand, we illustrated some of the efforts to reduce the computational burden of Bayesian inference in order to decrease the large gap with ML approaches. However, accommodating phylogenetic uncertainty by averaging over all plausible evolutionary histories will always remain restrictive relative to ML estimation.

References

Allen, B. L. and Steel, M. (2001). Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees. *Annals of Combinatorics*, 5(1):1–15.

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A., and Suchard, M. A. (2019). BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.*, 68(6):1052–1061.

Ayres, D. L., Lemey, P., Baele, G., and Suchard, M. A. (2020). Beagle 3 high-performance computational library for phylogenetic inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.4, pages 5.4:1–5.4:9. No commercial publisher | Authors open access book.

Baele, G., Lemey, P., Rambaut, A., and Suchard, M. A. (2017a). Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. *Bioinformatics*, 33(12):1798–1805.

Baele, G., Suchard, M. A., Rambaut, A., and Lemey, P. (2017b). Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol.*, 66(1):e47–e65.

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., Daniels, R. S., Gunasekaran, C. P., Hurt, A. C., Kelso, A., Klimov, A., Lewis, N. S., Li, X., McCauley, J. W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D. J., Suchard, M. A., Tashiro, M., Wang, D., Xu, X., Lemey, P., and Russell, C. A. (2015). Global

circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523:217–220.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, 41(D1):D36–D42.

Bielejec, F., Baele, G., Vrancken, B., Suchard, M. A., Rambaut, A., and Lemey, P. (2016). Spread3: interactive visualisation of spatiotemporal history and trait evolutionary processes. *Mol. Biol. Evol.*, 33(8):2167–2169.

Blok, J. (1985). Genetic relationships of the dengue virus serotypes. *J. Gen. Virol.*, 66:1323–1325.

Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342:1337–1342.

Bromham, L. (2020). Substitution rate analysis and molecular evolution. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.4, pages 4.4:1–4.4:21. No commercial publisher | Authors open access book.

Bromham, L., Duchêne, S., Hua, X., Ritchie, A. M., Duchêne, D. A., and Ho, S. Y. W. (2018). Bayesian molecular dating: opening up the black box. *Biological Reviews*, 93(2):1165–1191.

Chernomor, O., Minh, B. Q., Forest, F., Klaere, S., Ingram, T., Henzinger, M., Haeseler, A., and Freckleton, R. (2015). Split diversity in constrained conservation prioritization using integer linear programming. *Methods in Ecology and Evolution*, 6(1):83–91.

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, 65(6):997–1008.

Chor, B. and Tuller, T. (2005). Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, 21(Suppl 1):i97–i106.

Collins, T. M., Wimberger, P. H., and Naylor, G. J. P. (1994). Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. *Systematic Biology*, 43(4):482.

Dinh, V., Darling, A. E., and Matsen IV, F. A. (2018). Online Bayesian phylogenetic inference: theoretical foundations via sequential monte carlo. *Syst. Biol.*, 67(3):503–517.

Drummond, A., Forsberg, R., and Rodrigo, A. G. (2001). The Inference of Stepwise Changes in Substitution Rates Using Serial Sequence Samples. *Molecular Biology and Evolution*, 18(7):1365–1371.

Drummond, A., Oliver G., Pybus, and Rambaut, A. (2003a). Inference of Viral Evolutionary Rates from Molecular Sequences. *Advances in Parasitology*, 54:331–358.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003b). Measurably evolving populations. *Trends Ecol. Evol.*, 18(9):481–488.

Duchêne, S., Di Giallonardo, F., and Holmes, E. C. (2016a). Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales. *Molecular Biology and Evolution*, 33(1):255–267.

Duchêne, S., Geoghegan, J. L., Holmes, E. C., and Ho, S. Y. (2016b). Estimating evolutionary rates using time-structured data: a general comparison of phylogenetic methods. *Bioinformatics*, 32(22):btw421.

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D'Ambrozio, J., Dellicour, S., Caro, A. D., Diclaro II, J. D., Durrafour, S., Elmore, M. J., Fakoli III, L. S., Faye, O., Gilbert, M. L., Gevao, S. M., Gire, S., Gladden-Young, A., Gnirke, A.,

Goba, A., Grant, D. S., Haagmans, B. L., Hiscox, J. A., Jah, U., Kargbo, B., Kugelman, J. R., Liu, D., Lu, J., Malboeuf, C. M., Mate, S., Matthews, D. A., Matranga, C. B., Meredith, L. W., Qu, J., Quick, J., Pas, S. D., Phan, M. V. T., Pollakis, G., Reusken, C. B., Sanchez-Lockhart, M., Schaffner, S. F., Schieffelin, J. S., Sealfon, R. S., Simon-Loriere, E., Smits, S. L., Stoecker, K., Thorne, L., Tobin, E. A., Vandi, M. A., Watson, S. J., West, K., Whitmer, S., Wiley, M. R., Winnicki, S. M., Wohl, S., Wölfel, R., Yozwiak, N. L., Andersen, K. G., Blyden, S. O., Bolay, F., Carroll, M. W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G. F., Garry, R. F., Goodfellow, I., Günther, S., Happi, C. T., Holmes, E. C., Keïta, S., Kellam, P., Koopmans, M. P. G., Kuhn, J. H., Loman, N. J., Magassouba, N., Naidoo, D., Nichol, S. T., Nyenswah, T., Palacios, G., Pybus, O. G., Sabeti, P. C., Sall, A., Ströher, U., Wurie, I., Suchard, M. A., Lemey, P., and Rambaut, A. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, 544:309–315.

Edwards, C. J., Suchard, M. A., Lemey, P., Welch, J. J., Barnes, I., Fulton, T. L., Barnett, R., O'Connell, T. C., Coxon, P., Monaghan, N., Valdiosera, C. E., Lorenzen, E. D., Willerslev, E., Baryshnikov, G. F., Rambaut, A., Thomas, M. G., Bradley, D. G., and Shapiro, B. (2011). Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr. Biol.*, 21(15):1251–1258.

Estabrook, G. F., McMorris, F. R., and Meacham, C. A. (1985). Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Systematic Biology*, 34(2):193–200.

Faria, N. R., Kraemer, M. U. G., Hill, S. C., Goes de Jesus, J., Aguiar, R. S., Iani, F. C. M., Xavier, J., Quick, J., du Plessis, L., Dellicour, S., Thézé, J., Carvalho, R. D. O., Baele, G., Wu, C.-H., Silveira, P. P., Arruda, M. B., Pereira, M. A., Pereira, G. C., Lourenço, J., Obolski, U., Abade, L., Vasylyeva, T. I., Giovanetti, M., Yi, D., Weiss, D. J., Wint, G. R. W., Shearer, F. M., Funk, S., Nikolay, B., Fonseca, V., Adelino, T. E. R., Oliveira, M. A. A., Silva, M. V. F., Sacchetto, L., Figueiredo, P. O., Rezende, I. M., Mello, E. M., Said, R. F. C., Santos, D. A., Ferraz, M. L., Brito, M. G., Santana, L. F., Menezes, M. T., Brindeiro, R. M., Tanuri, A., dos Santos, F. C. P., Cunha, M. S., Nogueira, J. S., Rocco, I. M., da Costa, A. C., Komninakis, S. C. V., Azevedo, V., Chieppe, A. O., Araujo, E. S. M., Mendonça, M. C. L., dos Santos, C. C., dos Santos, C. D., Mares-Guia, A. M., Nogueira, R. M. R., Sequeira, P. C., Abreu, R. G., Garcia, M. H. O., Abreu, A. L., Okumoto, O., Kroon, E. G., de Albuquerque, C. F. C., Lewandowski, K., Pullan, S. T., Carroll, M., de Oliveira, T., Sabino, E. C., Souza, R. P., Suchard, M. A., Lemey, P., Trindade, G. S., Drumond, B. P., Filippis, A. M. B., Loman, N. J., Cauchemez, S., Alcantara, L. C. J., and Pybus, O. G. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*, 361(6405):894–899.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G., and Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61.

Faria, R., Quick, J., Morales, I., Thézé, J., Jesus, J., Giovanetti, M., Kraemer, M. U. G., Hill, S. C., Black, A., da Costa, A. C., Franco, L. C., Silva, S. P., Wu, C.-H., Raghwani, J., Cauchemez, S., du Plessis, L., Verotti, M. P., de Oliveira, W. K., Carmo, E. H., Coelho, G. E., Santelli, A. C. F. S., Vinhal, L. C., Henriques, C. M., Simpson, J. T., Loose, M., Andersen, K. G., Grubaugh, N. D., Somasekar, S., Chiu, C. Y., Muñoz-Medina, J. E., Gonzalez-Bonilla, C. R., Arias, C. F., Lewis-Ximenez, L. L., Baylis, S., Chieppe, A. O., Aguiar, S. F., Fernandes, C. A., Lemos, P. S., Nascimento, B. L. S., Monteiro, H. A. O., Siqueira, I. C., de Queiroz, M. G., de Souza, T. R., Bezerra, J. F., Lemos, M. R., Pereira,

G. F., Loudal, D., Moura, L. C., Dhalia, R., França, R. F., Magalhães, T., Marques, E. T., Jaenisch, T., Wallau, G. L., de Lima, M. C., Nascimento, V., de Cerqueira, E. M., de Lima, M. M., Mascarenhas, D. L., Moura Neto, J. P., Levin, A. S., Tozetto-Mendoza, T. R., Fonseca, S. N., Mendes-Correa, M. C., Milagres, F., Segurado, A., Holmes, E. C., Rambaut, A., Bedford, T., Nunes, M. R. T., Sabino, E. C., Alcantara, L. C. J., Loman, N., and Pybus, O. G. (2017). Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature*, 546(7658):406–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., and The WHO Rapid Pandemic Assessment Collaboration (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, 324(5934):1557–1561.

Gascuel, O. and Steel, M. (2014). Predicting the Ancestral Character Changes in a Tree is Typically Easier than Predicting the Root State. *Systematic Biology*, 63(3):421–435.

Gascuel, O. and Steel, M. (2019). A Darwinian Uncertainty Principle. *Systematic Biology*.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gill, M. S., Lemey, P., Bennett, S. N., Biek, R., and Suchard, M. A. (2016). Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Syst. Biol.*, 65(5):1041–1056.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.*, 30(3):713–724.

Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A., and Baele, G. (2020). Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. *Molecular Biology and Evolution*. msaa047.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332.

Guindon, S., Dufayard, J.-F. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59(3).

Haario, H., Saksman, E., and Tamminen, J. (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Statist.*, 14:375–395.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernouilli*, 7(2):223–242.

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular biology and evolution*, 27(9):1988–99.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174.

Holmes, E. and Twiddy, S. S. (2003). The origin, emergence and evolutionary genetics of dengue virus. *Infect. Genet. Evol.*, 3(1):19–28.

Hordijk, W. and Gascuel, O. (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, 21(24):4338–4347.

Hué, S., Pillay, D., Clewley, J. P., and Pybus, O. G. (2005). Genetic analysis reveals the complex structure of hiv-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A*, 102(12):4425–9.

Ishikawa, S. A., Zhukova, A., Iwasaki, W., and Gascuel, O. (2019). A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution*, 36(9):2069–2085.

Jiménez-Silva, C. L., Carreño, M. F., Ortiz-Baez, A. S., Rey, L. A., Villabona-Arenas, C. J., and Ocazionez, R. E. (2018). Evolutionary history and spatio-temporal dynamics of dengue virus serotypes in an endemic region of Colombia. *PLOS ONE*, 13(8):e0203090.

Jones, B. R. and Poon, A. F. Y. (2017). node.dating: dating ancestors in phylogenetic trees in R. *Bioinformatics*, 33(6):932–934.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of Protein Molecules. In *Mammalian Protein Metabolism*, pages 21–132. Elsevier.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Kozlov, A. M. and Stamatakis, A. (2020). Using raxml-ng in practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.3, pages 1.3:1–1.3:25. No commercial publisher | Authors open access book.

Langley, C. H. and Fitch, W. M. (1974). An examination of the constancy of the rate of molecular evolution. *Journal of molecular evolution*, 3(3):161–77.

Lefort, V., Longueville, J.-E., and Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, 34(9):2422–2424.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finding its roots. *PLoS Comp. Biol.*, 5(9):e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.*, 27(8):1877–1885.

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128.

Membrebe, J. V., Suchard, M. A., Rambaut, A., Baele, G., and Lemey, P. (2019). Bayesian inference of evolutionary histories under time-dependent substitution rates. *Mol Biol Evol*, 36(8):1793–1803.

Messina, J. P., Brady, O. J., Scott, T. W., Zou, C., Pigott, D. M., Duda, K. A., Bhatt, S., Katzelnick, L., Howes, R. E., Battle, K. E., Simmons, C. P., and Hay, S. I. (2014). Global spread of dengue virus types: mapping the 70 year history. *Trends Microbiol.*, 22(3):138–146.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091.

Minin, V. M., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.*, 25(7):1459–1471.

Nascimento, F. F., dos Reis, M., and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nat. Ecol. Evol.*, 1:1446–1454.

Naveca, F. G., Claro, I., Giovanetti, M., de Jesus, J. G., Xavier, J., Iani, F. C. d. M., do Nascimento, V. A., de Souza, V. C., Silveira, P. P., Lourenço, J., Santillana, M., Kraemer, M. U. G., Quick, J., Hill, S. C., Thézé, J., Carvalho, R. D. d. O., Azevedo, V.,

Salles, F. C. d. S., Nunes, M. R. T., Lemos, P. d. S., Candido, D. d. S., Pereira, G. d. C., Oliveira, M. A. A., Meneses, C. A. R., Maito, R. M., Cunha, C. R. S. B., Campos, D. P. d. S., Castilho, M. d. C., Siqueira, T. C. d. S., Terra, T. M., Albuquerque, C. F. C. d., Cruz, L. N. d., Abreu, A. L. d., Martins, D. V., Simoes, D. S. d. M. V., Aguiar, R. S. d., Luz, S. L. B., Loman, N., Pybus, O. G., Sabino, E. C., Okumoto, O., Alcantara, L. C. J., and Faria, N. R. (2019). Genomic, epidemiological and digital surveillance of chikungunya virus in the brazilian amazon. *PLOS Neglected Tropical Diseases*, 13(3):1–21.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Pett, W. and Heath, T. A. (2020). Inferring the timescale of phylogenetic trees from fossil data. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.1, pages 5.1:1–5.1:18. No commercial publisher | Authors open access book.

Pollett, S., Melendrez, M., Maljkovic Berry, I., Duchêne, S., Salje, H., Cummings, D., and Jarman, R. (2018). Understanding dengue virus evolution to support epidemic surveillance and counter-measure development. *Infection, Genetics and Evolution*, 62:279–295.

Price, M. N., Dehal, P. S., Arkin, A. P., Rojas, M., and Brodie, E. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490.

Pupko, T., Pe, I., Shamir, R., and Graur, D. (2000). A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896.

Quick, J., Loman, N., Durrafour, S., Simpson, J., Severi, E., Cowley, L., Bore, J., Koundouno, R., Dudas, G., Mikhail, A., Ouedraogo, N., Afrough, B., Bah, A., Baum, J., Becker-Ziaja, B., Boettcher, J., Cabeza-Cabrerizo, M., Camino-Sanchez, A., Carter, L., Doerrbecker, J., Enkirch, T., Garcia-Dorival, I., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L., Koropugui, M., Kosgey, A., Kuisma, E., Logue, C., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L., Portmann, J., Repits, J., Rickett, N., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R., Zekeng, E., Racine, T., Bello, A., Faye, O., Faye, O., Magassouba, N., Williams, C., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Camara, A., Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K., Diarra, A., Savane, Y., Pallawo, R., Gutierrez, G., Milhano, N., Roger, I., Williams, C., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D., Pollakis, G., Hiscox, J., Matthews, D., O'Shea, M., Johnston, A., Wilson, D., Hutley, E., Smit, E., Caro, A. D., Wölfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S., Koivogui, L., Diallo, B., Keïta, S., Rambaut, A., Formenty, P., Günther, S., and Carroll, M. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4):395–399.

Rambaut, A., Lam, T. T., Carvalho, L. M., and Pybus, O. G. (2016a). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.*, 2(1):vew007.

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016b). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus evolution*, 2(1):vew007.

Ratmann, O., Wymant, C., Colijn, C., Danaviah, S., Essex, M., Frost, S., Gall, A., Gaseitsiwe, S., Grabowski, M. K., Gray, R., Guindon, S., von Haeseler, A., Kaleebu, P., Kendall, M., Kozlov, A., Manasa, J., Minh, B. Q., Moyo, S., Novitsky, V., Nsubuga, R., Pillay, S.,

Quinn, T. C., Serwadda, D., Ssemwanga, D., Stamatakis, A., Trifinopoulos, J., Wawer, M., Brown, A. L., de Oliveira, T., Kellam, P., Pillay, D., Fraser, C., and on behalf of the PANGEA-HIV Consortium (2017). Hiv-1 full-genome phylogenetics of generalized epidemics in sub-saharan africa: Impact of missing nucleotide characters in next-generation sequences. *AIDS Research and Human Retroviruses*, 33(11):1083–1098.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive MCMC. *J. Appl. Prob.*, 44:458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *J. Comp. Graph. Stat.*, 18:349–367.

Robinson, D. (1971). Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11(2):105–119.

Rota-Stabelli, O., Lartillot, N., Philippe, H., and Pisani, D. (2013). Serine Codon-Usage Bias in Deep Phylogenomics: Pancrustacean Relationships as a Case Study. *Systematic Biology*, 62(1):121–133.

Sagulenko, P., Puller, V., and Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, 4(1).

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. S. (2014). tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–2080.

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 37(Database issue):D5—-15.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L., and Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–502.

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.*, 23(1):7–9.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

Suchard, M., Lemey, P., Baele, G., Ayres, D., Drummond, A., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25:1370–1376.

Swofford, D., Olsen, G., Waddell, P., and Hillis, D. (1996). Phylogenetic Inference. In Hillis, D., Moritz, C., and Mable, B., editors, *Molecular Systematics*, pages 407–514. Oxford University Press, Incorporated.

Tan, K.-K., Zulkifle, N.-I., Sulaiman, S., Pang, S.-P., NorAmdan, N., MatRahim, N., Abd-Jamil, J., Shu, M.-H., Mahadi, N. M., and AbuBakar, S. (2018). Emergence of the Asian lineage dengue virus type 3 genotype III in Malaysia. *BMC Evolutionary Biology*, 18(1):58.

Telford, M. J. (2007). Phylogenomics. *Current Biology*, 17(22):R945–R946.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. (2016). Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic biology*, 65(1):82–97.

Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., Pauwels, R., Alcantara, L. C., Vanden Eynden, E., Vandamme, A.-M., Deforche, K., and de Oliveira, T. (2019). Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics*, 35(5):871–873.

Volz, E. M. and Frost, S. D. W. (2017). Scalable relaxed clock phylogenetic dating. *Virus Evolution*, 3(2).

Walimbe, A. M., Lotankar, M., Cecilia, D., and Cherian, S. S. (2014). Global phylogeography of Dengue type 1 and 2 viruses reveals the role of India. *Infection, Genetics and Evolution*, 22:30–39.

Wertheim, J. O. and Kosakovsky Pond, S. L. (2011). Purifying selection can obscure the ancient age of viral lineages. *Mol Biol Evol*, 28(12):3355–65.

Woolley-Meza, O., Thiemann, C., Grady, D., Lee, J., Seebens, H., Blasius, B., and Brockmann, D. (2011). Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. *Eur. Phys. J. B.*, 84:589–600.

Worobey, M., Rambaut, A., and Holmes, E. C. (1999). Widespread intra-serotype recombination in natural populations of dengue virus. *Proc Natl Acad Sci U S A*, 96(13):7352–7.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yebra, G., Hodcroft, E. B., Ragonnet-Cronin, M. L., Pillay, D., Brown, A. J. L., Consortium, P. H., and Project, I. (2016). Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. *Scientific Reports*, 6(39489).

Yoder, A. D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.*, 17(1):1081–1090.

Zhang, J. and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution*, 44(S1):S139–S146.

Zhou, X., Shen, X.-X., Hittinger, C. T., and Rokas, A. (2018). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution*, 35(2):486–503.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins*, pages 97–166.