A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios

Sohta Ishikawa, Anna Zhukova, Wataru Iwasaki, Olivier Gascuel

A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios

Sohta A. Ishikawa,[†,1,2,3] Anna Zhukova,[†,1] Wataru Iwasaki,[2] and Olivier Gascuel*[,1]

[1]Unité Bioinformatique Evolutive, Institut Pasteur, C3BI USR 3756 IP & CNRS, Paris, France

[2]Department of Biological Sciences, The University of Tokyo, Tokyo, Japan

[3]Evolutionary Genomics of RNA Viruses, Virology Department, Institut Pasteur, Paris, France

***Corresponding author:** E-mail: olivier.gascuel@pasteur.fr.

[†]These authors contributed equally to this work.

Associate editor: Tal Pupko

Abstract

The reconstruction of ancestral scenarios is widely used to study the evolution of characters along phylogenetic trees. One commonly uses the marginal posterior probabilities of the character states, or the joint reconstruction of the most likely scenario. However, marginal reconstructions provide users with state probabilities, which are difficult to interpret and visualize, whereas joint reconstructions select a unique state for every tree node and thus do not reflect the uncertainty of inferences.

We propose a simple and fast approach, which is in between these two extremes. We use decision-theory concepts (namely, the Brier score) to associate each node in the tree to a set of likely states. A unique state is predicted in tree regions with low uncertainty, whereas several states are predicted in uncertain regions, typically around the tree root. To visualize the results, we cluster the neighboring nodes associated with the same states and use graph visualization tools. The method is implemented in the PastML program and web server.

The results on simulated data demonstrate the accuracy and robustness of the approach. PastML was applied to the phylogeography of Dengue serotype 2 (DENV2), and the evolution of drug resistances in a large HIV data set. These analyses took a few minutes and provided convincing results. PastML retrieved the main transmission routes of human DENV2 and showed the uncertainty of the human-sylvatic DENV2 geographic origin. With HIV, the results show that resistance mutations mostly emerge independently under treatment pressure, but resistance clusters are found, corresponding to transmissions among untreated patients.

Key words: **phylogenetics, ancestral character reconstruction, maximum likelihood, marginal and joint posterior probabilities, maximum a posteriori, Brier scoring rule, simulations, Dengue, HIV, phylogeography, drug resistance mutations.**

Introduction

A central issue in biology is to recover and understand the evolutionary history of biological entities. These may be of different nature and scale, ranging from DNA and protein sequences to communities, going through biological systems, organs, strains, individuals, species, and populations. The characteristics and evolution of these objects are measured using a variety of "characters," including molecular properties (e.g., Werner et al. 2014; Bickelmann et al. 2015; Busch et al. 2016), gene contents of genomes (e.g., Iwasaki and Takagi 2007), morphological and phenotypic characteristics (e.g., Endress and Doyle 2009; Marazzi et al. 2012; Beaulieu et al. 2013; Sauquet et al. 2017), ecological traits (e.g., Maor et al. 2017), and geographic locations (e.g., Arbogast 2001; Wallace et al. 2007; Lemey et al. 2009, 2014; Edwards et al. 2011; Dudas et al. 2017; Magee et al. 2017). Ancestral character reconstruction (ACR) is central in all these studies to trace the origin and evolution of the character of interest. ACR relies first on the inference of phylogenetic relationships among the studied objects, that is, a phylogenetic tree, typically inferred from DNA or protein sequences. The character state is generally known for all (most) tips of the tree (some methods can accommodate for unknown or ambiguous state values). ACR is commonly used to reconstruct ancestral sequences corresponding to specific tree nodes (typically the tree root). ACR is also used to determine how the character of interest has changed on the tree from the root to the tips over evolutionary time, by assigning the most likely ancestral character states to every internal node. This global reconstruction over the whole tree describes the evolutionary history of the character and is commonly called an "ancestral scenario," which is the focus of this article. Several approaches have been proposed for ACR so far, including parsimony (Swofford and Maddison 1987), maximum likelihood (ML; Pagel 1999; Pupko et al. 2000; Felsenstein 2004; Ree and Smith 2008), and Bayesian methods (Huelsenbeck and Bollback 2001; Pagel et al. 2004).

Parsimony-based ACR provides quick and simple methods to infer ancestral scenarios. However, due to the oversimplification of evolutionary processes (e.g., not accounting for branch lengths and evolutionary times), parsimony has limited accuracy (Collins et al. 1994). ML and Bayesian approaches are based on probabilistic models of character evolution. ML methods were shown to perform better than parsimony, using both theoretical arguments and simulation studies under a variety of conditions (Zhang and Nei 1997; Gascuel and Steel 2014). Simulation results showed that even the simplest models (e.g., JC, Jukes and Cantor 1969) yield more accurate reconstructions than parsimony (Gascuel and Steel 2014), thanks to the consideration of evolutionary times and branch lengths, and are robust to moderate model violations and phylogenetic uncertainty (Hanson-Smith et al. 2010).

The size of the trees subjected to ACR has rapidly increased thanks to new generation sequencing technologies. Evolutionary and epidemiological analysis of pathogens, like human immunodeficiency virus (HIV), Influenza, and Ebola, is one of the hotspots of this problem, with data sets commonly comprising thousands of strains (Holmes et al. 2016; Ratmann et al. 2017; Durães-Carvalho and Salemi 2018). With such rapidly evolving pathogens, the links between evolutionary and epidemiological processes raise essential public health questions with important practical issues, notably the routes and patterns of pathogen spread (Wallace et al. 2007; Faria et al. 2014; Lemey et al. 2014; Gräf et al. 2015; Dudas et al. 2017; Magee et al. 2017) and the emergence of drug resistances (Mourad et al. 2015; Zhukova et al. 2017). In these studies (and many others), ACR was a major tool, aiming to map ancestral states of pathogen characters (e.g., sampling location, risk group of the host, and presence of drug resistance) on the tree inferred from DNA or protein sequences.

Bayesian methods (Huelsenbeck and Bollback 2001; Pagel et al. 2004; Drummond et al. 2012) are commonly used in this context, notably in phylogeography studies (Lewis et al. 2015; Magee et al. 2017). The main approach is to infer the joint posterior distributions of ancestral character states, phylogenetic tree, and model parameters, using a Markov chain Monte Carlo (MCMC) procedure. This involves complex probabilistic models describing the evolution of the sequences, the molecular clock (possibly relaxed and correlated), the demography, and last but not least the evolution of the studied character. Stochastic mapping (Nielsen 2002; Huelsenbeck et al. 2003) is simpler and commonly used to generate and compare alternative, plausible evolutionary histories of the studied character on a given tree. The character evolution model can be very simple, typically symmetrical with a few states, but the current trend is to rely on increasingly complex models, nonsymmetrical, with latent variables, dozens of character states, and evolution over time (Stadler and Bonhoeffer 2013; Lambert et al. 2014; Leventhal et al. 2014; Kühnert et al. 2014, 2016). The Bayesian approach is very popular because of this wealth of options and flexibility, via famous software programs like BEAST (Drummond et al. 2012). However, Markov chain Monte Carlo based methods have a high computational cost, and the joint inference of all the tree, parameter, and character distributions cannot be achieved for large data sets. Even the stepwise approach where we first infer the tree distribution, and then the distribution of the studied character along the most likely trees is hardly applicable to medium data sets (500–1,000 tips), requiring high performance computing units (Graphics Processing Units [GPUs]), sophisticated parallel implementations (Ayres et al. 2012, 2019), and days to weeks of computation. In contrast, the ML approach is less computationally demanding as it gives point estimates for the parameters of interest, instead of distributions. For example, TreeTime (Sagulenko et al. 2018) is able to deal with large trees comprising thousands of tips and perform fast ML-based ACR in a few minutes or even a few seconds.

However, there are still potential limitations in applying standard ML-based ACR to large data sets and trees. These limitations are related to the inference of the character states, the uncertainty that is inherent to such inference and that of the phylogeny, as well as the visualization and interpretation of the (large) resulting ancestral scenario. Two main approaches are used in ML-based ACR:

- Either we compute the marginal posterior probabilities of every state for each of the tree nodes (Felsenstein 2004; Yang 2007). Then, we usually select the state with the highest posterior. This maximum a posteriori (MAP) selection is independent from one node to another, which could induce globally inconsistent scenarios (typically: two very close nodes with incompatible predictions). This possible shortcoming formed the basis of criticisms against MAP-based ACR.
- Or we compute the joint ancestral scenario with the maximal posterior probability (Pupko et al. 2000) using dynamic programing. This approach has some global consistency guarantee but returns a unique scenario, and thus does not reflect the fact that with real data and large trees, billions of scenarios may have similar posterior probabilities.

Our simulations (Gascuel and Steel 2014; see also results below) showed that the predictions and accuracy of MAP and joint approaches are very close. This advocates for the use of the full marginal approach, which not only indicates the most likely state for each node (as predicted by MAP) but also returns the marginal posterior probabilities of all states, thus reflecting the uncertainties of node predictions. However, interpreting and using these probabilistic outputs is difficult, for example when two states have similar posteriors. Another difficulty is to visualize and summarize the resulting, global scenario, which commonly involves thousands of probability distributions attached to each of the tree nodes.

Here, we propose a simple and fast approach to overcome these limitations. We use decision theoretic concepts and tools to infer a limited set of likely states for each of the tree nodes, which best approximate the marginal posterior probabilities. In the easy regions of the tree (typically close to the tips [Gascuel and Steel 2014]), this approach predicts a unique state, whereas in the difficult parts (typically close to the root) it may predict several likely states reflecting the

uncertainty of the inferences. Such results are typically encountered in phylogeography, where the deep origin of the studied species or virus cannot be determined with certainty, whereas its recent history is almost certain (see application to Dengue below). When one is mostly interested in the best guess, this approach will highlight that some of the nodes are predicted with high confidence, whereas some others are not (as with numerical values and confidence intervals). This approach is generic and could be used to reconstruct ancestral sequences (see Oliva et al. [2019] for a closely related method), but we restrict ourselves here to unique, discrete characters, for example geographical or morphological. To summarize and visualize the resulting scenarios in this framework, we cluster the neighboring nodes with identical predictions and reuse some of the ideas we developed in parsimony-based PhyloType software (Chevenet et al. 2013). This way we obtain a compact, tree-shaped and easily interpretable graphical representation of the most likely ancestral scenarios, which is robust to phylogenetic uncertainties and sampling rate variations. In the following, we first describe the different components of the method, then the results with simulated data along with comparisons with other ACR methods, and lastly two real data analyses on the phylogeography of Dengue and the evolution of drug resistances in a large HIV data set. All methods developed and studied in this article are implemented in the PastML software, which is freely available in several versions and interfaces, including a web server (https://pastml.pasteur.fr/).

New Approaches

Preamble

The method can be decomposed into three main steps: 1) ML-based rescaling of the tree and estimation of the model parameters, 2) ancestral reconstruction of the most likely character states, and 3) compression and visualization of the inferred ancestral scenario. These three steps are described in turn in the following. In this section, we describe the input data, notation, model and global framework, and goals.

The input of the method is a rooted tree denoted as T, where every tip is associated with a character state. The number of tree tips is denoted as n and the tree root as R. T may be not fully resolved, the method applies to both binary and nonbinary trees. In most cases, T is obtained from a multiple alignment of sequences (DNA or proteins) using some standard phylogenetic software. Then, the branch lengths are expressed in number of substitutions per site. As we shall see, the input tree is rescaled to fit the evolution of the studied character, and thus all branch-length measures are acceptable. Most interesting results will be obtained with time scaled trees, where branch lengths are expressed in years. Then, the rescaling factor estimated from the input data represents the average number of character changes per year.

The studied character may be of various natures, as discussed in the Introduction. Here, we consider discrete characters with values taken from a finite, nonordered set of states; for example: {Africa, America, Asia, Australia, Europe} in phylogeography, or {Sensitive, Resistant} when studying drug resistances. S denotes the set of possible states, with size s. A tree tip (or leaf) is denoted as l, and $c(l) \in S$ is the character state associated with l. The method is able to accommodate tips with unknown character values, denoted as $c(l) = X$, as well as ambiguities when for a given tip several states are possible. Then, $c(l)$ is a subset of S, and consistently $c(l) = X$ is equivalent to $c(l) = S$.

Continuous-time Markov models are commonly used to represent the evolution of characters, notably with sequences where all sites of the studied multiple alignment are usually assumed to evolve according to the same model (with different rates when using rates across sites models, e.g., gamma distributed). In this setting, especially with DNA having four states only, one is able to accurately estimate the parameters of relatively complex models, for example GTR (Tavaré 1986) having ten parameters and 8 degrees of freedom with DNA. Here, we have a unique observation describing the evolution of the studied character through the tips values, and accurately estimating the parameters of complex models is usually difficult, if not impossible, especially when s is large (Gascuel and Steel 2018). We therefore use simple s-state JC-like and F81-like models, which generalize to s states the 4-state JC and F81 models for DNA (Jukes and Cantor 1969; Felsenstein 1981). With JC-like models all rates of changes from state i to state j ($i \neq j$) are equal, whereas with F81-like models, the rate of changes from i to j ($i \neq j$) is proportional to the equilibrium frequency of j, denoted as π_j. JC-like models are special cases of F81-like ones, with all equilibrium frequencies equal to $1/s$. Several studies advocate the use of s-state F81-like models. We showed (Gascuel and Steel 2014), using simulations, that even the simpler JC-like version performs nearly as well as the true model, with DNA-like data generated using an HKY model (Hasegawa et al. 1985) with high transition/transversion rate and heterogeneous nucleotide frequencies. Moreover, Dudas et al. (2017) showed that the origin and destination population sizes (represented by π_i and π_j, respectively) are two of the main factors explaining Ebola dissemination in West Africa. This finding is in accordance with the use of s-state F81-like models, where the expected number of changes from i to j is proportional to $\pi_i\pi_j$. Another advantage of F81-like models is that the probability of changes along a branch of length t is simply expressed as

$$\begin{aligned} \mathrm{PC}(i \to j/t) &= (1 - e^{-\mu t})\pi_j \qquad \text{if } j \neq i \\ &= e^{-\mu t} + (1 - e^{-\mu t})\pi_i \quad \text{otherwise,} \end{aligned}$$

where μ is the normalization factor:

$$\mu = 1/(1 - \sum \pi_i^2).$$

An s-state F81-like model has s parameters ($s - 1$ degrees of freedom) corresponding to the equilibrium frequencies of the s states. In our software, these frequencies can be user supplied, roughly estimated from the state frequencies

observed at the tree tips (not recommended, see Gascuel and Steel [2018]), or estimated by ML as explained in the next section.

Tree Rescaling and Editing, Parameter Estimation

Beyond the state equilibrium frequencies, the whole model involves one additional parameter, namely the "global rate," denoted as ρ. With a unique character, as is the case here, estimating all branch lengths in the tree is just impossible. We therefore assume that the number of character changes along the tree is proportional to the branch lengths of the input tree. Every branch length t is turned into ρt, which is interpreted as the expected number of character changes along the given branch. Moreover, we assume that ρ is constant across the tree over evolutionary time, which is a similar assumption to the one-rate model (Mooers and Schluter 1999). With dated input trees, the original branch lengths are measured in years and ρ in number of state changes per year. The estimated value of ρ is then highly informative about the global evolutionary rate of the studied character along the tree.

Both dated and molecular trees may have branches of length zero. For example, when two input sequences are identical (quite common with virus strains), we expect that any reasonable phylogenetic method infers a cherry with null branches connecting the two sequences (a cherry is a rooted subtree of two taxa). Similar configurations may happen in dated trees due to temporal constraints (To et al. 2016). However, two identical sequences may have been observed in different countries, thus giving rise to two different character states linked by a path of length zero, and the same may happen with other types of characters (e.g., phenotypic). In a standard phylogenetic setting, the likelihood of any scenario containing such a configuration is null and no ML-based ancestral reconstruction is possible. To circumvent this difficulty, we edit the input tree. First, all internal branches of length zero are turned into polytomies. Then, for each node having daughter tips with branches of length zero, we compute the union of their states, which is assigned to all these tips (e.g., consider a node v with three daughter tips: x with state i and zero branch length, y with state j and zero branch length, and z with state i and nonzero branch length; then, in the edited tree x and y are associated with state set $\{i,j\}$ and z is unchanged). This simply expresses that the corresponding taxa were observed with different states (e.g., the same sequences in different countries). The likelihood is computed as usual (Materials and Methods) from this edited (and rescaled) tree, in order to estimate the model parameters and infer the ancestral states. Before outputting the state predictions and visualizing the results, the zero-tip configurations are reedited back to their original state.

To estimate the model parameters (ρ, the equilibrium frequencies with F81-like models, and κ the transition/transversion ratio with the HKY model used in simulations), we compute the scenario likelihood using the standard pruning algorithm (Felsenstein 1981), and use the limited-memory bounded BFGS optimization routine to obtain ML estimates (L-BFGS-B [Byrd et al. 1995; Zhu et al. 1997], available in Python SciPy library [Oliphant 2007]). This is achieved in two steps (both using the pruning and L-BFGS-B algorithms): first we obtain a rough estimate of ρ assuming a JC-like model, then we estimate all parameters together, starting from the previously estimated value of ρ and equal frequencies for all the states (and $\kappa = 4$ with HKY). The L-BFGS-B algorithm allows for constraints. To avoid very large or very low values of the global rate ρ (possible when the tip states are very similar or highly divergent), we impose: $\beta^{-1} \times 0.001 \leq \rho \leq \beta^{-1} \times 10$, where β is the average length of nonzero branches. This means that in the rescaled tree the number of changes along a branch with original length β is in between 0.001 and 10 (note that with 10 changes per branch reconstructing ancestral scenarios is just impossible). These bounds on the ρ estimate performed well in all our applications to real and simulated data. To ensure that the sum of frequencies is 1.0, one of the frequencies is kept equal to $1/s$, the others are freely optimized (but constrained to be positive), and all frequencies are normalized in likelihood computations and when outputting the results.

Discrete Approximation of the State Marginal Posterior Probabilities

Our method is based on a discrete approximation of the marginal posterior probabilities of the character states, attached to the internal nodes of the tree. The computation of these probabilities is standard and used under different forms in most if not all ML-based phylogenetic programs. However, the complete description of the procedure is rarely available and is included in the Materials and Methods for the sake of completeness. To summarize, we first use the pruning algorithm (Felsenstein 1981), which performs a bottom up, postorder tree traversal, and accounts for the information of the descendants of every tree node; then, we perform a top-down, preorder tree traversal, which adds to the previous calculations the information coming from the rest of the tree. We thusly obtain for every tree node N and state i, the marginal posterior probability of i for N, Marginal(N,i), which accounts for the state value of all tree tips. This procedure has a time complexity in $O(ns^2)$, where n is the number of tips and s the number of states. It is therefore linear in n and able to process trees with dozens of thousands of tips in a few seconds. It is equivalent (but faster) to the procedure consisting in iteratively rerooting the tree with every internal node and applying the pruning algorithm. The reconstruction accuracy is clearly higher than that obtained with the pruning algorithm (without rerooting) and the descendant information only (Gascuel and Steel 2014).

Let N be any given internal node of T. Based on the marginal posterior probabilities Marginal(N,i), we have to decide which states are predicted for N and which ones are discarded because their posteriors are too low. We could use some thresholding approach, but the choice of the threshold values and decision procedure would be very subjective, without any formal guarantee on the accuracy of the predictions (see Oliva et al. 2019 for simulation results). We therefore used concepts and tools from decision theory and supervised classification (Brier 1950; Gneiting and Raftery 2007).

Assume that the true evolutionary model (tree, branch lengths, and model of character changes) is fully known; then, a standard result (Guiasu 1977), known as the Bayes decision rule, is that the most accurate prediction for N is obtained by selecting the state with the highest posterior (MAP). In this framework, we predict a unique state, and the accuracy is simply measured by the probability of correctly predicting the true ancestral state. However, in our framework, a unique prediction per node is often unsatisfactory, especially when several states corresponding to different scenarios have similar posteriors. Then, a refined approach involves using probabilistic predictions, where states are assigned to probabilities instead of binary, mutually exclusive decisions as with the Bayes rule. When the true evolutionary model is fully known, the marginal posterior probabilities can be shown to be optimal among all probabilistic predictors (Gneiting and Raftery 2007). Various scoring criteria (or scoring rules) have been proposed to measure the accuracy of probabilistic predictors. The most used is the logarithmic scoring criterion from information theory. This is the negative of "surprisal," which is commonly used in Bayesian inference. However, this scoring criterion is not appropriate here, where we have state predictions with null probabilities (see below). Our approach is derived from the Brier quadratic scoring criterion. Let $\mathrm{PPr}(N,i)$ be the predicted probability of state i for node N and $\mathrm{Truth}(N,i)$ be the "truth" of i for N, which is equal to 1 when the ancestral state of N is i, and 0 otherwise. The Brier score can be expressed as

$$\mathrm{Brier}(N) = \sum_{i \in S} [\mathrm{PPr}(N,i) - \mathrm{Truth}(N,i)]^2.$$

In this form, the Brier score is simply the squared Euclidean distance between $\mathrm{PPr}(N)$ and $\mathrm{Truth}(N)$ (the lower the better). In practice, the truth is rarely known, except when analyzing simulation results and past events (e.g., in weather forecasting). In this context, the Brier score is commonly used to measure the accuracy of probabilistic predictors. For instance, assuming that we assign probability 1 to the true state, then $\mathrm{Brier}(N) = 0$. On the opposite, if we assign probability 1 to an incorrect state, then $\mathrm{Brier}(N) = 2$, which is the worst possible value of the score. Assume now that we have no information on the ancestral state of N. Then, we have two natural solutions: 1) assign probability $1/s$ to every state, then $\mathrm{Brier}(N) = (1-1/s)^2 + (s-1)(1/s)^2 = 1 - 1/s$; 2) randomly, uniformly predict one of the states, then the expected value of $\mathrm{Brier}(N)$ is equal to $0 \times 1/s + 2 \times (s-1)/s = 2 - 2/s$. In other words, random predictions are worse than the recognition of our ignorance.

As already stated, when the model is fully known predicting the marginal posterior probabilities of the states is optimal, regarding the Brier criterion (and other proper scoring rules, as the logarithmic one). Thus, we use a discrete approximation of the posteriors, which is consistently selected using the Euclidean distance. The goal is to add as little error as possible to the (unknown) optimal value of $\mathrm{Brier}(N)$. Thanks to the triangle inequality, this ensures a form of minimization of the Brier score (see below). Assuming that we decide to retain k states (among s) in the predictions, then each of these has probability $1/k$, whereas the discarded states have probability 0. These probabilities are used in the selection of state subsets but are implicit. The method returns a set of likely states without any associated probabilities. To define the state subsets to be explored, we rank the states based on their posteriors: i_1 $(=\mathrm{MAP})$ is best and i_s has the lowest posterior. Then, we select the best subset $SS_k = \{i_1, i_2 \ldots i_k\}$ ($k = 1$ to s) by minimizing the Euclidean distance $D_k(N)$ between $\mathrm{Marginal}(N)$ and the probability vector defined by $\Pr(i_{1 \text{ to } k}) = 1/k$ and $\Pr(i_{k+1 \text{ to } s}) = 0$.

This method is named Marginal Posterior Probabilities Approximation (MPPA) in the following. Let $\mathrm{BrierMPP}(N)$ be the Brier score of the marginal posteriors probabilities, and $\mathrm{BrierMPPA}(N)$ the Brier score of MPPA. Both scores are unknown, since the truth is unknown, but we have (triangle inequality): $\sqrt{\mathrm{BrierMPPA}(N)} - \sqrt{\mathrm{BrierMPP}(N)} \leq D_k(N)$. In other words, by minimizing $D_k(N)$, we minimize the gap between the two Brier scores, where BrierMPP is optimal assuming a fully correct model. We shall see with simulated data that BrierMPPA is close to BrierMPP, especially with analyzes performed under the true model.

This method is both simple and fast, with time complexity in $O(ns^2)$ again. Its accuracy strongly depends on the accuracy of the marginal posteriors, and thus on the severity of model violations, which are inevitable with real data. A possible shortcoming could be that these computations are performed independently for each of the nodes. However, we observed with simulations (Gascuel and Steel 2014; results below) that joint reconstructions have no clear advantage over marginal ones. Actually, in a large fraction of cases, the MAP and joint states are the same, which explains this finding. Moreover, the MAP state is always included in the set of states resulting from MPPA. For the rare cases ($\ll$1% in our simulations below) where the joint state is not included in MPPA state set, a program option adds it systematically to ensure that the joint (most likely) scenario is included in MPPA solution.

Tree Compression and Visualization

On large phylogenies with hundreds or thousands of tips, once the ancestral states are reconstructed on each node, it might be difficult for a human eye to visualize and interpret the results. To overcome this issue, we provide a compressed representation of the ancestral scenarios, which highlights the main facts and hides minor details. This representation is calculated in two steps: 1) "vertical merge" that clusters together the parts of the tree where no state change happens and 2) "horizontal merge" that clusters independent events of the same kind. Algorithmically, the two merges are performed in the following way:

- Vertical merge (vertical arrow in fig. 1) : while there exists a parent–child couple such that the child's set of input/predicted states is the same as the parents' one, merge them. Note that tips can be merged as well as internal

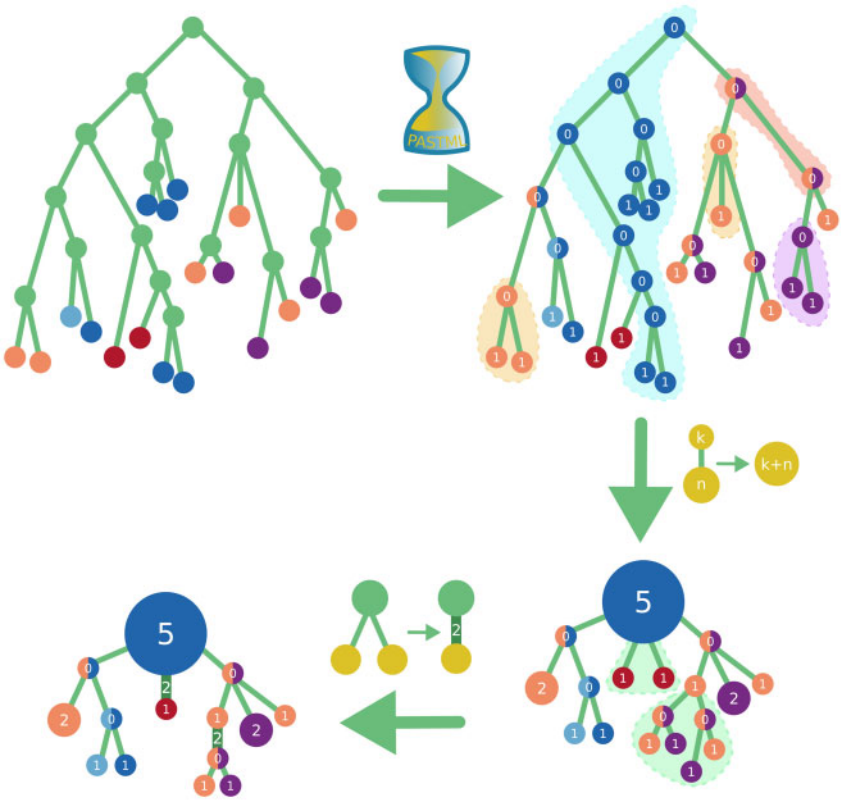

FIG. 1. ACR and visualization steps. Starting from the initial tree with annotated tips (top left, different annotations correspond to different colors), we reconstruct the ancestral node states (top right; colored sectors are used for ambiguous nodes, e.g., orange and purple), and then perform a two-step compression: the vertical compression (bottom right) clusters together the regions of the tree where no state change happens and puts the number of tips collapsed into each node as its size (e.g., the blue root cluster of size 5 in the bottom right tree corresponds to the part of the top-right tree highlighted blue and containing 5 tips), whereas the horizontal compression (bottom left) merges identical subtree configurations, keeping their number as branch sizes (e.g., the two red tip children of the bottom right root are merged into a red tip attached with a branch of size 2 in the bottom left tree).

nodes in this process. We compute the size of a cluster so obtained, by the number of tips contained in the cluster, as the tips correspond to the input data units used for tree and ancestral scenario reconstructions. Accordingly, in the initial tree each tip has a size of 1, each internal node has a size of 0, and when merging two nodes we sum their sizes. In the compressed tree, a cluster is a node or "phylotype" following Chevenet et al. (2013).

- Horizontal merge (horizontal arrow in fig. 1): starting at the root and going top-down toward the tips, at each node we compare its child subtrees. If two or more identical subtrees are found, we keep just one representative and assign their number to the size of the branch that connects the kept subtree to the current node. Hence, a branch size corresponds to the number of times its subtree is found in the initial tree. Before the horizontal merge all branches have a size of 1. Two trees are considered *identical* if they have the same topology and their corresponding nodes have the same state(s) and sizes.

These two routines are illustrated in figure 1. In the case of a transmission tree with states representing countries, the vertical merge will cluster together the transmissions happening within the same country and having the same source within that country; for instance, see the clouds colored in blue, orange and purple in figure 1. Then, the horizontal merge detects independent transmissions from a country A to a country B; for instance in figure 1, the two red nodes (=B) that branch independently from a big blue circle (=A).

For large trees with many state changes, even after compression the visualization might contain too many details. To address this issue, a program option makes it possible to specify the desired number of tips shown in the compressed tree (15 by default), which is then achieved by performing a relaxed horizontal merge and hiding less important nodes. In a relaxed horizontal merge, the definition of identical trees is updated: instead of requiring identical sizes of the corresponding nodes, we allow for nodes of sizes of the same order (log 10); for instance, now a node in state A of size 3 can correspond to a node in state A of any size between 1 and 9, and a node in state B of size 25 can correspond to a node in state B of any size between 10 and 99.

If even after a relaxed horizontal merge the compressed representation contains too many details, we trim less important tips as follows. For each node, we calculate its importance by multiplying its size by the sizes of all the branches on the path to the root, therefore obtaining the number of tips of the original tree that are represented by this node; for

instance, a node of size 2 connected to the root via branches of sizes 1, 3, and 5 gets an importance of 30. We call a node blocked by its descendant if its importance is smaller than the descendant's one. The intuition behind is that this node cannot be removed from the tree unless the importance threshold allows the removal of its descendant first. We then set the cutoff threshold to the 15th largest nonblocked node's importance (a parameter that can be adjusted), and iteratively trim all tips with smaller importance (once a tip is removed its parent becomes an unblocked tip itself and is also considered for trimming). Finally, we rerun relaxed horizontal merge as some of the previously different topologies might have become identical after trimming. This trimming procedure and its parameters are somewhat empirical, but we observed that it performed well on all data sets we analyzed here and in other studies.

To simplify the ancestral state analysis of phylogenetic trees with multiple character data, we developed a pipeline that combines the results for different characters, for example geographical location and resistance to drugs. We first apply ancestral reconstruction separately for each character to obtain their ancestral states, and visualize each character on the tree nodes as sectors (see application to HIV below). If we could not choose a unique state for a character, we keep the corresponding sector uncolored (i.e., white). Once the tree is colored and each node is assigned its combined states (pie of colors), we compress the tree as described in the previous section.

When sampling dates are available for the tree tips, it is possible to visualize a timeline: we split the time between the oldest and the most recent samples into five intervals, and add a visualization slider to navigate in time. At each milestone we hide the subtrees for which all tips were sampled later. When sampling dates are not available, the timeline is based on root-to-tip distances instead.

Software and Utilities

PastML takes as input a rooted tree and a tip state annotation table (for one or more characters). It produces a table with predicted ancestral states, and an interactively modifiable visualization (an html file that can be viewed in a browser). PastML is implemented in python 3, uses Scipy/Numpy (Oliphant 2007) for parameter optimization/estimation, and Cytoscape.js library (Franz et al. 2016) for tree visualization. An additional option performs an automatic upload of the full tree, annotated with ACR predictions, to iTOL, an interactive tool for tree management and visualization (Letunic and Bork 2016). PastML is available as a python library/command-line program on pip3. We also provide a docker container that includes all the functionality and does not require installing python: evolbioinfo/pastml. Last but not least, a user friendly web application is available to perform ACR and visualization of ancestral scenarios. Several ACR methods are available: our new MPPA algorithm; the standard marginal posterior probability approach (both MAP and full probabilistic predictions); the joint posterior probability estimation algorithm of Pupko et al. (2000); and the three usual variants of parsimony: ACCTRAN, DELTRAN, and DOWNPASS (Maddison and Maddison 2000). All this material (source code, docker container, web server, etc.) is available from https://pastml.pasteur.fr.

Results: Method Comparison Using Simulated Data

Simulation Protocol

In this study, we basically followed the simulation procedure used in Gascuel and Steel (2014). We generated pure-birth trees with $n = 1{,}000$ tips. To obtain a broad range of ACR difficulties, we used 16 values of the speciation/evolutionary rate ratio (ω) ranging from 0.2 to 8.0, which correspond to an average number of state changes per branch of 2.5 and 0.0625, respectively (Steel and Mooers 2010). With a high number of state changes per branch (e.g., 2.5) ACR is very difficult, especially for the tree root, whereas with a low number of changes (e.g., 0.0625) ACR becomes easy as all tips and nodes tend to have the same state value. For each value of ω, 50 trees were generated, and for each tree we simulated the evolution of 50 unique characters with 4 states, and 50 with 20 states. To ease the implementation, reproducibility and interpretation of the results, we used DNA and protein models, although the method and software are intended for unique characters. We generated 4-state data sets using Seq-Gen v1.3.2 (Rambaut and Grass 1997) and the HKY model (Hasegawa et al. 1985) with equilibrium frequencies of A, C, G, and T being equal to 0.2, 0.1, 0.3, and 0.4, respectively, and a transition/transversion ratio κ of 8.0. These relatively extreme values were chosen to challenge ACR when using the F81 model implemented in PastML. Likewise, we generated 20-state data sets using Seq-Gen and the JTT model (Jones et al. 1992) with its default amino-acid equilibrium frequencies. We thusly obtained 16 (ω values) $\times$ 50 (1,000-tip trees) $\times$ 50 (number of characters) $\times$ 2 (4-state/20-state) data sets to assess the accuracy of ACR methods. During the simulation procedure with Seq-Gen, we recorded the ancestral state of the character seen at each internal node, including the root. Thus, the "true" ancestral scenario was known. All these data sets are available from https://pastml.pasteur.fr/.

Methods Being Compared

These simulated trees and tip state values were then subjected to ACR with five methods:

- Parsimony: We computed the most parsimonious states for all nodes including the root. This computation was performed using UPPASS (Fitch-Hartigan) and then DOWNPASS algorithms, which combine the state information from all tree tips (analogous to the calculation of posteriors, Materials and Methods). DOWNPASS returns all most parsimonious states for all nodes, as opposed to ACCTRAN and DELTRAN which solve part of the ancestral ambiguities heuristically (Maddison and Maddison 2000). This method returns a set of possible states for each node, and thus shares common points with MPPA.
- Joint: We used the dynamic programing algorithm described in Pupko et al. (2000) to infer the most likely ancestral scenario over all the tree and possible state

values. For each of the nodes, we thus obtained a joint estimation of the most likely state. This method returns a unique state for each node.

- Marginal: we computed the marginal likelihoods and posterior probabilities of all states for all internal nodes including the root (see Materials and Methods). This method returns full probabilistic predictions (each state is assigned a probability) for all nodes. Marginal is the best possible probabilistic predictor (as measured by the Brier score, and other proper scoring rule) when the model is fully known.
- MAP: Using previous computations, we assigned the state with the highest marginal posterior to each node. MAP thus returns a unique state for each of the nodes, just as Joint.
- MPPA: We used the method described above to approximate the state posteriors and return for every node a subset of likely states.

With the ML-based methods (i.e., Joint, Marginal, MAP, and MPPA), the simulated data sets were analyzed under the four following conditions, corresponding to various model violations intended to measure the robustness of the ACR methods being compared:

(1) True model and branch lengths: the true evolutionary model (i.e., the model used in simulations, HKY with 4 states and JTT with 20 states) was used for ACR. The model parameters were estimated from the data (i.e., with HKY: the global rate ρ, the nucleotide frequencies, and κ; with JTT: ρ only, we used JTT frequencies as usual). The (true) input tree was rescaled using the estimated value of ρ before performing ACR (standard usage of PastML, used in all conditions). In this setting, there is no model violation and Marginal is expected to be optimal regarding the Brier score. The goal was to check that the results were only slightly degraded with model violations.

(2) True model and noised branch lengths: the true model was used, as in previous setting (including parameter estimation and rescaling), but the branch lengths of the input trees were perturbed. The goal was to account for approximate branch-length estimation and molecular-clock violation, two common features with usual data sets. To obtain the input tree, all branch lengths of the true tree were multiplied by independent, lognormal variables, with mean 1.0 and standard deviation 0.5 (i.e., uncorrelated, lognormal molecular-clock model with standard deviation similar to that observed in HIV data [To et al. 2016]).

(3) F81 model and true branch lengths: we used F81 and F81-like models for 4- and 20-state data sets, respectively. The model parameters (ρ and equilibrium frequencies) were estimated from the input data. This setting corresponds to the default option of PastML. The goal was to check that the loss of accuracy was low, compared with the perfect "true model and branch lengths" setting.

(4) F81 model and noised branch lengths: we combined 4- and 20-state F81-like models and optimizations, with noised input trees (see conditions 2 and 3 above). This setting can be seen as realistic as it combines the default evolutionary model in PastML with branch-length perturbation.

Comparison Criteria

To compare the accuracy of the various ACR methods being tested, we used the Brier score of the predicted states against the known, true scenario. In the above Brier score formula $\mathrm{Truth}(N,i)$ was equal to 1 when the true state of node N was i, and 0 otherwise. Accordingly, $\mathrm{PPr}(N,i)$ was equal to 1 (i is predicted) or 0 (i is not predicted) for the methods predicting a unique state (i.e., Joint and MAP); $\mathrm{PPr}(N,i)$ was equal to $1/k$ (i is predicted) or 0 (i is not predicted) with Parsimony and MPPA when k states were predicted; with Marginal, $\mathrm{PPr}(N,i)$ was simply equal to the marginal posterior probability of state i for node N. The Brier scores of the nodes were then averaged, and we returned the average score over 2,500 trials (50 trees $\times$ 50 trials) for each of the simulation conditions. The same criterion was also applied to the tree root, as ACR is expected to be more difficult with the root than with other tree nodes, especially those being close to the tips (Gascuel and Steel 2014).

To compare the performance of the various methods in producing consistent predictions across the tree, we applied the Brier score to edges instead of nodes. The goal was to check that the predictions for the two extremities of any given edge were compatible and close to the truth, thus establishing, or not, the superiority of global predictions as produced by Joint, over independent predictions as produced by MAP or MPPA (see also Gascuel and Steel 2014). An edge E was perfectly predicted $[\mathrm{PPr}(E,i,j) = \mathrm{Truth}(E,i,j) = 1]$ when its two extremity states i, j were the same as the true ones. In case of multiple predictions, k on one extremity and p on the other, $\mathrm{PPr}(E,i,j)$ was equal to $1/kp$ when both states were included in the predicted states at both edge extremities, and 0 otherwise. With Marginal we simply used for $\mathrm{PPr}(E,i,j)$ the product of the state posteriors of i and j at both extremities of E. The formula to compute the Brier score was the same as for nodes, but considering a state space of size s^2.

For Parsimony and MPPA, we counted the average number of predicted states per node in the various simulation conditions. The goal was to check the advantage of predicting several states instead of a single one, in case of uncertainty. The same was also applied to the tree root, where a larger number of states is expected as predictions are more difficult.

Accuracy of the Various ACR Approaches

The results are displayed in figure 2 (4 states) and figure 3 (20 states) for the most relevant simulation conditions. Additional results are provided in supplementary figures S1–S5, Supplementary Material online, for all simulation conditions and the root and edge predictions. We observe that

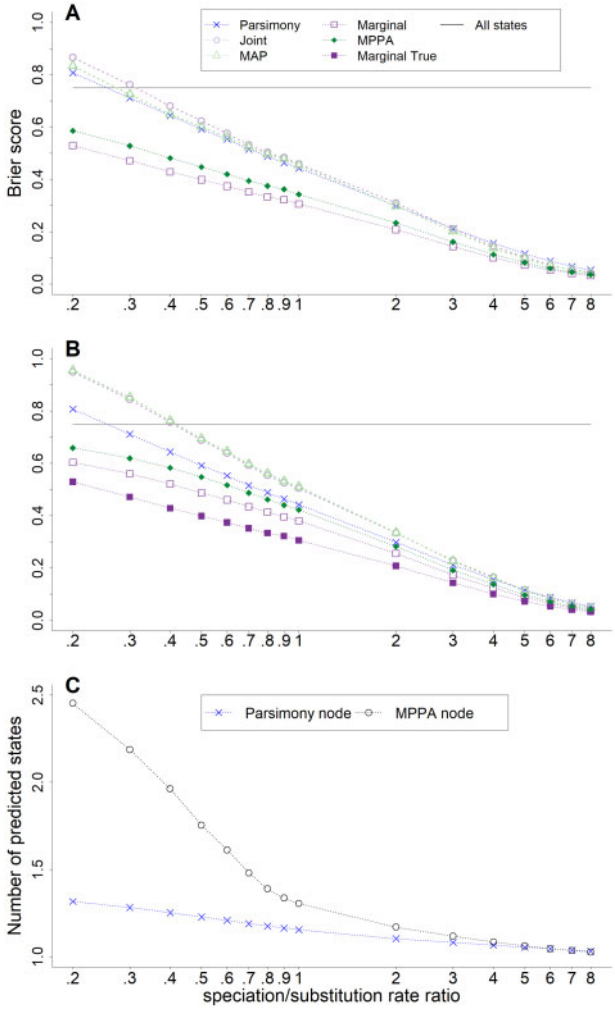

FIG. 2. Accuracy of ACR methods with 4-state DNA-like simulated data. *x* axis: speciation/substitution (or evolutionary) rate ratio. *y* axis: Brier score (the lower the better; panels *A* and *B*); Number of predicted states per node (panel *C*). (*A*) ML-based methods use the true model (HKY, with estimated parameters) and branch lengths (rescaled by PastML). (*B*) ML-based methods use the F81 model (with estimated equilibrium frequencies) and noisy branch lengths (rescaled by PastML). (*C*) Number of nodes predicted by parsimony and MPPA. "All states": all states are predicted with equal probability (=1/4). "Marginal True": best possible accuracy, obtained with Marginal using the true model and tree, as in panel *A*. See text for details.

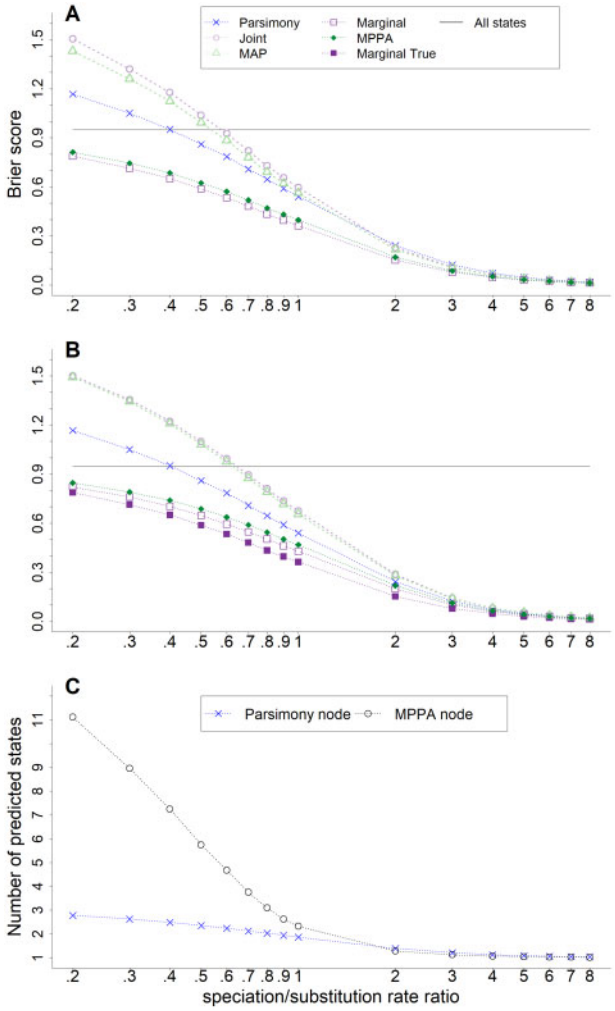

FIG. 3. Accuracy of ACR methods with 20-state protein-like simulated data. (*A*) True model (JTT) and tree (rescaled by PastML). (*B*) 20-state F81-like model (with estimated frequencies) and noisy branch lengths (rescaled by PastML). With "All states" every state is predicted with probability 1/20. See note to figure 2 and text for details.

- *Evolutionary rate*: As expected, predictions are very difficult with the lowest ρ value (speciation/evolutionary rate ratio, 0.2, i.e., 2.5 changes per branch in average); then, all methods have similar or worse accuracy to/than the agnostic method predicting all states with equal probability. With higher ρ values, predictions become easy, as few mutations occur in the tree. With the highest ρ values ($\geq$5.0, i.e., $\leq$0.1 changes/branch) all methods succeed (Brier score $\approx$0) and are equivalent.
- *Root prediction*: As expected, predicting the root is much more difficult (supplementary figs. S1 and S2, Supplementary Material online): when ρ is $\leq$0.5 ($\geq$1.0 changes/branch), even the best method (Marginal with the correct model) does not improve much over the agnostic method, and with the highest ρ value (8.0, easiest condition), the Brier score is still substantially larger than 0.
- *Number of states*: As expected again, predictions are more difficult with 20 states (fig. 3, Brier score $\approx$1.5 for the worst methods and conditions) than with 4 states (fig. 2, Brier score $\approx$0.95 for the worse methods and conditions). Moreover, the gap between the best and worse methods is larger with 20 states than with 4 states. However, the ranking of the various methods is the same in both settings, and this holds for root prediction as well.
- *Ranking of ACR methods*: Marginal is the best method regarding the Brier score, as expected with decision theory, and its advantage still holds with model violations (figs. 2 and 3 and supplementary fig. S5, Supplementary Material online). Joint and MAP are the worst, due to the fact that they predict a unique state and do not account

for uncertainty. Their accuracies are similar, with a slight advantage for MAP in certain conditions (e.g., in the absence of model violations, figs. 2A and 3A). This result still holds with the edge Brier score (supplementary figs. S3 and S4, Supplementary Material online), thus indicating again that Joint and global predictions have no advantage compared with the more local calculations of marginal posterior probabilities (still highly dependent for neighboring nodes). Moreover, the ranking of all methods with the edge Brier score is the same as that obtained with the node Brier score.

- *Multiple/single state predictions*: Thanks to multiple predictions in uncertain configurations, Parsimony has a clear advantage over Joint and MAP (figs. 2 and 3). The advantage of MPPA is even larger, due to the fact that MPPA predicts more states than Parsimony, and that these states are predicted using a rigorous probabilistic approach. With medium ρ value (1.0), the number of predicted state by MPPA is $\sim$1.3 and $\sim$3.0, for 4 and 20 states, respectively. This indicates that the large accuracy gain of MPPA, compared with unique state prediction methods (Joint, MAP), is obtained thanks to a relatively low number of predicted states, which eases the interpretability and visualization of the global scenarios returned by MPPA. However, in difficult conditions (low ρ values and/or root prediction, figs. 2 and 3 and supplementary figs. S1 and S2, Supplementary Material online) the number of states predicted by MPPA is larger (and much larger than parsimony's), indicating that MPPA "recognizes its ignorance." We shall see that these findings are confirmed with real data.
- *Model violations*: In terms of accuracy, MPPA is close to Marginal in all conditions, especially with 20 states, and is the second best method (figs. 2 and 3). Moreover, MPPA's accuracy remains nearly identical with noisy branch lengths, compared with true branch lengths (supplementary fig. S5, Supplementary Material online). When violations occur in the model describing state changes (i.e., F81-like models are used to analyze data simulated with HKY and JTT) we observe different results depending on the number of states. With 4 states, there is a low but visible gap between "Marginal true" and "Marginal F81," which added to the gap between MPPA and "Marginal F81" indicates that model violations have a substantial impact. However, MPPA is still better than Parsimony, especially in the difficult region (low ρ values). With 20 states, all these differences are low and the impact of model violations is negligible. This advocates for the use of F81-like models, especially with a large number of states, where estimating all the parameters of complex models is likely unfeasible (but see Lemey et al. 2014).

To summarize, MPPA performs well in this simulation study, with an accuracy close to the fully probabilistic Marginal method, but outputs that are much easier to interpret and visualize. Moreover, the F81-like model seems to be a relevant choice, especially with 20 states, as it yields accurate ancestral predictions while avoiding difficult estimations of the relative rates of changes from one state to another. Results of Oliva et al. (2019) confirm these findings. The method proposed in that study , named MPPE, shares common points with MPPA (e.g., ranking of the character states using their posteriors, and selection of the best k states) but uses a different selection criterion based on the expected classification error. The accuracies of MPPE and MPPA are very close, and clearly better than the accuracy of all other methods considered in that study.

Results: Phylogeography of Dengue Epidemics

To demonstrate the performance of PastML on real data, we reconstructed the ancestral history of Dengue serotype 2 (DENV2) epidemics. We used a medium-size data set of 356 sequences, obtained from Ayres et al. (2019). This data set is annotated with: sampling years (between 1944 and 2014); countries, which we grouped in 10 regions (Central America, Caribbean, South America, Europe, Africa, Western Asia, Eastern Asia, South-eastern Asia, and Oceania); and genotypes (13 sequences of Sylvatic lineage, and 343 sequences of 5 endemic genotypes: American, Asian-American, Asian I, Asian II, and Cosmopolitan). We inferred a ML tree from the DNA sequences, and dated and rooted this tree (based on dates) using LSD (To et al. 2016). To check the robustness of PastML inferences against phylogenetic uncertainty, the tree reconstruction was performed with three ML tools: RAxML (Stamatakis 2014), PhyML (Guindon et al. 2010), and FastTree (Price at al. 2010), resulting in three trees with substantial topological differences (mean normalized bipartition distance $\sim$10%). However, the global topological information was preserved (e.g., the genotypes and sylvatic lineage were perfectly identified and supported in the three trees). To further study the impact of poorly supported branches and topological uncertainties, we created a fourth tree by collapsing the 40 poorly supported branches of the PhyML tree (SH-like support $<$50%). This collapsed tree was dated and rooted as the three others. The phylogeography of DENV2 epidemics was reconstructed from these four trees and location annotations using PastML with default options (MPPA with F81-like model). We also checked the robustness of ACR results regarding state sampling variations. For this purpose, we generated five new DENV2 alignments, each by picking 356 sequences randomly with replacement from the original alignment. This way we obtained five randomized alignments of the same size as the original one, but with some sequences removed and some present multiple times, which in turn perturbed the numbers of samples per location. We then reconstructed the phylogeography of these five resampled data sets using RaxML and the same approach as for the original alignment.

PastML results with the RAxML tree are shown in figure 4. The location of the root is unresolved between five regions: the Caribbean, Eastern Asia, Oceania, South America, and South-eastern Asia. Provided that the root represents the common ancestor of the sylvatic and endemic/epidemic strains (dated 1209 [1147–1249]), this result is not surprising.

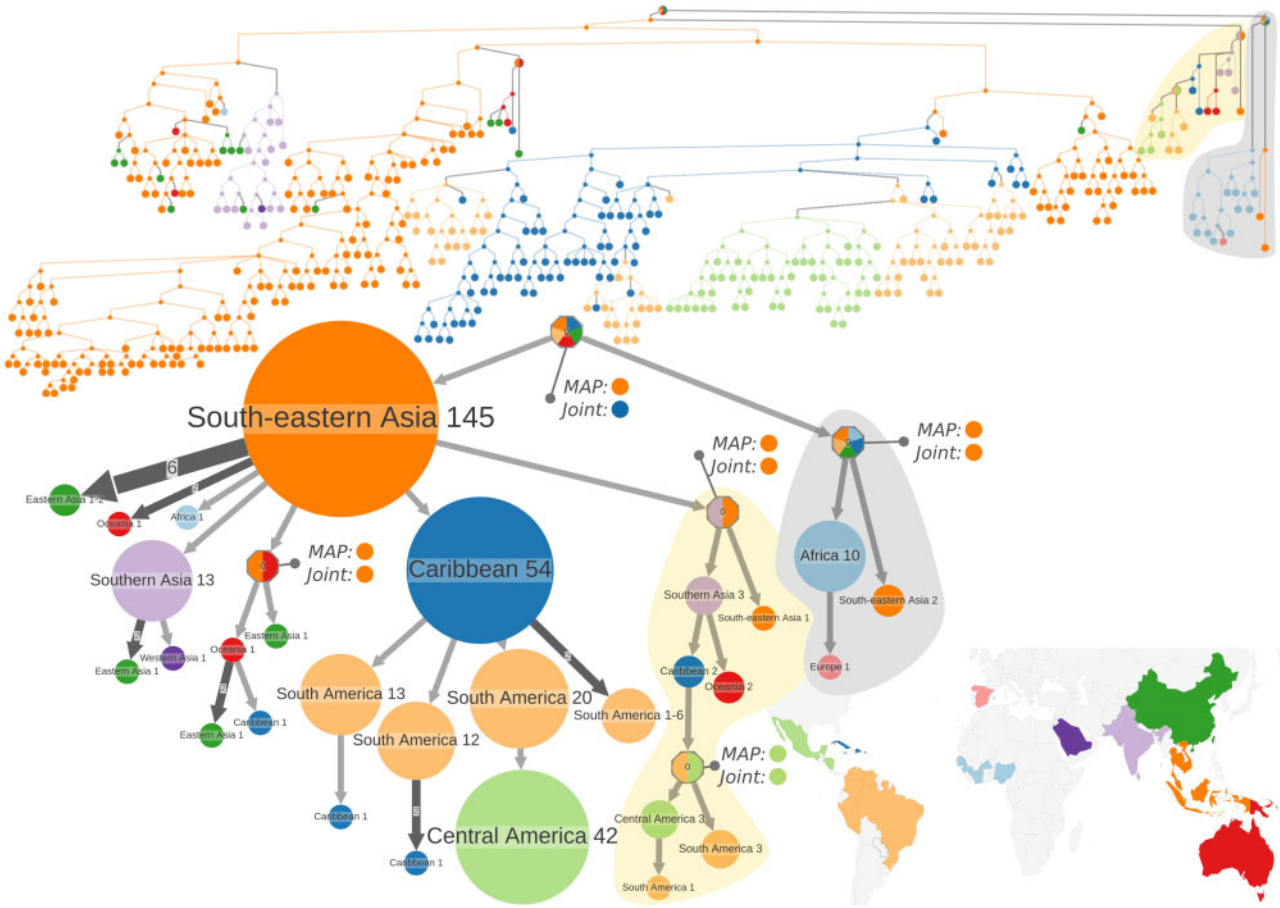

FIG. 4. Ancestral reconstruction of DENV2 epidemic locations. The figure shows the full tree (top) and compressed (bottom left) visualizations produced by PastML using MPPA with an F81-like model. Different colors correspond to different geographical regions as shown in the map in the bottom right corner. The sylvatic and American genotype subtrees are shown with grey and light-yellow backgrounds, respectively. The Joint and MAP predictions are shown for the uncertain nodes. MAP and Joint disagree on the tree root, but their predictions are included in MPPA predictions for all nodes (including those with unique MPPA prediction).

If instead we focus on the endemic/epidemic subtree (left subtree of the root), its root (dated 1744 [1723–1759]) is resolved to South-eastern Asia. The South-eastern Asian endemic/epidemic cluster has grown over years and by 2014 contained 145 strains from our data set (large orange node on the compressed representation). From there, the epidemic was spread to Southern Asia (lilac), Eastern Asia (six independent introductions represented as an edge of size six leading to a green node), Oceania (red), Africa (light blue), and the Caribbean (blue). From Caribbean, it was further spread to South America (multiple introductions leading to clusters of various sizes in light-orange), and from there to Central America (light-green). These findings agree with the study of global Dengue type 2 phylogeography by Walimbe et al. (2014) performed using a Bayesian approach on 307 E-gene sequences (sampled between 1944 and 2011). They also could not pinpoint the ancestral location for DENV2 sylvatic and endemic/epidemic strains, and detected South-eastern Asia as the origin for Asian, Asian-American, and Cosmopolitan endemic genotypes. They also found multiple migrations from Caribbean countries to the American mainland. However, in some cases, PastML predictions remain more cautious. For example, for sylvatic strains Walimbe et al. predicted West Africa as the most probable ancestral location, whereas PastML hesitates between five possible locations, including Africa (gray subtree). For the American endemic genotype (light-yellow subtree), Walimbe et al. predicted India to be the ancestral location, whereas PastML hesitates between Southern (including India) and South-eastern Asia.

PastML compressed scenarios for RAxML, FastTree, PhyML, and PhyML-collapsed trees are almost identical, with a few minor differences that can be mostly eliminated by resolving some unresolved nodes (supplementary fig. S6, Supplementary Material online). This illustrates the robustness of PastML against phylogenetic uncertainty. The same holds regarding sampling variations: PastML ancestral scenarios with resampled data sets are very similar to the scenario inferred using the original alignment (supplementary fig. S7, Supplementary Material online). Figure 4 shows the advantage of the approach, where uncertain inferences are not solved arbitrarily but made explicit. For example, the MAP and Joint inferences for the tree root are Eastern Asia and South-eastern Asia, respectively, whereas MPPA predicts five possible origins including both MAP and Joint predictions. However, for most of the nodes MPPA predicts a unique

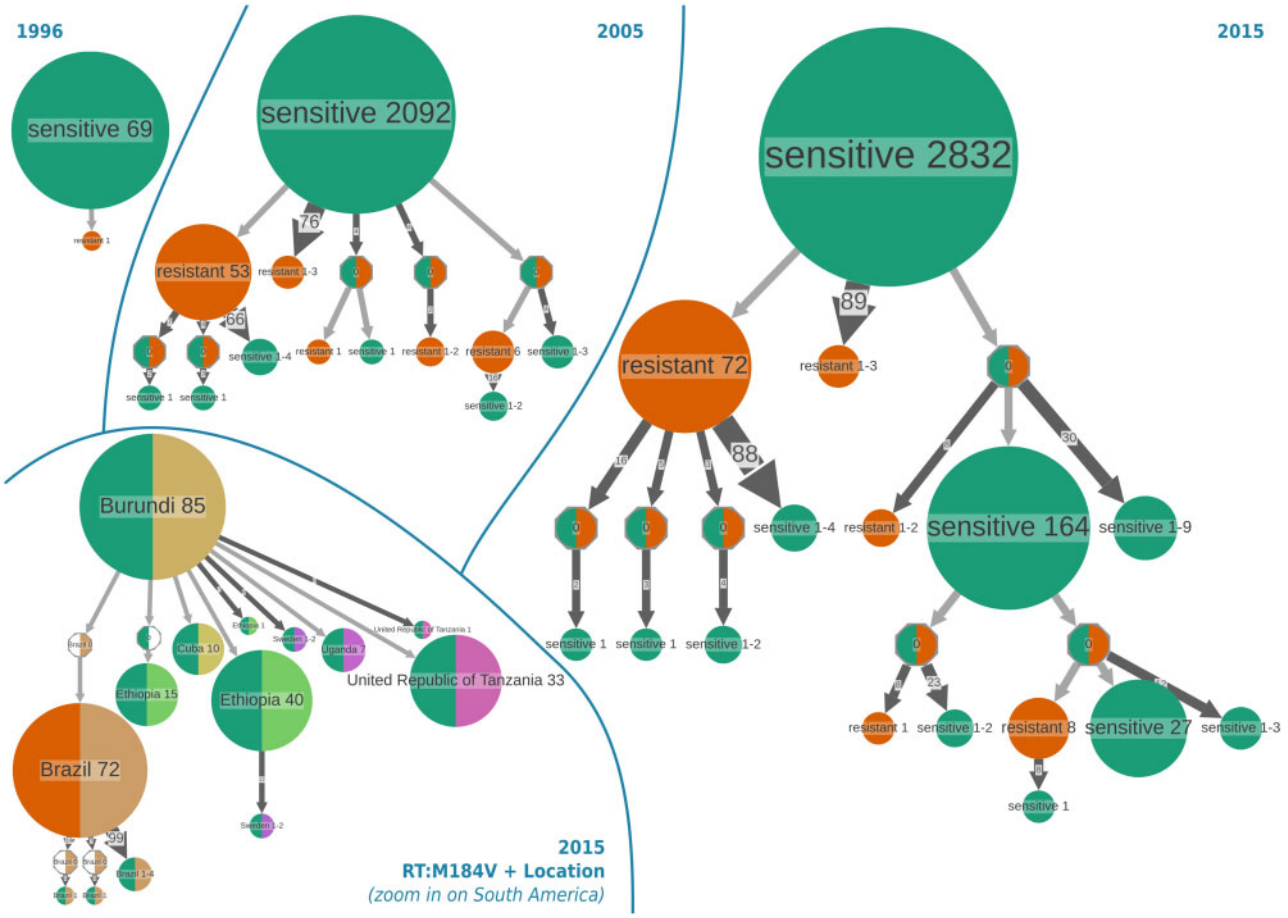

FIG. 5. Ancestral state reconstruction of the presence/absence of DRM M184V over time (top), and combined with location data (bottom left). The reconstruction was done by PastML with MPPA + F81 option. For the timeline, the tree was pruned at each year to remove the tips sampled after that year. In the bottom left panel, M184V presence/absence is combined with location data: M184V state is shown color-coded in the left half of each node (green when mutation is absent, and orange for resistant strains), countries are color-coded in the right half of each node, and shown in the labels.

location (average number of states per node ~1.03). The log-likelihood of the Map, Joint, and MPPA scenarios is equal to −197.3, −197.2, and −193.8, respectively (as expected Joint is better than MAP, and MPPA is even better as it includes several states for some of the nodes). PastML thus represents a larger fraction of the data, while producing a scenario which is almost fully resolved.

Results: Drug Resistance Mutations in HIV

To demonstrate the performance of PastML with large data sets, we studied the emergence, transmission and reversion of drug resistance mutations (DRMs, Bennett et al. 2009) in HIV. DRMs emerge under the pressure of drug treatments, and then may be transmitted to drug naïve patients (Zhukova et al. 2017). An essential public health issue is to detect potential drug resistant subepidemics, which could become prevalent and pose major problems, as is already the case with other pathogens and diseases (e.g., malaria). Parsimony-based ancestral reconstructions were already used fruitfully in this context, with patients from the UK (Mourad et al. 2015). We focused here on the subtype C of HIV1 (HIV1-C), the most prevalent subtype around the world (~50% of HIV infections), originating from Central Africa, spread in Southern and Eastern Africa, and then in Europe and the Americas, with multiple introductions (Vidal et al. 2000; Hemelaar 2012; Faria et al. 2014).

We used a large data set of 3,619 HIV-1C *pol* sequences, obtained from Jung et al. (2012), Chevenet et al. (2013), and the latest (2017) *pol* alignment of the Los Alamos HIV database. This data set is annotated with sampling dates and countries. We built a tree from the DNA sequences using PhyML (Guindon et al. 2010), and rooted it using non-C sequences. As in Mourad et al. (2015), we first computed the presence of the studied DRM in the sequences, thus obtaining the value of a 2-state character (sensitive/resistant) for every tree tip. Then, we used PastML with default options to reconstruct the ancestral resistance status of the tree nodes and root. We also performed analyses through time, to study the dynamics of DRM emergence, diffusion and reversion, and combined DRM analyses with phylogeography to reveal the spatial propagation of resistances. Five of the most prevalent DRMs were analyzed.

Results for M184V (the most prevalent DRM in our data set) and the combination with phylogeography for the largest resistance cluster are displayed in figure 5. M184V is a major

nonnucleoside RT inhibitor (NRTI) mutation selected in patients receiving Lamivudine (3TC) and Emtricitabine (FTC) (Gallant 2006). 3TC was approved for medical use in the United States in 1995, and FTC in 2006. They are both used worldwide nowadays. According to the study of Castro et al. (2013) on the persistence of DRMs, in the absence of drug-selective pressure M184V is lost relatively quickly (median time to loss ∼1.0 [0.5–2.0] years).

Ancestral state reconstruction allows us to detect potentially acquired (ADR) and transmitted drug resistance (TDR) patterns. An acquired drug resistance is represented by a single-patient resistant node in the compressed visualization, which implies a state change from a sensitive parent node. Potential TDRs are represented by cluster(s) of resistant patients, where internal edges correspond to transmissions of resistant strains. Note that these simple statements still hold with incomplete sampling (Mourad et al. 2015): transmissions of DRMs within resistant clusters are then indirect, whereas a one-patient resistant node may correspond to a small resistance cluster, the root of which acquired the DRM.

We analyzed the reconstructed transmission tree at different time points, each time pruning the tree to remove the tips and nodes sampled after the corresponding year. Figure 5 shows the results for 1996 (the sampling year corresponding to the first sequences with M184V in our data set), 2005 and 2015 (last sampling year in our data set). We see the emergence and growth of potential TDR clusters over time. In 2005, the main configurations included a major TDR cluster of 53 patients, and 76 cases of independent DRM emergence (ADR, i.e., 1-patient node, or small resistant clusters of 2–3 patients). There are also multiple cases of reversion of the DRM (e.g., 66 cases of patients having a sensitive virus at the time of sampling, which originate from the major resistant cluster). By 2015 the main TDR cluster grew (to 72 patients), and so did the numbers of cases of ADR and DRM reversion. Importantly, the main resistance cluster in 2005 is included in the one of 2015, which demonstrates the potential of the method in surveying the emergence of problematic resistant subepidemics, as the 2015 cluster was already predictable in 2005.

To further investigate the largest TDR cluster, we combined the ancestral state reconstruction for M184V with the location. The result located the whole resistant cluster in South America. We then increased the geographical resolution by replacing the regions with the countries and focused on the subtree with the root in East Africa, as this is from where the virus was spread to South America according to our reconstruction. The results are shown in the bottom left panel of figure 5, and suggest that the resistance cluster is located in Brazil, and that it originated from either a sensitive or a resistant case in Brazil (the parent node of the TDR cluster is a Brazilian node with unresolved M184V state). The reconstruction also shows that the virus was introduced to Brazil from Burundi, from where it was also spread to Tanzania, and Ethiopia. This geographical result agrees with the study on HIV-1C epidemics in Eastern Africa and Southern Brazil by Mir et al. (2018), although the latter was performed using a Bayesian approach and a different data set, which included more Brazilian sequences.

The results for the second, third, and fourth most prevalent DRMs in our data set (K103N, D67N, and K70R) are similar to those for M184V: they show emergence and growth of TDR clusters over time, and well as growth of the number of ADR and reversions to sensitive state. The largest TDR clusters for these three DRMs are located in Brazil, just as with M184V. With the decrease of DRM prevalence, the size of TDR clusters in our data decreases, from a 72-patient TDR cluster for M184V to a 7-patient one for K70R. The analysis of K103N can be found in supplementary figure S8, Supplementary Material online. The results for the fifth most prevalent DRM (Y181C) are different: We hardly see any TDR clusters (their size is at most four patients), and the largest TDR cluster is located in India (supplementary fig. S9, Supplementary Material online). This could be the very start of TDR spread for this mutation hence making it a candidate for closer surveillance, or it could simply be due to the quick reversion time of Y181C (median of 1.3 years, cf., Castro et al. [2013]) and hence inability to form TDR clusters.

Figure 5 and the results illustrate again the advantage of MPPA reconstructions. For example, the Joint and MAP predictions for the root of the Brazilian subtree (unresolved by MPPA) disagree, they are respectively resistant and sensitive. The same occurs with 34 other nodes, where MPPA predictions contain again both states predicted by MAP and Joint. In total, ∼2.5% of the tree nodes cannot be fully resolved by MPPA, which corresponds to ∼1.025 states per node in average. The total log-likelihood of MPPA is equal to −834.9, whereas the log-likelihood of Joint is −861.8, and of MAP is −872.4 (as expected substantially lower than Joint). MPPA thus represents a much larger fraction of the data, while producing a scenario which is almost fully resolved.

PastML analyses were performed on a laptop with a 4-core 2.50GHz CPU. ACR and visualization of Dengue phylogeography (356 tips, 10 states) as well as of HIV-1C DRM dynamics (3,619 tips, 2 states) took ∼1 min per tree.

Conclusions

We presented a new, simple, and fast approach to reconstruct ancestral scenarios, deal with the uncertainty of ancestral inferences in the difficult regions of the tree (typically around the tree root), and visualize and edit interactively a tree-shaped graphical representation of the most likely ancestral scenarios. The results obtained with the Dengue and a large HIV data sets are congruent with previous studies. Moreover, these results are robust against phylogenetic uncertainty and sampling variations.

The method proposed here to account for uncertainties in ancestral inferences could easily be adapted to sequences and especially proteins, where its ability to provide an inventory of a limited number of alternative amino acids should be very useful to examine likely variants and determine the best candidate based on the structure and physicochemical properties. Directions for further research include the exploration of other loss functions, in place of the Brier scoring rule, for

example in line of Oliva et al. (2019) or reusing previous concepts and results from decision theory (Gneiting and Raftery 2007). Moreover, the current version of PastML is based on JC-like and F81-like models, due to the limited information available with unique characters. Some refinements (e.g., in the line of Lemey et al. [2014], and Dudas et al. [2017]) should be useful, not only to improve the accuracy and ancestral reconstructions but also to provide users with a global view of the evolutionary processes at stake (strain flow between regions and countries, acquisitions and losses of molecular characters, dynamics of ecological character changes, etc.). Lastly, an interesting direction would be to develop methods and tools to compare our tree-shaped, compressed representations of ancestral scenarios, calculate some distance between two scenarios, extract the common parts and the differences, and propose some consensus.

Materials and Methods

Computation of the Marginal Posterior Probabilities

Let N be a given internal node of the tree T, and U, V, and F be the left descendant, right descendant, and father of N, respectively, with corresponding rescaled branch lengths denoted as u, v, and f. Moreover, let $\text{Down}(N)$ be the vector of state conditional likelihoods induced by the state values of the tips of the "*down*" subtree rooted with N. $\text{Down}(N,i)$ is equal to the likelihood of having state i in N given the states observed in the extant descendants of N. $\text{Down}(N)$ is computed recursively using the pruning algorithm (Felsenstein 1981), which combines a postorder tree traversal with the following formula:

$$\text{Down}(N,i) = \Big[\sum_j \text{PC}(i \to j/u)\text{Down}(U,j)\Big] \times \Big[\sum_j \text{PC}(i \to j/v)\text{Down}(V,j)\Big],$$

$$\text{and for a tip } l: \text{ if } c(l) = i \text{ or } X, \text{ then } \text{Down}(l,i) = 1, \text{ else } \text{Down}(l,i) = 0.$$

This algorithm proceeds in a bottom-up fashion, first computing the conditional likelihoods of the nodes close to the tips and progressing until the tree root. The conditional likelihoods so obtained can be used to compute marginal posterior probabilities and then predict the ancestral states attached to every tree node. Several ancestral reconstruction programs use this approach. However, a more accurate method does exist (Yang 2007; Gascuel and Steel 2014). Indeed, when using $\text{Down}(N)$ we only account for the information contained in the tips descending from N, and not for the information contained in the rest of the tree.

To account for all tree information, we define a second vector of conditional likelihoods attached to N, $\text{Up}(N)$, where $\text{Up}(N,i)$ denotes the conditional likelihood of having i in N given the tip values observed in the "*up*" subtree of N. To define this subtree, assume that T is rerooted with N; then N has three direct descendants: U, V, and F, each associated with a subtree. The *up* subtree of N is defined as the subtree associated with F including the branch (of length f) from F to N. In other words, the *up* subtree contains all branches, nodes and tips which are not included in the *down* subtree of N. To compute the *Up* conditional likelihoods we use the following formula (applied to node U to simplify the notation, but the same formula applies to V, N and all tree nodes):

$$\text{Up}(U,i) = \Big\{\sum_j \text{PC}(i \to j/u)\text{Up}(N,j) \times \Big[\sum_k \text{PC}(j \to k/v)\text{Down}(V,k)\Big]\Big\}.$$

This formula is exploited recursively thanks to a top-down, preorder tree traversal. We start from the tree root R, having $\text{Up}(R,i) = 1$ for all states i, and progress toward the tips; for example, $\text{Up}(N)$ is computed after $\text{Up}(F)$ and before $\text{Up}(U)$ and $\text{Up}(V)$, as seen in the formula. Moreover, this formula uses the *Down* likelihoods, which have to be computed first. Both *Down* and *Up* calculations are easily extended to polytomies: the *Down* formula contains as many sum terms as N has descendants (instead of 2 above with U and V); the *Up* formula contains as many internal sum terms as U has brother nodes (instead of 1 above with V).

Once $\text{Down}(N)$ and $\text{Up}(N)$ have been computed, the state marginal posterior probabilities of N are computed using (remember that for the tree root $\text{Up}(R,i) = 1$):

$$\text{Marginal}(N,i) = \frac{\pi_i \text{Down}(N,i)\text{Up}(N,i)}{\text{TotalProba}(N)}, \text{ where}$$

$$\text{(law of total probability): } \text{TotalProba}(N) = \sum_{j \in S} \pi_j \text{Down}(N,j)\text{Up}(N,j).$$

These algorithms have a time complexity in $O(ns^2)$, where n is the number of tree tips and s the number of states. The whole procedure is thus linear in n and remarkably fast.

All along these calculations, some of the conditional likelihood values may be extremely small when n is large, and smaller than the minimum value permitted for double floating numbers. As in other ML programs, if the conditional likelihoods of N are smaller than a given threshold, then all conditional likelihoods of N are multiplied by a power of 2. This numerical trick does not change the marginal posterior probabilities, as the relative values of the conditional likelihoods are preserved.

DENV2 Data and Analyses

We used a data set of 356 aligned sequences of Dengue serotype 2, obtained from Ayres et al. (2019). We built ML trees from the DNA sequences using three inference tools: RAxML v8.2.11-sse3 (Stamatakis 2014), PhyML v3.3.20180621 (Guindon et al. 2010), and FastTree v2.1.10 (Price at al. 2010), all with GTR + Γ6 substitution model and default options. The trees were then dated and rooted (based on dates) using LSD v0.3beta-55183ca9d0 (To et al. 2016). PastML was used with default options (MPPA+F81) to

reconstruct the ancestral locations of all tree nodes and root, among ten regions (character states) present in the data set.

HIV-1C Data and Analyses

We used an HIV-1C *pol* sequence data set, previously used in Jung et al. (2012), and then (in a slightly updated version) in Chevenet et al. (2013). We extended the latter alignment with HIV-1C *pol* sequences from the latest (2017) *pol* alignment in the Los Alamos HIV database (https://www.hiv.lanl.gov/content/index; Last accessed June 1, 2019), hence adding 583 sequences. Addition of the new sequences was performed using MAFFT multiple sequence alignment program with the –add option (Katoh and Standley 2013). The final alignment contains 3,619 HIV-1C *pol* sequences, plus 35 outgroup reference sequences from the non-C subtypes. The data set is annotated with sampling dates and countries.

We detected the Surveillance Drug Resistance Mutations (Bennett et al. 2009), using the Sierra web service of the Stanford HIV drug resistance database (Liu and Shafer 2006). We removed the Surveillance Drug Resistance Mutation positions from the alignment and reconstructed five most parsimonious trees using TNT (Goloboff and Catalano 2016), which were used as starting trees for five runs of PhyML (Guindon et al. 2010) with GTR+I + Γ6 substitution model and aLRT SH-like branch supports. The most likely tree was retained for further analyses and ancestral reconstructions. This tree was rooted with the outgroup sequences, which were subsequently removed from the tree. The branches of length zero and aLRT SH-like support <0.5 were collapsed into polytomies.

PastML was then used to reconstruct, along this tree, the ancestral scenarios describing the emergence, diffusion, and reversion in some cases, of DRMs. We analyzed DRMs with high prevalence in our data set: M184V with a prevalence of 0.07 (highest prevalence in our data set), K103N (prevalence $= 0.05$, second highest prevalence), and Y181C (prevalence $=$ 0.03, fifth highest prevalence). Results (available at https://pastml.pasteur.fr) for the third and fourth highest prevalence DRMs are similar to those of M184V and K103N. PastML was used with MPPA + F81 and a node importance threshold of 10. We performed analyses through time to study the dynamics of DRM emergence, diffusion and reversion. In this context, we have two character states: the DRM is absent or present, and the corresponding strain (tip, node) is sensitive or resistant, respectively. Results for the most prevalent DRM (M184V) are provided in figure 4. Results for the two other DRMs are in supplementary figures S8 and S9, Supplementary Material online.

Data Availability

All of our data, trees, ACRs, and Snakemake pipelines (Köster and Rahmann 2012) used to reconstruct the trees and analyze them are available from https://pastml.pasteur.fr; Last accessed June 1, 2019.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Sincere thanks to L. Blassel, S. Cosentino, S. Duchêne, F. Lemoine, M. Matsui, H. Ménager, S. Mestack, and K. Theys for their help and comments. We also thank Tal Pupko and two anonymous reviewers for their suggestions to improve the first submitted manuscript. This work was supported by the EU-H2020 Virogenesis project (Grant No. 634650, O.G.), by the INCEPTION project (PIA/ANR-16-CONV-0005, A.Z. and O.G.), and by Postdoctoral Fellowship and KAKENHI 282725 (S.A.I.), 16H06279 (W.I.), and 16H06154 (W.I.) from Japan Society for the Promotion of Science.

References

Arbogast BS. 2001. Phylogeography: the history and formation of species. *Am Zool*. 41(1):134–135.

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA. 2019. BEAGLE 3: improved usability for a high-performance computing library for statistical phylogenetics, *Syst Biol.* (Advanced Access, Published: 23 April 2019, https://doi.org/10.1093/sysbio/syz020).

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, et al. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol*. 61(1):170–173.

Beaulieu JM, O'Meara BC, Donoghue MJ. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Syst Biol*. 62(5):725–737.

Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, et al. 2009. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* 4(3):e4724.

Bickelmann C, Morrow JM, Du J, Schott RK, Hazel I, Lim S, Müller J, Chang BS. 2015. The molecular origin and evolution of dim-light vision in mammals. *Evolution* 69(11):2995–3003.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev*. 78(1):1–3.

Busch F, Rajendran C, Heyn K, Schlee S, Merkl R, Sterner R. 2016. Ancestral tryptophan synthase reveals functional sophistication of primordial enzyme complexes. *Cell Chem Biol*. 23(6):709–715.

Byrd RH, Lu P, Nocedal J, Zhu C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 16(5):1190–1208.

Castro H, Pillay D, Cane P, Asboe D, Cambiano V, Phillips A, Dunn DT, Aitken C, Asboe D, Webster D, et al. 2013. Persistence of HIV-1 transmitted drug resistance mutations. *J Infect Dis*. 208(9):1459–1463.

Chevenet F, Jung M, Peeters M, De Oliveira T Gascuel, O. 2013. Searching for virus phylotypes. *Bioinformatics* 29(5):561–570.

Collins TM, Wimberger PH, Naylor GJ. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst Biol*. 43(4):482–496.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*. 29(8):1969–1973.

Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, Park DJ, Ladner JT, Arias A, Asogun D, et al. 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 544(7650):309.

Durães-Carvalho R, Salemi M. 2018. In-depth phylodynamics, evolutionary analysis and in silico predictions of universal epitopes of Influenza A subtypes and Influenza B viruses. *Mol Phylogenet Evol*. 121:174–182.
Edwards CJ, Suchard MA, Lemey P, Welch JJ, Barnes I, Fulton TL, Barnett R, O'Connell TC, Coxon P, Monaghan N. 2011. Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr Biol*. 21(15):1251–1258.
Endress PK, Doyle JA. 2009. Reconstructing the ancestral angiosperm flower and its initial specializations. *Am J Bot*. 96(1):22–66.
Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, et al. 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346(6205):56–61.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer.
Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. 2016. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32(2):309–311.
Gallant JE. 2006. The M184V mutation: what it does, how to prevent it, and what to do with it when it's there. *AIDS Reader* 16(10):556–559.
Gascuel O, Steel M. 2014. Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Syst Biol*. 63(3):421–435.
Gascuel O, Steel M. 2018. A Darwinian uncertainty principle. *BioRxiv*. doi:https://doi.org/10.1101/506535.
Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*. 102(477):359–378.
Goloboff PA, Catalano SA. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* 32(3):221–238.
Gräf T, Vrancken B, Junqueira DM, de Medeiros RM, Suchard MA, Lemey P, de Matos Almeida SE, Pinto AR. 2015. Contribution of epidemiological predictors in unraveling the phylogeographic history of HIV-1 subtype C in Brazil. *J Virol*. 89(24):12341–12348.
Guiasu S. 1977. Information theory with applications. New York: McGraw-Hill.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.
Hanson-Smith V, Kolaczkowski B, Thornton J. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol*. 27(9):1988–1999.
Hasegawa M, Kishino H, Yano TA. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22(2):160–174.
Hemelaar J. 2012. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med*. 18(3):182–192.
Holmes EC, Dudas G, Rambaut A, Andersen KG. 2016. The evolution of Ebola virus: insights from the 2013–2016 epidemic. *Nature* 538(7624):193.
Huelsenbeck JP, Bollback JP. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol*. 50(3):351–366.
Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. *Syst Biol*. 52(2):131–158.
Iwasaki W, Takagi T. 2007. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics* 23(13):i230–i239.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8(3):275–282.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. *Mamm Protein Metab*. 3(21):132.
Jung M, Leye N, Vidal N, Fargette D, Diop H, Toure Kane C, Gascuel O, Peeters M. 2012. The origin and evolutionary history of HIV-1 subtype C in Senegal. *PLoS One* 7(3):e33579.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.
Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28 (19):2520–2522.
Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *J R Soc Interface* 11(94):20131106.
Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Mol Biol Evol*. 33(8):2102–2116.
Lambert A, Alexander HK, Stadler T. 2014. Phylogenetic analysis accounting for age-dependent death and sampling with applications to epidemics. *J Theor Biol*. 352:60–70.
Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, Russell CA, Smith DJ, Pybus OG, Brockmann D, et al. 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog*. 10(2):e1003932.
Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLoS Comput Biol*. 5(9):e1000520.
Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 44(W1):W242–W245.
Leventhal GE, Günthard HF, Bonhoeffer S, Stadler T. 2014. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol Biol Evol*. 31(1):6–17.
Lewis NS, Verhagen JH, Javakhishvili Z, Russell CA, Lexmond P, Westgeest KB, Bestebroer TM, Halpin RA, Lin X, Ransier A, et al. 2015. Influenza A virus evolution and spatio-temporal dynamics in Eurasian wild birds: a phylogenetic and phylogeographical study of whole-genome sequence data. *J Gen Virol*. 96(8):2050–2060.
Liu TF, Shafer RW. 2006. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis*. 42(11):1608–1618.
Maddison WP, Maddison DR. 2000. MacClade, version 4.0. Sunderland (MA): Sinauer.
Magee D, Suchard MA, Scotch M. 2017. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. *PLoS Comput Biol*. 13(2):e1005389.
Maor R, Dayan T, Ferguson-Gow H, Jones KE. 2017. Temporal niche expansion in mammals from a nocturnal ancestor after dinosaur extinction. *Nat Ecol Evol*. 1(12):1889–1895.
Marazzi B, Ané C, Simon MF, Delgado-Salinas A, Luckow M, Sanderson MJ. 2012. Locating evolutionary precursors on a phylogenetic tree. *Evolution* 66(12):3918–3930.
Mir D, Gräf T, Esteves de Matos Almeida S, Pinto AR, Delatorre E, Bello G. 2018. Inferring population dynamics of HIV-1 subtype C epidemics in Eastern Africa and Southern Brazil applying different Bayesian phylodynamics approaches. *Sci Rep*. 8(1):8778.
Mooers AØ, Schluter D. 1999. Reconstructing ancestor states with maximum likelihood: support for one-and two-rate models. *Syst Biol*. 48(3):623–633.
Mourad R, Chevennet F, Dunn DT, Fearnhill E, Delpech V, Asboe D, Gascuel O, Hue S. 2015. A phylotype-based analysis highlights the role of drug-naive HIV-positive individuals in the transmission of antiretroviral resistance in the UK. *AIDS* 29(15):1917–1925.
Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol*. 51(5):729–739.
Oliphant TE. 2007. Python for scientific computing. *Comput Sci Eng*. 9(3):10–20.
Oliva A, Pulicani S, Lefort V, Brehelin L, Gascuel O, Guindon S. Forthcoming 2019. Accounting for ambiguity in ancestral sequence reconstruction. *Bioinformatics*. Advanced Access, Published: 12 April 2019, https://doi.org/10.1093/bioinformatics/btz249

Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst Biol*. 48(3):612–622.
Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*. 53(5):673–684.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9590.
Pupko T, Pe I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*. 17(6):890–896.
Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13(3):235–238.
Ratmann O, Hodcroft EB, Pickles M, Cori A, Hall M, Lycett S, Colijn C, Dearlove B, Didelot X, Frost S, et al. 2017. Phylogenetic tools for generalized HIV-1 epidemics: findings from the PANGEA-HIV methods comparison. *Mol Biol Evol*. 34(1):185–203.
Ree RH, Smith SA. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Syst Biol*. 57(1):4–14.
Sagulenko P, Puller V, Neher RA. 2018. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol*. 4(1):vex042.
Sauquet H, von Balthazar M, Magallón S, Doyle JA, Endress PK, Bailes EJ, Barroso de Morais E, Bull-Hereñu K, Carrive L, Chartier M, et al. 2017. The ancestral flower of angiosperms and its early diversification. *Nat Commun*. 8:16047.
Stadler T, Bonhoeffer S. 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos Trans R Soc B Biol Sci*. 368(1614):20120198.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
Steel M, Mooers A. 2010. The expected length of pendant and interior edges of a Yule tree. *Appl Math Lett*. 23 (11):1315–1319.
Swofford DL, Maddison WP. 1987. Reconstructing ancestral character states under Wagner parsimony. *Math Biosci*. 87(2):199–229.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci*. 17(2):57–86.
To HT, Jung M, Lycett S, Gascuel O. 2016. Fast dating using least-squares criteria and algorithms. *Syst Biol*. 65(1):82–97.
Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B, Delaporte E. 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J Virol*. 74(22):10498–10507.
Walimbe AM, Lotankar M, Cecilia D, Cherian SS. 2014. Global phylogeography of Dengue type 1 and 2 viruses reveals the role of India. *Infect Genet Evol*. 22:30–39.
Wallace R, Hodac H, Lathrop R, Fitch W. 2007. A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci U S A*. 104(11):4473–4478.
Werner GD, Cornwell WK, Sprent JI, Kattge J, Kiers ET. 2014. A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. *Nat Commun*. 5:4087.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol*. 44(1):139–146.
Zhu C, Byrd RH, Lu P, Nocedal J. 1997. Algorithm 778: l -BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans Math Softw*. 23(4):550–560.
Zhukova A, Cutino-Moguel T, Gascuel O, Pillay D. 2017. The role of phylogenetics as a tool to predict the spread of resistance. *J Infect Dis*. 216(Suppl 9):S820–S823.