A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios

Sohta A. Ishikawa†, Anna Zhukova†, Wataru Iwasaki, Olivier Gascuel*
* olivier.gascuel@pasteur.fr

Supplementary Material

Data availability

All the data used and produced as well our analysis pipelines are available for download at https://pastml.pasteur.fr.

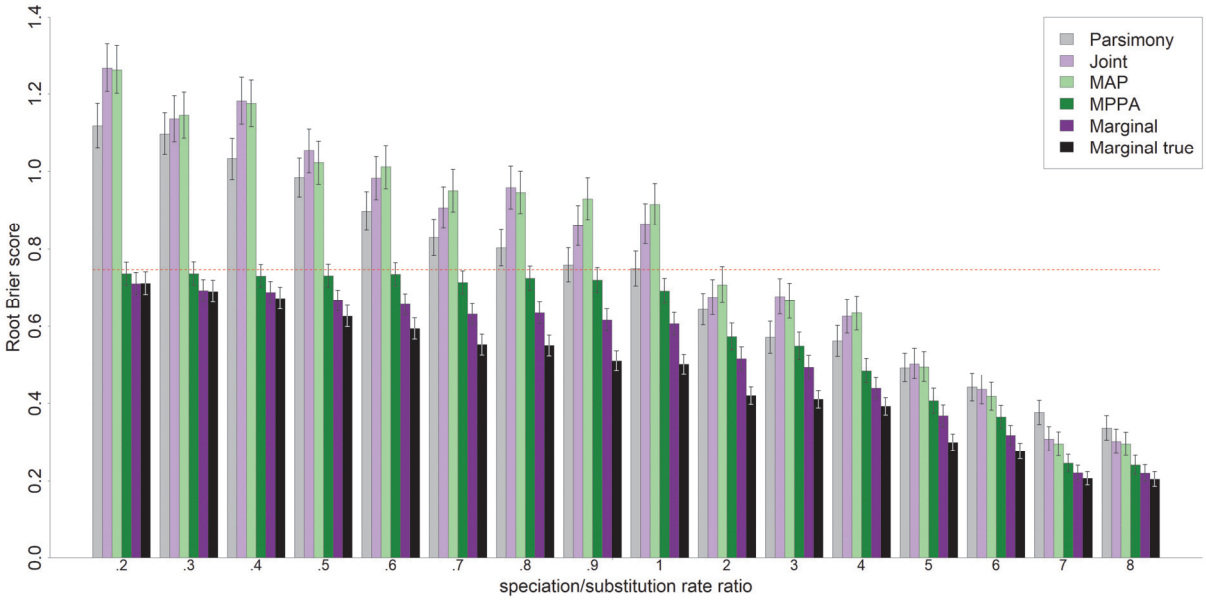

Figure S1. ACR accuracy comparison: tree root with 4-state DNA-like data. Predicting the root state is more difficult than predicting any given node in the tree, especially those that are close to the tips (Gascuel and Steel 2014). We assess here the accuracy of the various methods in this task and their ranking, using the same simulation setting as in Figure 2B (F81-like model and noisy branch lengths). X-axis: speciation/substitution rate ratio; Y-axis: Brier score; red dashed line: every state is predicted with probability 1/4 ('All states'). See text and note to Figure 2 for details and explanations. As we have less (2,500) measurements per rate ratio than when computing the average Brier score per node and per rate ratio in Figure 2 (2,500 trials x 999 nodes), we provide here 95% confidence intervals. We observe that:

- As expected, predicting the root state is more difficult (e.g. parsimony is not any better than 'All states' with a rate ratio of 1, while with node prediction in Figure 2 there is a clear gap).
- Consistently, the number of predicted states (not shown) is larger for MPPA (e.g. with a rate ratio of 1, MPPA predicts ~2.45 states in average, versus ~1.30 in Figure 2). Interestingly, we do not observe such a difference with parsimony (~1.25 versus ~1.15).
- The ranking of the methods is the same as in Figure 2 with node prediction.
- MPPA is close to Marginal for all rate ratios, with non-significant differences in a large number of cases.

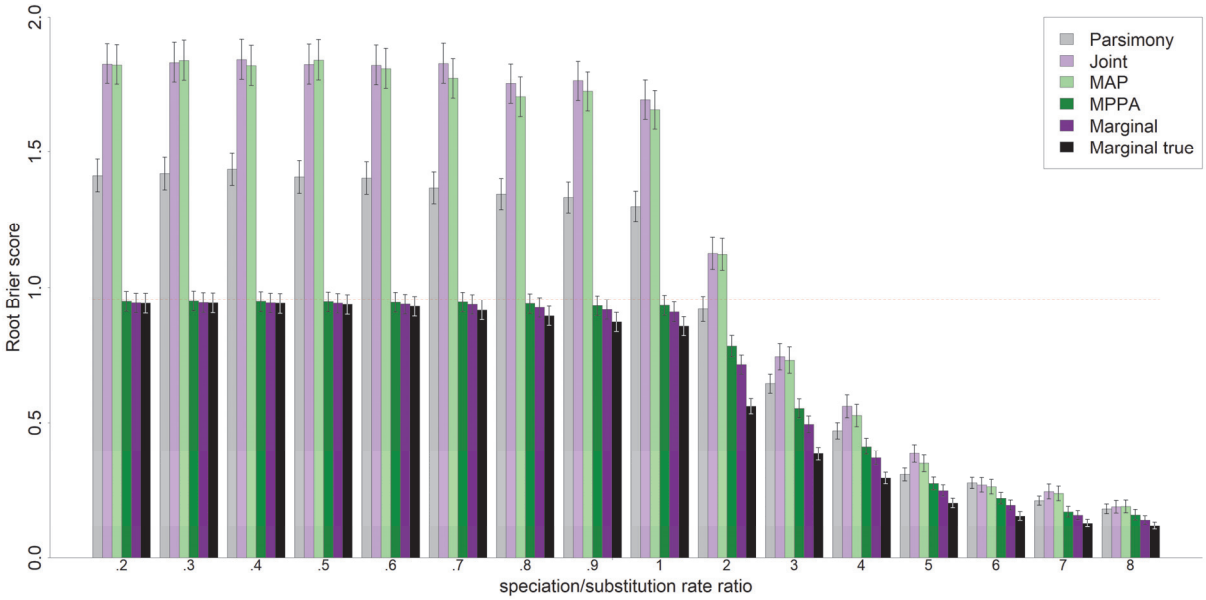

Figure S2. ACR accuracy comparison: tree root with 20-state protein-like data. See text and notes to Fig. 3 and S1 for details and explanations. The simulation setting is the same as in Figure 3B (F81-like model and noisy branch lengths). With 'All states' every state is predicted with probability 1/20. Conclusions are basically the same as with 4 states (Fig. S1):

- Root prediction is even more difficult with 20 states (e.g. no method is substantially better than 'All states' with a rate ratio of 1).
- Consistently, MPPA predicts a large number of states (~13.5 with a rate ratio of 1, versus ~3 with parsimony, which explains the gap in accuracy between both methods).
- The ranking of the methods is again the same as in Fig. 2-3 and S1, but with a higher contrast between Joint/MAP together, parsimony, and the others.
- There is no significant difference between MPPA and Marginal, and the difference between MPPA/MAP and 'Marginal true' is low, indicating that the model violations have a low impact.

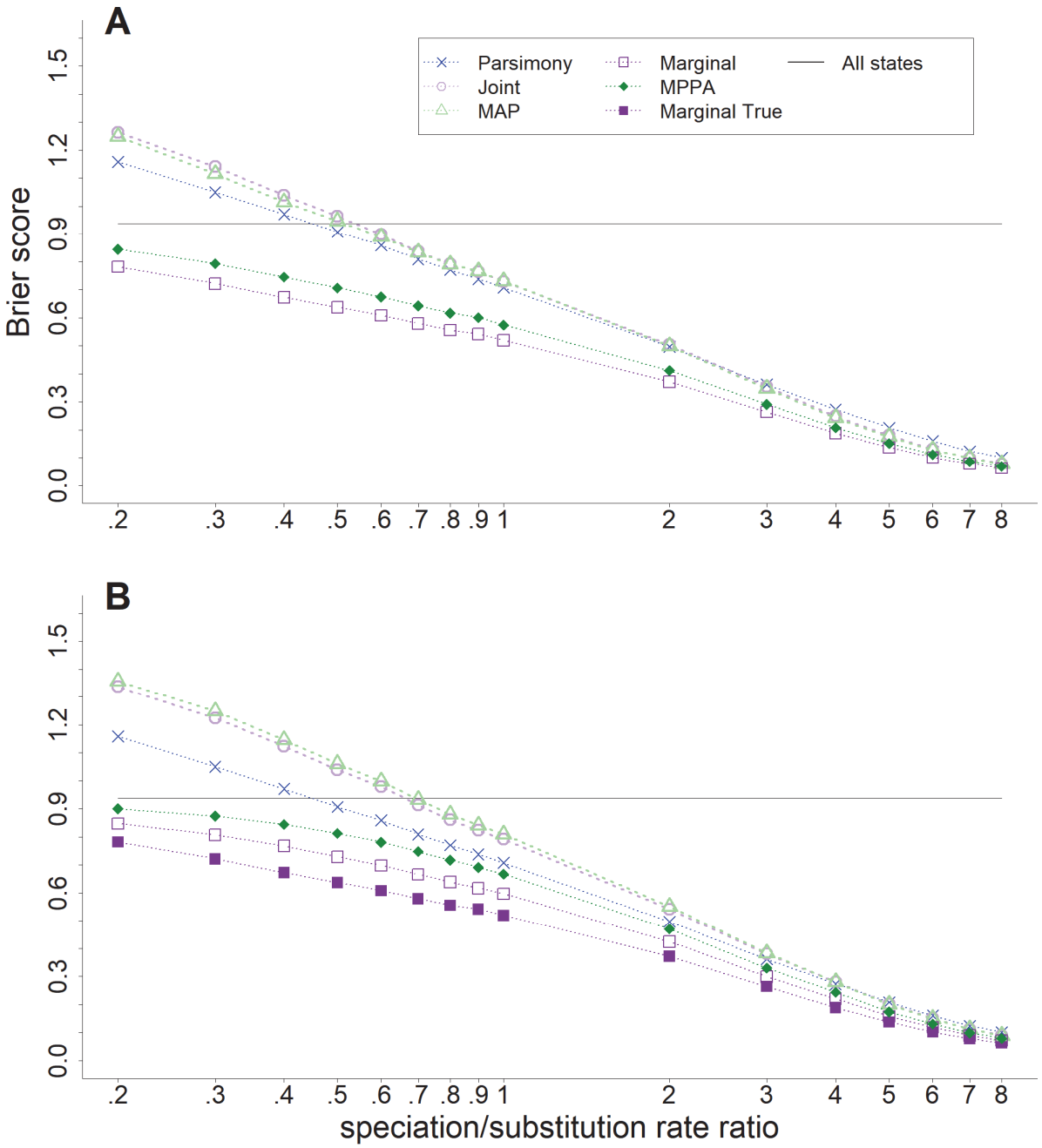

Figure S3. ACR accuracy comparison: edges with 4-state DNA-like data. The edge Brier score was used to check that the predictions for the two extremities of any given edge were compatible and close to the truth, thus establishing, or not, the superiority of global predictions as produced by Joint, over "independent" predictions as produced by MAP and MPPA. Simulation conditions are the same as in Figure 2 (A: true model and tree, B: F81 model and noisy branch lengths); see note to Figure 2 and text for details and explanations. We observe in this figure that Joint does not have any superiority over MAP (and MPPA). The ranking of the methods remains the same as in Figure 2 using node prediction accuracy.

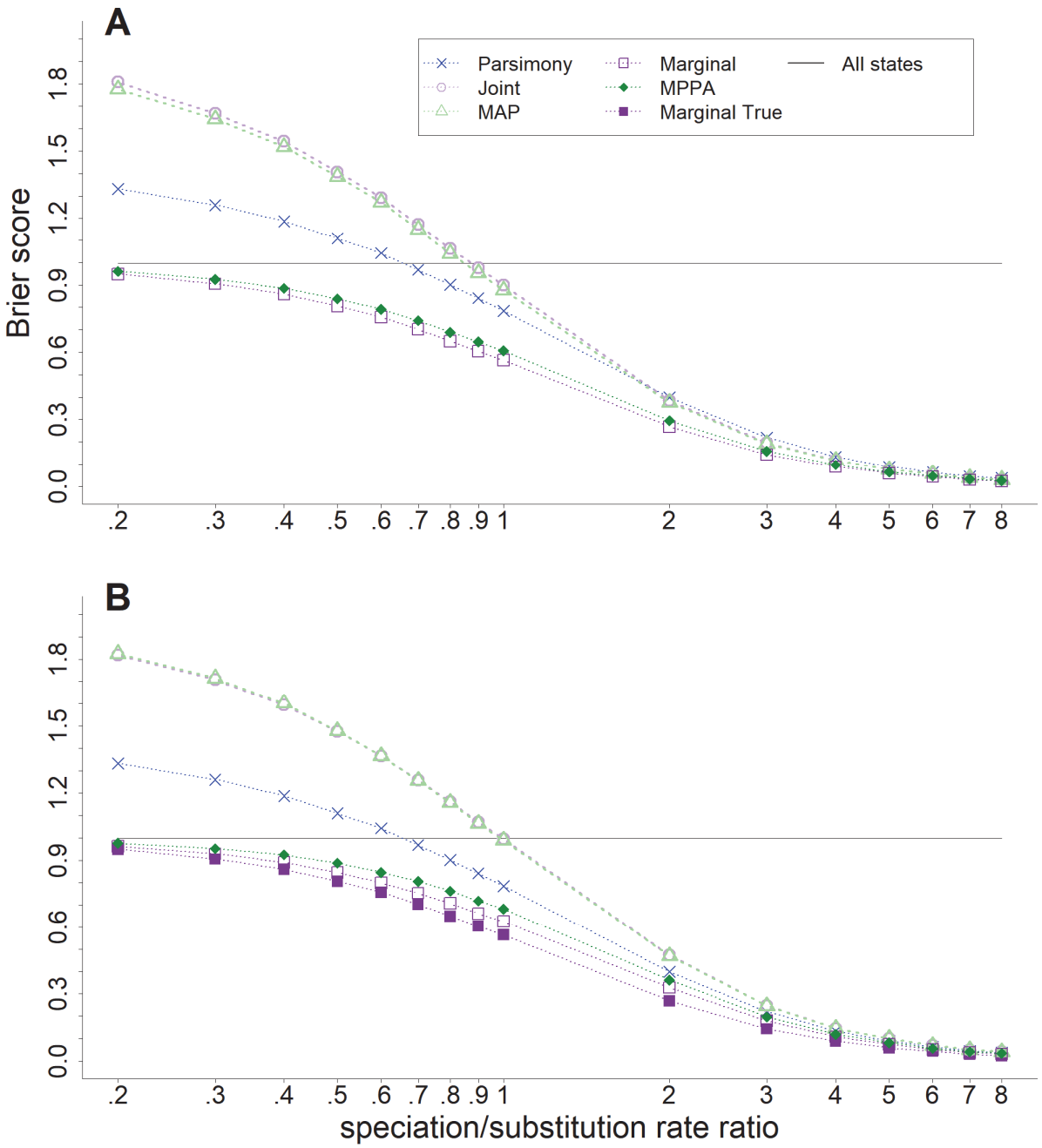

Figure S4. ACR accuracy comparison: edges with 20-state protein-like data. Simulation conditions are the same as in Figure 3 (A: true model and tree, B: F81-like model and noisy branch lengths); see text and notes to Figure 2-3 and S3 for details and explanations. Joint and MAP have very similar accuracies again, and the ranking of the methods remains the same as in Figure 3.

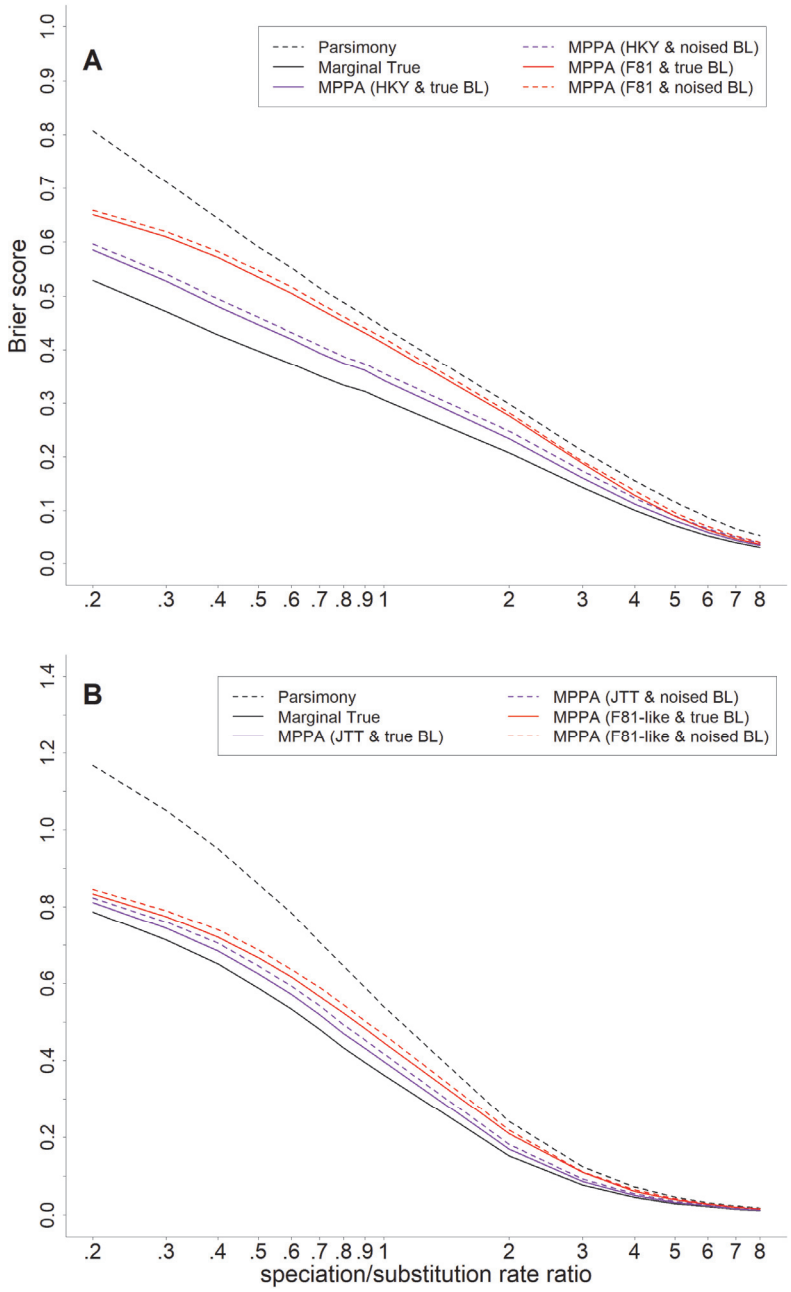

Figure S5. MPPA accuracy under different levels of model violation: A: 4-state DNA-like data; B: 20-state protein-like data. Here we use the four simulation settings described in the main text (True model and branch lengths; True model and noisy branch lengths; F81-like model and true branch lengths; F81-like model and noisy branch lengths). 'Marginal true' represents the best possible accuracy, as measured by the Brier score. See text and notes to Figure 2 and 3 for details and explanations. With the true model (4 state: HKY, 20 state: JTT), the gap between MPPA and 'Marginal true' is substantial with 4 states, but negligible with 20 states. Similarly, with 4 states there is a substantial gap for all methods between the true and F81-like models, but with 20 states this gap is negligible. Inaccurate/noisy branch lengths do not significantly degrade the performance of MPPA in any condition. Importantly, MPPA is more accurate than parsimony in all conditions, even with approximate model and branch lengths.

Robustness of PastML reconstructions with Dengue data

As described in the main text, to demonstrate the performance of PastML on real data, it was applied to the phylogeography of Dengue serotype 2 (DENV2) epidemics, using a medium-size dataset of 356 sequences, obtained from (Ayres et al. 2019). 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 ~10%). We constructed a fourth tree by collapsing the 40 poorly supported (SH-like supports <50%) branches in the PhyML tree. The phylogeography of DENV2 epidemics was reconstructed from these 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. The tree topologies inferred by RAxML were also substantially different (average bipartition distance for common taxa between the original and resampled trees of ~8%). We then reconstructed the phylogeography of these five resampled datasets and trees using the same approach as for the original alignment.

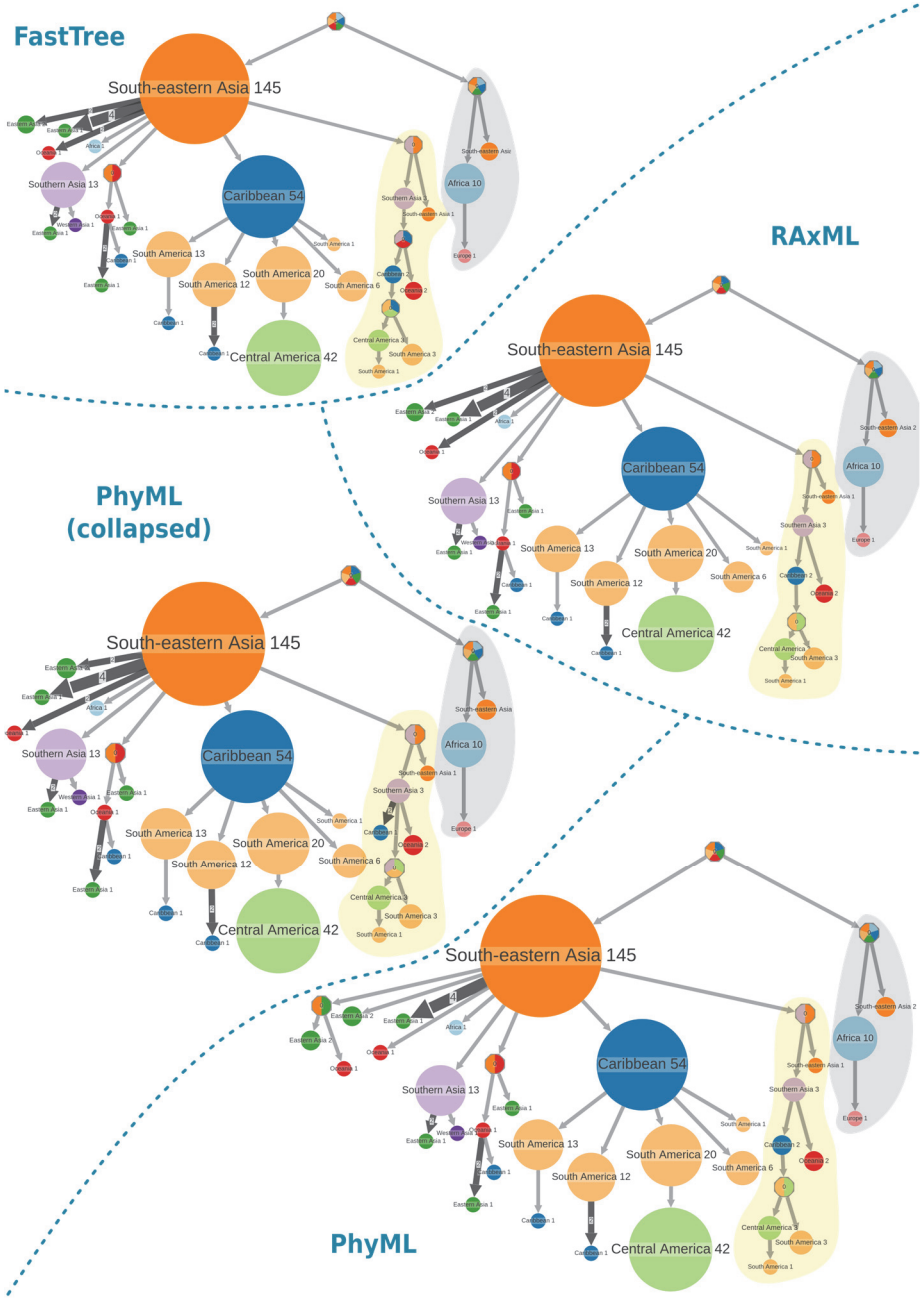

Figure S6. ACR of locations with four ML trees with different topologies. The average normalized bipartition distance among FastTree, PhyML and RAxML tree topologies is ~10%. 'PhyML (collapsed)' corresponds to the PhyML tree with poorly supported branches collapsed (SH-like <50%; 40 branches among 354 in total). The global topological information is preserved in all trees (e.g. the genotypes and sylvatic lineage are perfectly identified and supported). Despite these substantial (local) topological differences, we observe very little differences between the corresponding compressed scenarios shown here. The only difference that cannot be eliminated by resolving ancestral annotations of some unresolved nodes, corresponds to the DENV2 American genotype (on yellow background) spread from Southern Asia (lilac) to Central (light-green) and South (light-orange) Americas: in all the non-collapsed trees it happens via Carribean (blue), while in the collapsed tree this passage is direct. All other differences can be eliminated by resolving ancestral annotations of some unresolved nodes. This illustrates the robustness of PastML against phylogenetic uncertainty.

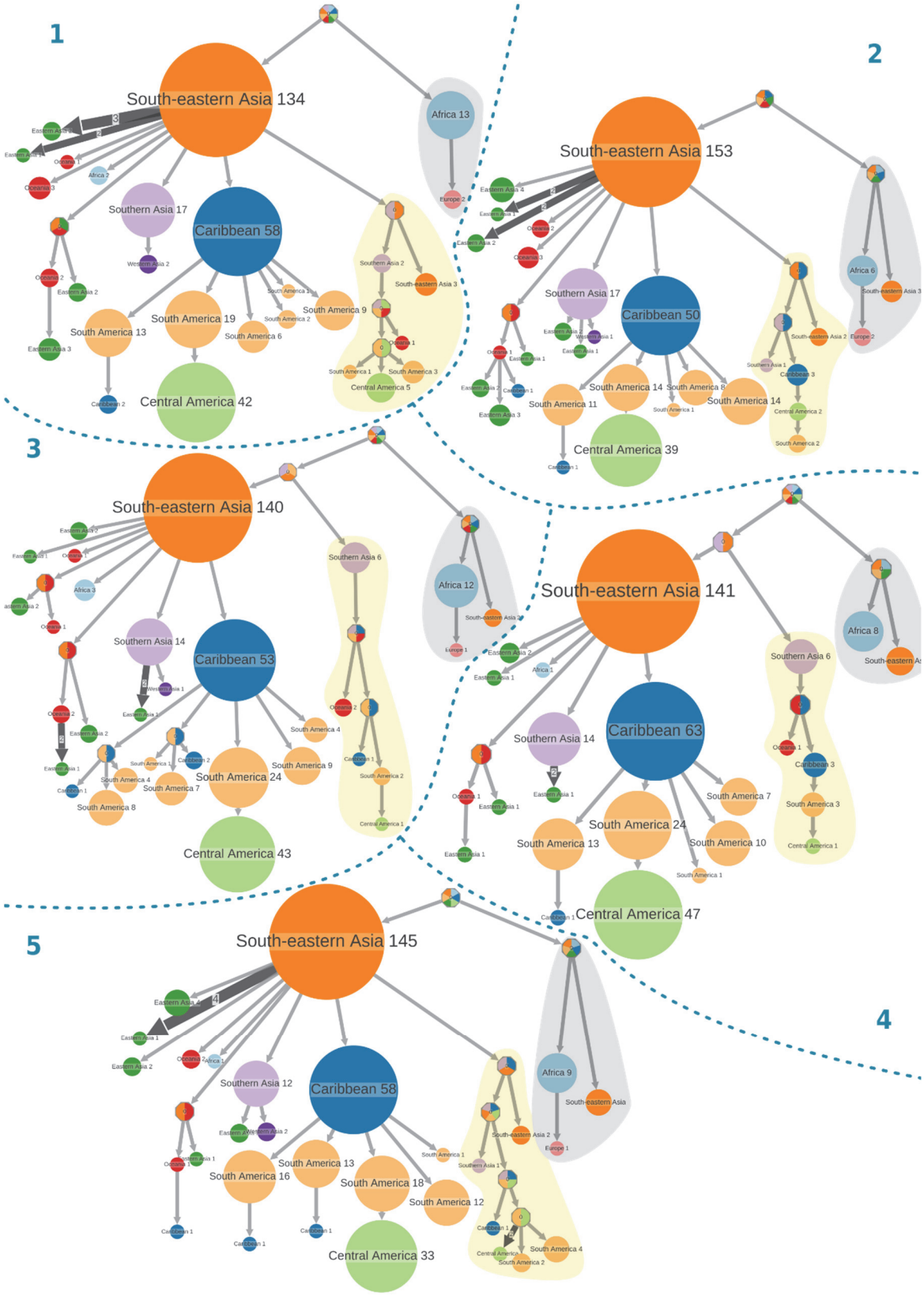

Figure S7. ACR of locations for five resampled trees. Globally all scenarios are compatible with the tree of Figure 3 using the original un-sampled dataset, despite the tree topology and sampling differences. Most of the differences can be resolved by resolving some unresolved nodes, in terms of ancestral annotation and/or topology.

ACR for two highly prevalent SDRMs

We reconstructed ancestral scenarios for the emergence, diffusion and reversion of three highly prevalent DRMs: M184V, K103N, and Y181C.

Results for the most prevalent DRM (M184V) are provided in the main text. The results for K103N are similar to those for M184V: they show emergence and growth of TDR clusters over time (e.g. the largest TDR cluster grows from 14 to 17 patients between 2005 and 2015), and well as growth of the number of ADR and reversions to sensitive state. The largest TDR cluster is located in Brazil. These results are shown in Figure S8.

The results for Y181C are different. We hardly see any TDR clusters: By 2005 a small TDR cluster of size 4 appears, but it does not grow by 2015. The ACR of location placed this cluster in India (Fig. S9). This could be the very start of TDR spread for this resistance mutation, hence making it a candidate for closer surveillance, or it could be due to its quick reversion time and hence inability to form large TDR clusters (median time of loss of 1.3 years, Castro et al. 2013).

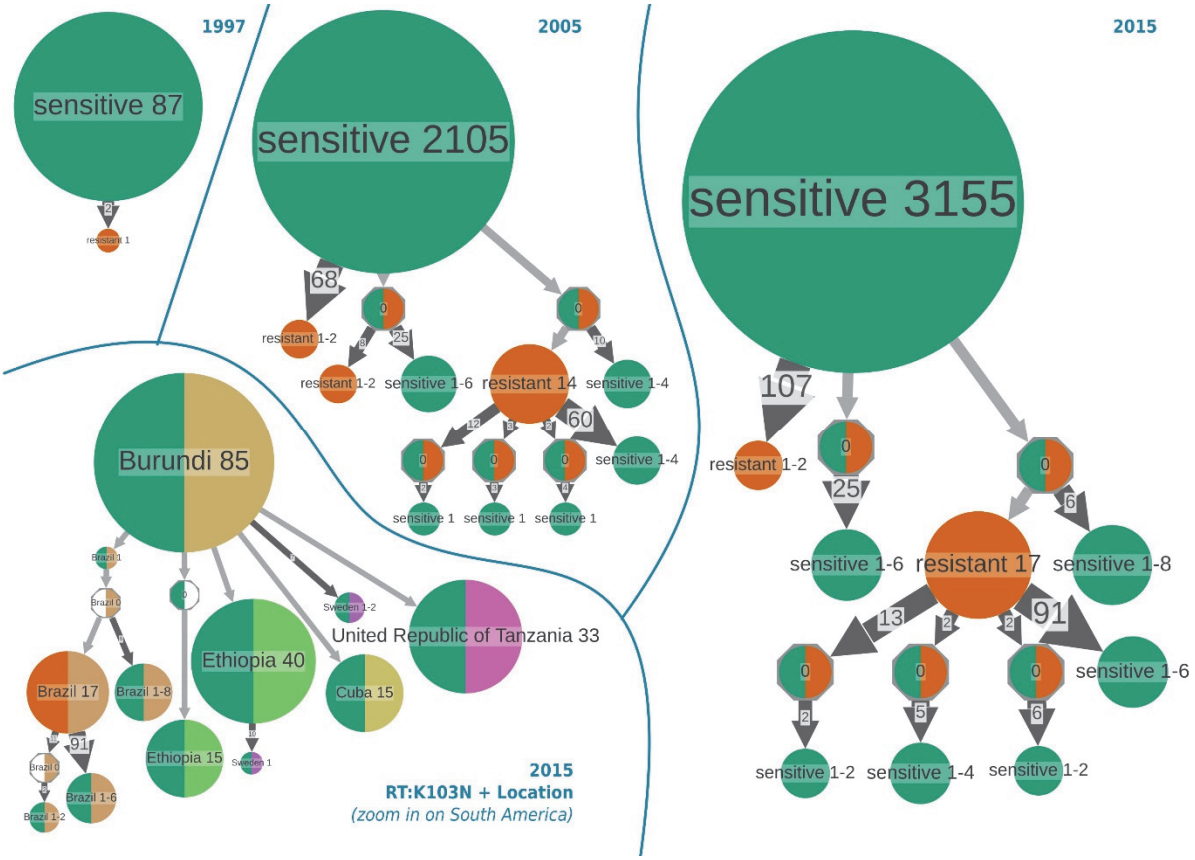

Figure S8: Ancestral state reconstruction of presence of DRM K103N over time (top) and combined with location data (bottom left). The reconstruction was done by PastML with default MPPA+F81 option. For the timeline at each year the tree was pruned to remove the tips sampled after that year prior to the reconstruction. In the bottom left figure the K103N presence/absence is combined with the location data: K103N 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.

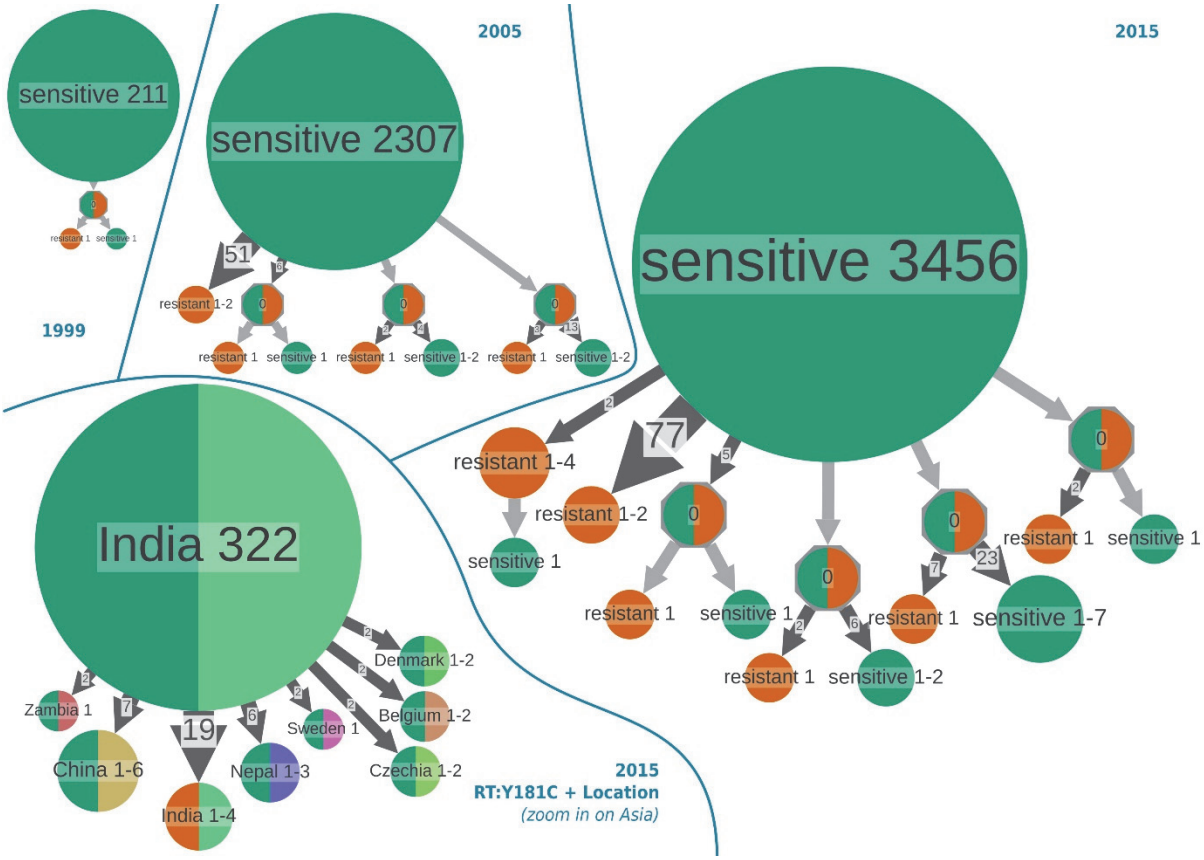

Figure S9: Ancestral state reconstruction of presence of DRM Y181C over time (top) and combined with location data (bottom left). The reconstruction was done by PastML with default MPPA+F81 option. For the timeline at each year the tree was pruned to remove the tips sampled after that year prior to the reconstruction. In the bottom left figure the Y181C presence/absence is combined with the location data: Y181C 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.