

Dated ancestral trees from binary trait data and its application to the diversification of languages:

[Supplementary material]

Geoff K. Nicholls

Department of Statistics, Oxford, UK.

Russell D. Gray

Department of Psychology, Auckland University, Auckland, New Zealand.

This document is available online at

<http://www.stats.ox.ac.uk/~nicholls/linkfiles/papers/NichollsGray06-SUPP.pdf>

Section 2. A model of binary trait evolution

A realization of the NOUNIQUE observation process introduced in Section 2 (main paper) is shown in Supplement-Fig. 1. The sets of traits branch independently at rate θ , new traits are born independently into each set at rate λ and each set element in each set dies independently at rate μ . Event +8 and the birth event +2 (which must occur on $\langle R, R^* \rangle$) so that 2 appears in the root-set $H(t_R, R)$ generate a trait which survives into no taxa at time $t = 0$. Events +3 and +7 generate a trait which survives into just a single taxon.

Section 4. Prior models on trees

Here are some further remarks on the tree prior f_R . Recall that

$$f_R(g|T) \propto \mathbb{I}_{t_R < T} \prod_{i \in S} (t_R - s_i)^{-1}$$

is a prior defined in Section 4 (main paper) on the space Γ' of admissible trees with fixed (or for the case of ancient languages, constrained) leaf ages given in years BP. When there are no calibration constraints, and all leaves have age zero units of time, this prior reduces to a prior in which the root time has an exactly uniform marginal distribution between age 0 and age T (quite unlike the uniform distribution on all admissible trees, which favors more deeply rooted trees).

When calibration constraints are present, it is not immediately clear how t_R behaves. For all sufficiently large t_R , in particular, when $\max_{i \in V_A} (s_i) \ll t_R \leq T$, the marginal

Address for correspondence: GK Nicholls, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK

E-mail: nicholls@stats.ox.ac.uk

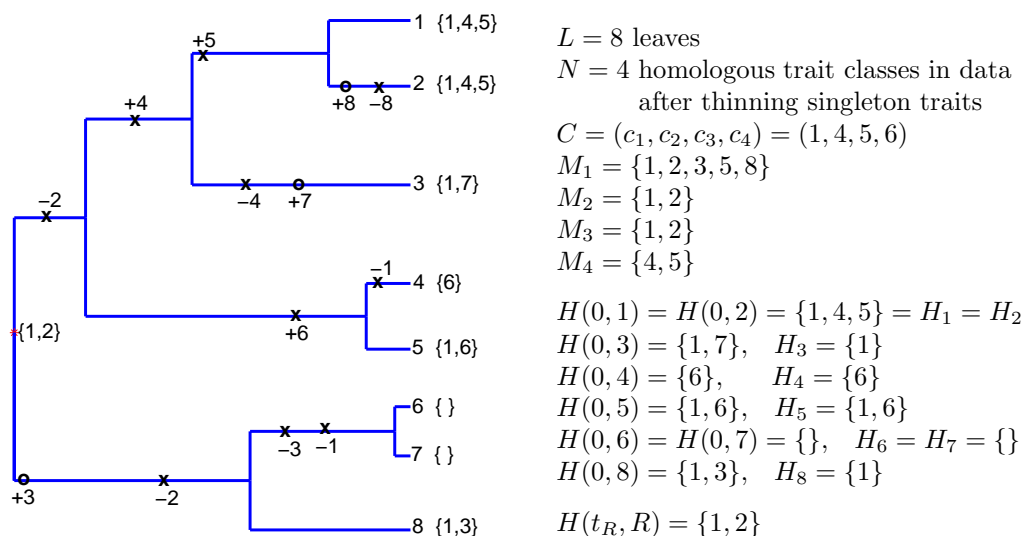


Fig. 1. Observation process and notation. Trait birth (+) and death (-) events are marked. Birth events marked “o” generate traits absent from the data. Leaves are labelled $i = 1, 2, \dots, 8$ from top to bottom. NOUNIQUE is illustrated.

density of the root will be well approximated by a uniform distribution. The marginal distribution of the root time was checked by simulation from the prior. In Supplement-Fig. 2 and Supplement-Fig. 3 we summarize simulations of the priors imposed in the Dyen/ $\{100, 200\}/30$ and Ringe/100/17 analyses respectively. The prior density for root times very close to the lowest achievable root time is a factor of 2 down from the maximum (rather than the factor 2^{L-2} we would get under the uniform distribution on admissible trees). The histograms in Supplement-Fig. 2 and Supplement-Fig. 3 have relatively large variability. Apart from the depletion at small age values, they are flat. The TraitLab MCMC is tuned to be efficient on the posterior, and is not efficient for simulation of the prior. Efficient prior simulation is itself straightforward, but would probably require implementation of some form of rejection algorithm.

In the case of f_G , an informative prior for θ generates calibration data for the rates μ and λ . In our analyses we used the prior $1/\theta$, which is improper. Simulation from the joint prior distribution of g and θ is not possible, as it is improper.

Subsection 7.2. Calibration data

At the phylogenetic level, statistical issues in the selection of language and species tree models are similar. In both applications, diversification is allowed where there is spatial isolation, or some other kind of segregation. An “isolation-event” process is not readily modeled by a tree prior with a physically plausible frequentist interpretation (as is the case in population genetics, where the Kingman coalescent may apply). Subjective Bayesian inference therefore features in both application areas. In both application areas, hypotheses concerning the age of the most recent common ancestor of groups of taxa are important,

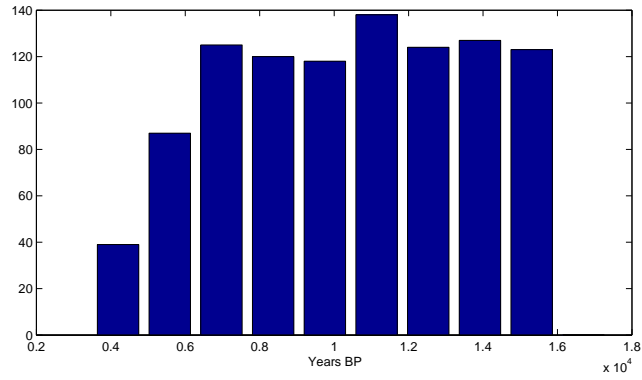


Fig. 2. Marginal prior distribution of the root age t_R in the prior f_R described in Section 4 (main paper and supplement). The prior simulated here is the prior imposed in the Dyen/{100, 200}/30 analyses.

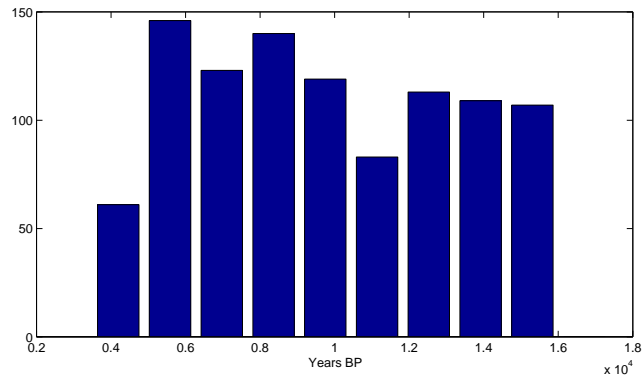


Fig. 3. Marginal prior distribution of the root age t_R in the prior f_R described in Section 4 (main paper and supplement). The prior simulated here is the prior imposed in the Ringe/100/17 analysis.

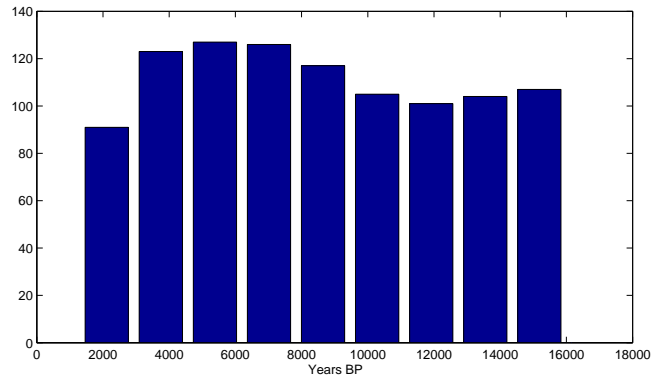


Fig. 4. Marginal prior distribution of the root age t_R in the prior f_R described in Section 4 (main paper and supplement). The prior simulated here is the prior for the clade structure of the synthetic tree shown in Fig. 5.

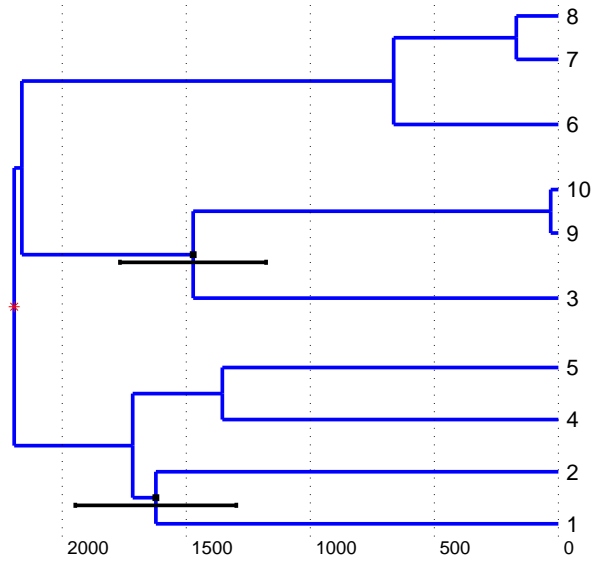


Fig. 5. This tree and clade structure were synthesized to give a simple example structure on which to test the prior. The marginal distribution of the root time is shown in Fig. 4.

and so priors which are non-informative with respect to dating hypotheses are relevant in both fields.

Sampled languages correspond to the sampled individuals, not the species, of phylogenetic species analysis. In our binary trait evolution model, a tree node can be thought of as marking a highly idealised “isolation” event, at which one group of speakers stopped talking to another. Nodes certainly do not represent some notion of loss of inter-language comprehensibility. In the model at least, the two languages are identical immediately below the split, since it is the split which allows differences to develop. In the species setting, under anagenic character substitution, individuals emerging from a tree node are genetically identical. However, nodes of (idealised) species trees do not represent an isolation event, since the coalescence of lineages occurs in the panmictic parent population.

One area of difference is the issue of word borrowing, modeled in Section 9.1 of the main paper. Interaction of languages occurs to an extent which causes linguists to question the usefulness of tree representations. The break-up of dialect groups into languages is not remotely binary. Whilst lateral transfer of genes between species may be important for certain groups of species, we believe the relative rate for this transfer process (compared to the species branching rate) is much lower for species than for cognate classes in language diversification. However, linguists do assert that the diversification of languages is tree-like on large time scales. Also, the characters we analyse evolve on a time scale (3000 years) that is large compared to the time scale of the break up of dialect continua (a few hundred). As a consequence, the great majority of characters sail straight through these network-nodes of the tree. The fraction of characters which are born or die in these break-up intervals is small. Essentially, the data does not resolve the problematic structures. This is our basis for returning the tree-based representation, and tree models formally similar to phylogenetic species models.

In contrast, as discussed in the last paragraph of Section 2 (main paper), when we consider the detailed evolution of our binary characters, the observation model here is unlike any standard character substitution model from genetics. The identity between our model and the gene-substitution model of Huson and Steel (2004) is an accident, which arises because both are motivated by applications which lead to the consideration of models of the time evolution of sets of binary traits. In this respect, the genetic setting of Huson and Steel (2004) is not essential.

Subsection 8.2. Results

In Supplement-Fig. 6 we give the posterior clade probability estimated under the branching process prior f_G (in contrast to the results in the main paper which are estimated under tree prior f_R). The support is very similar to that displayed in Fig. 1 (main paper). In Supplement-Fig. 7 we give the posterior mean age for the most recent common ancestor of the languages in each clade. The combined g, θ prior given by $\theta^{-1}f_G(g|\theta)$ used in this check is scale invariant, and improper. Supplement-Fig. 7 is very similar to Fig. 2 (main paper). Error bars are slightly longer here, and there is slightly more consistency in estimated ages.

In Supplement-Fig. 16 through Supplement-Fig. 24 (top tree in each figure), we display consensus tree reconstructions of topology and branch length for all the posterior distributions summarised in Fig. 1 and Fig. 2 (main-paper).

A tree edge, or “split”, determines a partition of the leaves into two sets. Our majority rule consensus trees show all splits present in 50% or more posterior samples, and indicate

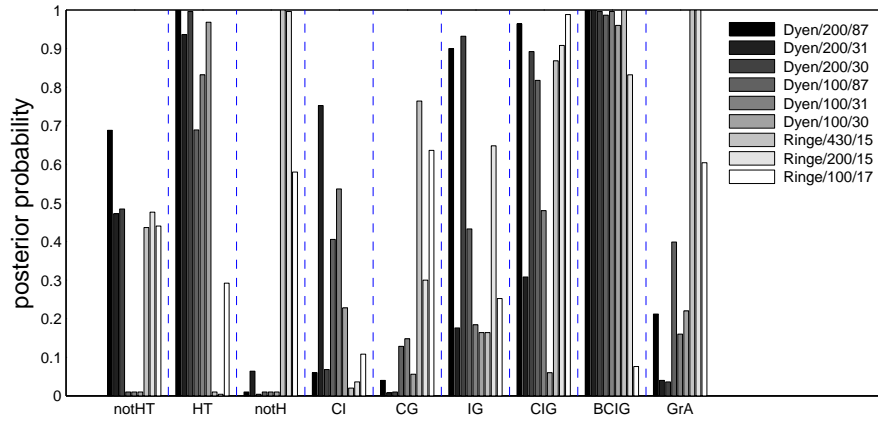


Fig. 6. Compare Fig. 1 (main paper). Posterior probabilities for selected clades, across data sets, fitting under the f_G tree prior with Jefferys prior for branching rate θ . Fig. 1 (main paper) gives results under f_R tree prior.

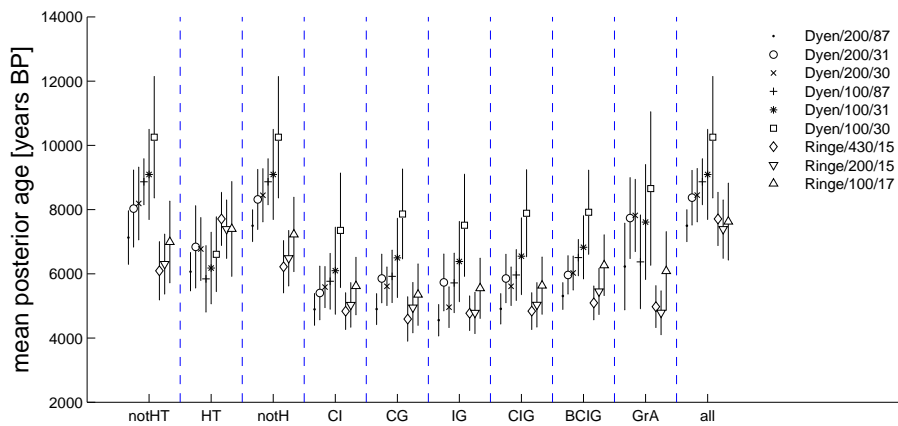


Fig. 7. Compare Fig. 2 (main paper). The mean posterior ages, in years BP, for the most recent common ancestor (MRCA, super-clade root) of languages in selected combinations of clades, fitting under the f_G tree prior with Jefferys prior for branching rate θ . Fig. 2 (main paper) gives results under f_R tree prior.

the percent support for splits with support less than 95%. Edge length depicted is the posterior mean elapsed time in years, conditioned on the existence of the corresponding split. The consensus tree is a popular central point estimate in phylogenetic analyses. However, splits with posterior support close to 100% in a Bayesian phylogenetic analysis often have far lower frequency in bootstrap analyses of the same data. Here, the Dyen/200/30 and Dyen/100/31 consensus trees (respectively Supplement-Fig. 19 and Supplement-Fig. 21) group Balto-Slav with Germanic, Celtic and Italic at above 95%, whilst Ringe/100/17 (Supplement-Fig. 22) puts Balto-Slav with Indo-Iranian at 93%.

The consensus tree is not a state in the sample space of trees. Prior constraints, and in particular leaf ages, are best represented on sampled states so in Supplement-Fig. 16 through Supplement-Fig. 24 (bottom tree in each figure) we give samples from the Ringe/100/17 and Ringe/200/15 distributions.

In an earlier version of this paper, the MCMC output was further processed using the R package Ape 1.8-2 of Paradis et al. (2004). Monte Carlo simulations were carried out using TraitLab, a freely available MatLab package written by Geoff Nicholls and David Welch. TraitLab now has a facility to compute and draw consensus trees, and this package was used to create all figures in this paper, bar the consensus networks. In an earlier version of this paper we followed Holland et al. (2005), and gave an additional representation of uncertainty in topology *via* a consensus network. The network analysis tool SplitsTree V4.4 of Huson and Bryant (2006) was used for this purpose. Consensus networks display all splits with posterior support above a given threshold set by the user. The consensus network contains the consensus tree as a subtree with branch lengths measuring posterior support. In this revised supplement we represent this uncertainty via Fig. 1 (main paper) and the consensus trees below.

Section 9. Model mis-specification

See Supplement-Fig. 8 for the true tree on which synthetic data was simulated for the mis-specification analysis of Section 9 (main paper).

Subsection 9.1. Global and local borrowing

Wang and Minett (2005) give a statistical method for measuring the frequency of borrowing in lexical data. Their method is based on analysis of summary statistics and avoids phylogenetic inference.

Subsection 9.2. Predictive distributions and external data

See Supplement-Fig. 9 for a comparison of posterior predictive distributions with data for each language (x -axis) against the number of singleton traits (y -axis) at each leaf. Supplement-Fig. 10 and Supplement-Fig. 11 mimic the analysis of Supplement-Fig. 9 on synthetic data simulated under borrowing (Supplement-Fig. 10) and rate heterogeneity (Supplement-Fig. 11).

Observe that rate heterogeneity distorts the distribution of the singleton traits in a way similar to borrowing. On the real data (Supplement-Fig. 9) the model misfit disappears as we move to shorter Swadesh word-lists. When we simulate synthetic data for comparison,

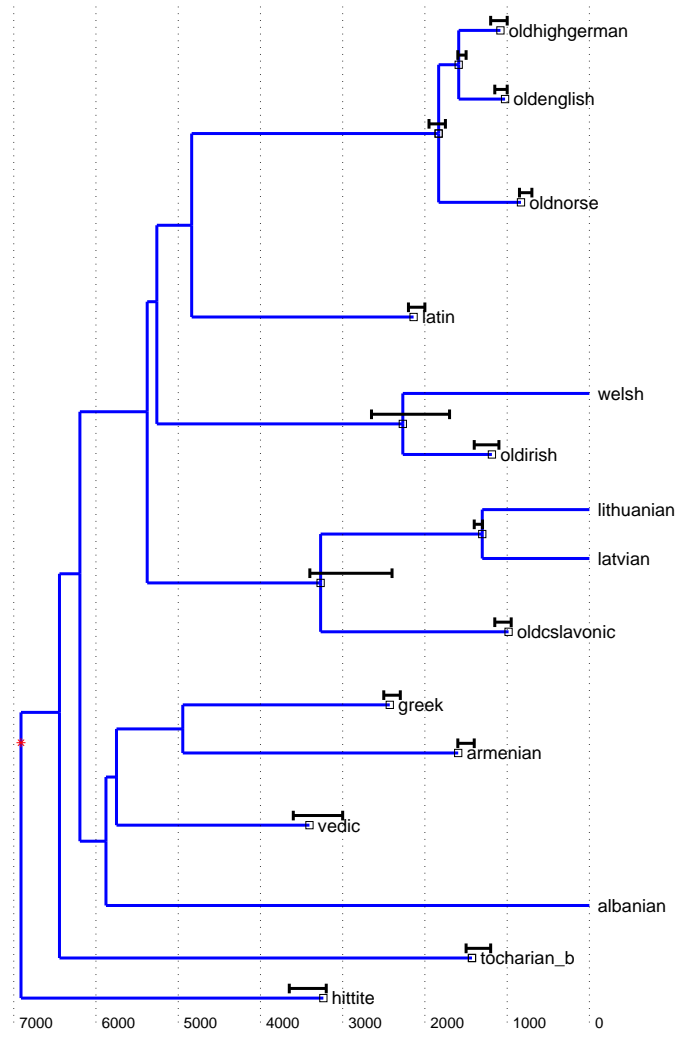


Fig. 8. The true tree in the synthetic studies reported in Section 9 (main paper)

in Supplement-Fig. 10 and Supplement-Fig. 11, we keep the list size fixed (at $K = 200$) and vary the magnitude of the synthesized model error component.

Subsection 9.3. Rate heterogeneity across traits

Supplement-Fig. 12 is the synthetic data study referred to in the 3rd paragraph of Section 9.3 (main paper). It can be seen that, at high levels of rate heterogeneity between traits (left graph in Supplement-Fig. 12) the posterior predictive distribution for the number of languages in which a given lexical trait appears (dashed envelope) does not well capture the corresponding distribution in the data (solid line). At lower levels of rate heterogeneity (right graph) the predictive distribution encloses the data statistic. A similar decrease in this aspect of misfit is seen in Supplement-Fig. 13 as we move from the Ringe et al. (2002) data gathered for the ‘longest’ word list ($K = 328$ meaning categories, left graph in Supplement-Fig. 13) to the shortest ($K = 100$, at right).

Subsection 9.4. Rate heterogeneity in space and time

In Supplement-Fig. 14 we give a cross-validation study for the Dyen/100/87 analysis. There are seven calibration constraints, plus three age intervals for ancient languages, which play a similar role. In each of 10 associated analyses, one calibration constraint is removed, and the corresponding age (the age of the root of the deconstrained clade) is estimated using the remaining nine calibration constraints plus the Dyen et al. (1997) data in the Swadesh-100 wordlist. One MCMC simulation of the 9-clade Dyen/100/87 posterior is needed for each vertical bar in Supplement-Fig. 14.

The predicted age of the root of the Balto-Slav clade and the age of Hittite are incorrect. The upper limit on the age of the most recent common ancestor of the languages in the Balto-Slav clade stands at 3400 BP. This limit asserts a prior belief that the ancestral Baltic and Slavic languages did not differ by any lexical traits before 3400 BP. Unlike the other upper limits present in the calibration constraints, this is based on no historical record (ie no written record), so we question this upper limit. This point was raised also in the final paragraph of Section 7.2 (main paper).

The Hittite data has been flagged, in the fourth paragraph of Section 8.2 (main paper), as possibly corrupt. This “lost” language was reconstructed from a small number of texts. It is possible that some terms present in this language are not attested in these particular texts. The number of lexical traits per language observed in the Dyen et al. (1997) data (Swadesh 200 word-list) is plotted in Supplement-Fig. 15. Note that lexical traits present in just a single language (singleton traits) have been removed from these data. We observe a form of polymorphism: languages with more than 200 lexical traits must have some meaning categories with more than 1 word, and this is not even including the singleton traits. At the other end of the histogram, Hittite has the least traits, just 69 (after thinning of singletons). We might ask how the Hittites made themselves understood, with just 69 words in the 200 core meaning categories. However, the number of singleton traits may be large, and these may populate the otherwise empty meaning categories (there are at least 131 of these). Thus, there are four modern languages with less than 120 lexical traits: Gypsy_Gk, Singhalese, Ossetic and Wakhi. For these languages we may assume full knowledge of the lexicon. Given this feature of Hittite, it is important to check which date estimates are sensitive to an artificially aged Hittite data vector. As we note in the final

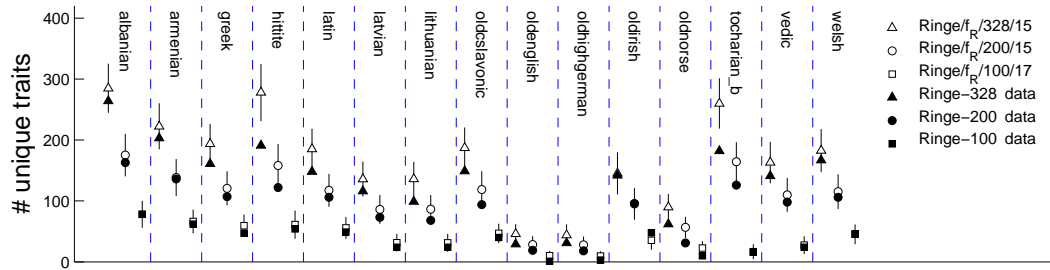


Fig. 9. Posterior predictive distributions for the number of singleton traits, external data from the Ringe et al. (2002): \triangle 328 meaning categories; \circ Swadesh-200 list; \square Swadesh 100 list.

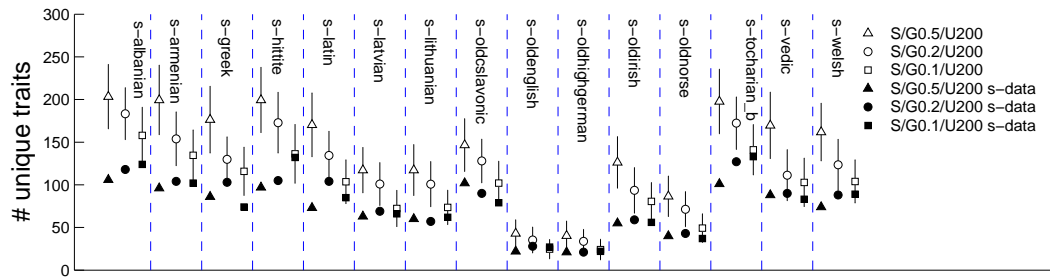


Fig. 10. Synthetic data incorporating borrowing yield posterior predictive distributions for the number of singleton traits, and show increasing singleton-trait depletion as the level of synthesized borrowing increases: \triangle borrowing rate $\beta = 0.5\mu$; \circ borrowing rate $\beta = 0.2\mu$; \square borrowing rate $\beta = 0.1\mu$.

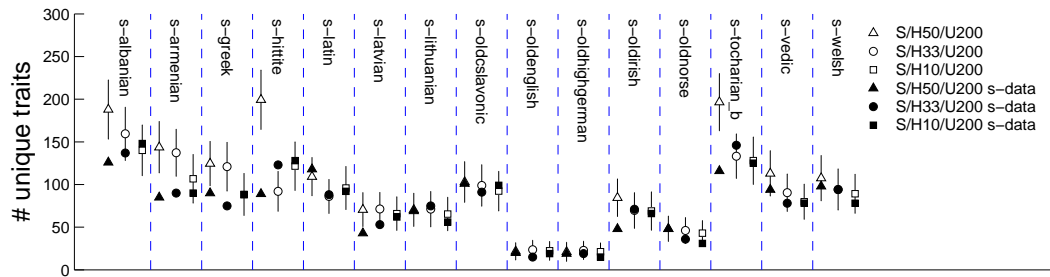


Fig. 11. Synthetic data incorporating rate heterogeneity yield posterior predictive distributions for the number of and show increasing singleton-trait depletion as the level of rate heterogeneity increases: \triangle rate heterogeneity with standard deviation from one branch to another of 0.5μ ; \circ standard deviation 0.33μ ; \square standard deviation 0.1μ .

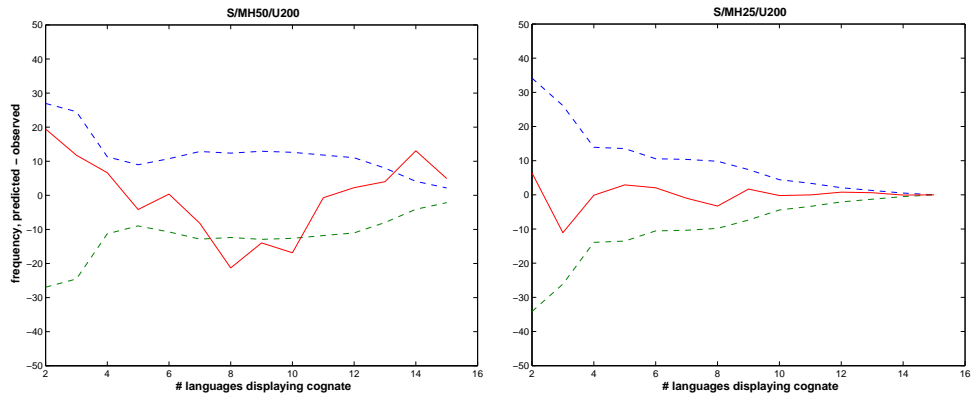


Fig. 12. Posterior predictive distributions (synthetic data) for $Y^{(n)}$, the number of traits displayed at exactly n leaves. y -axis $E(\tilde{Y}^{(n)}|D) - Y^{(n)}(D)$ and the envelope $\pm 2\text{std}(\tilde{Y}^{(n)}|D)$. x -axis $n = 2, 3, \dots, L$. (Left) Predictive distribution for $S/MH50/U200$ posterior. (Right) $S/MH25/U200$ posterior.

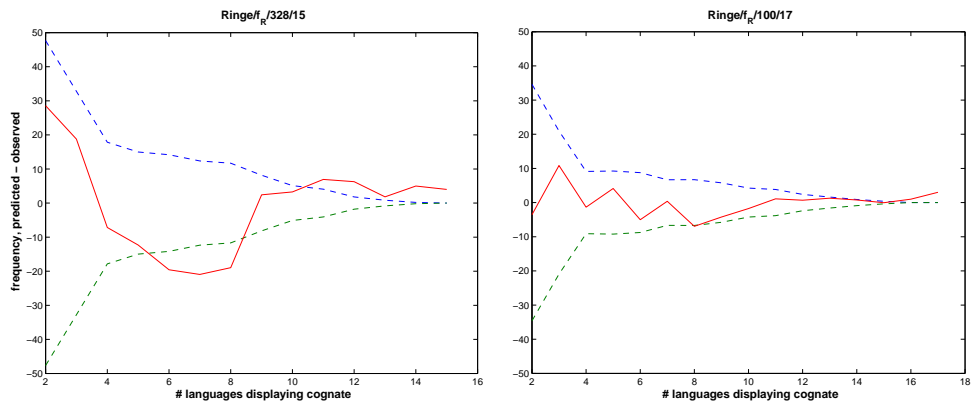


Fig. 13. Posterior predictive distributions (real data), otherwise as Fig. 12. (Left) Predictive distribution for Ringel/328/15 posterior. (Right) Ringel/100/17 posterior.

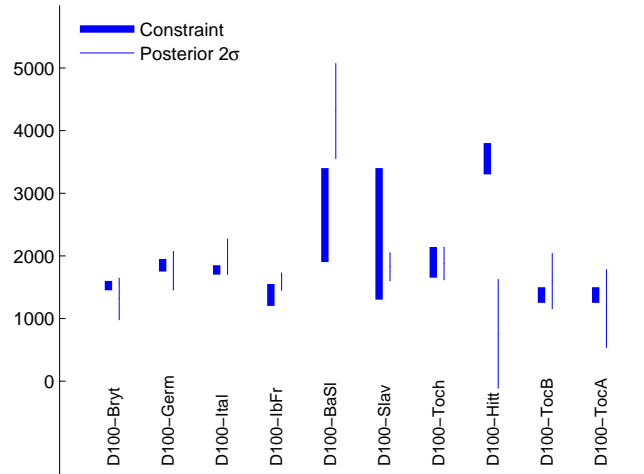


Fig. 14. A cross-validation study for the $D_{\text{yen}}/100/87$ analysis discussed in Subsection 9.4 of this supplement.

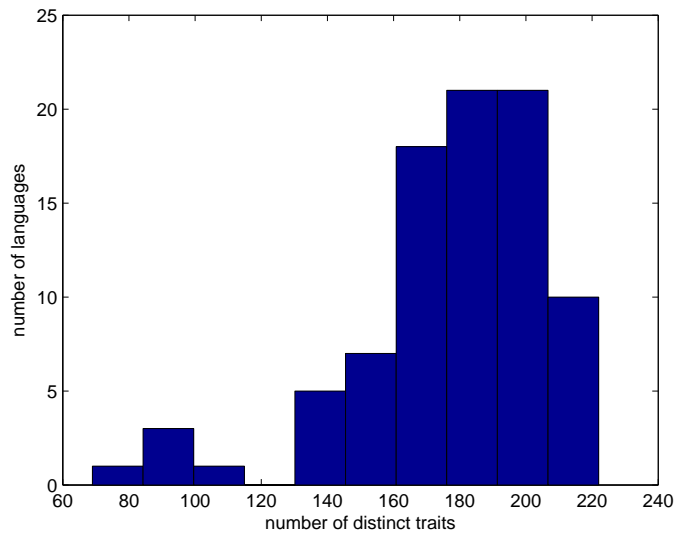


Fig. 15. The number of cognate instances (ie the number of distinct lexical traits, or words) per language, in the Swadesh-200 wordlist, from the Dyen et al. (1997) data (supplemented with Hittite and Tocharian A and B by Gray and Atkinson (2003)), after thinning of singleton traits. Hittite has 69 cognate lexical items, the smallest number.

paragraph of Section 9.4, it happens that the root age results in the Dyen/100/{87, 30, 31} and Ringe/100/17 analyses are not determined by the position of any single language, though this need not always be the case.

References

- Dyen, I., J. Kruskal, and P. Black (1997). FILE IE-DATA1, raw data available from <http://www.ntu.edu.au/education/langs/ielex/IE-DATA1>. Binary data available from <http://www.psych.auckland.ac.nz/psych/research/RusselsData.htm>.
- Gray, R. and Q. Atkinson (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426, 435–439.
- Holland, B. R., F. Delsuc, and V. Moulton (2005). Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to study the origins of placentals and hexapods. *Systematic Biology* 54, 66–76.
- Huson, D. and D. Bryant (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23, 254–267.
- Huson, D. and M. Steel (2004). Phylogenetic trees based on gene content. *Bioinformatics* 20(13), 2044–2049.
- Paradis, E., J. Claude, and K. Strimmer (2004). Ape: Analysis of Phylogenetics and Evolution in R language. *BIOINFORMATICS* 20, 289–290.
- Ringe, D., T. Warnow, and A. Taylor (2002). Indo-European and Computational Cladistics. *Transactions of the Philological Society* 100, 59–129. Data available from <http://www.cs.rice.edu/nakhleh/CPHL/#datasets>.
- Wang, W. S.-Y. and J. W. Minett (2005, August). Vertical and horizontal transmission in language evolution. *Transactions of the Philological Society* 103(2), 121–146.

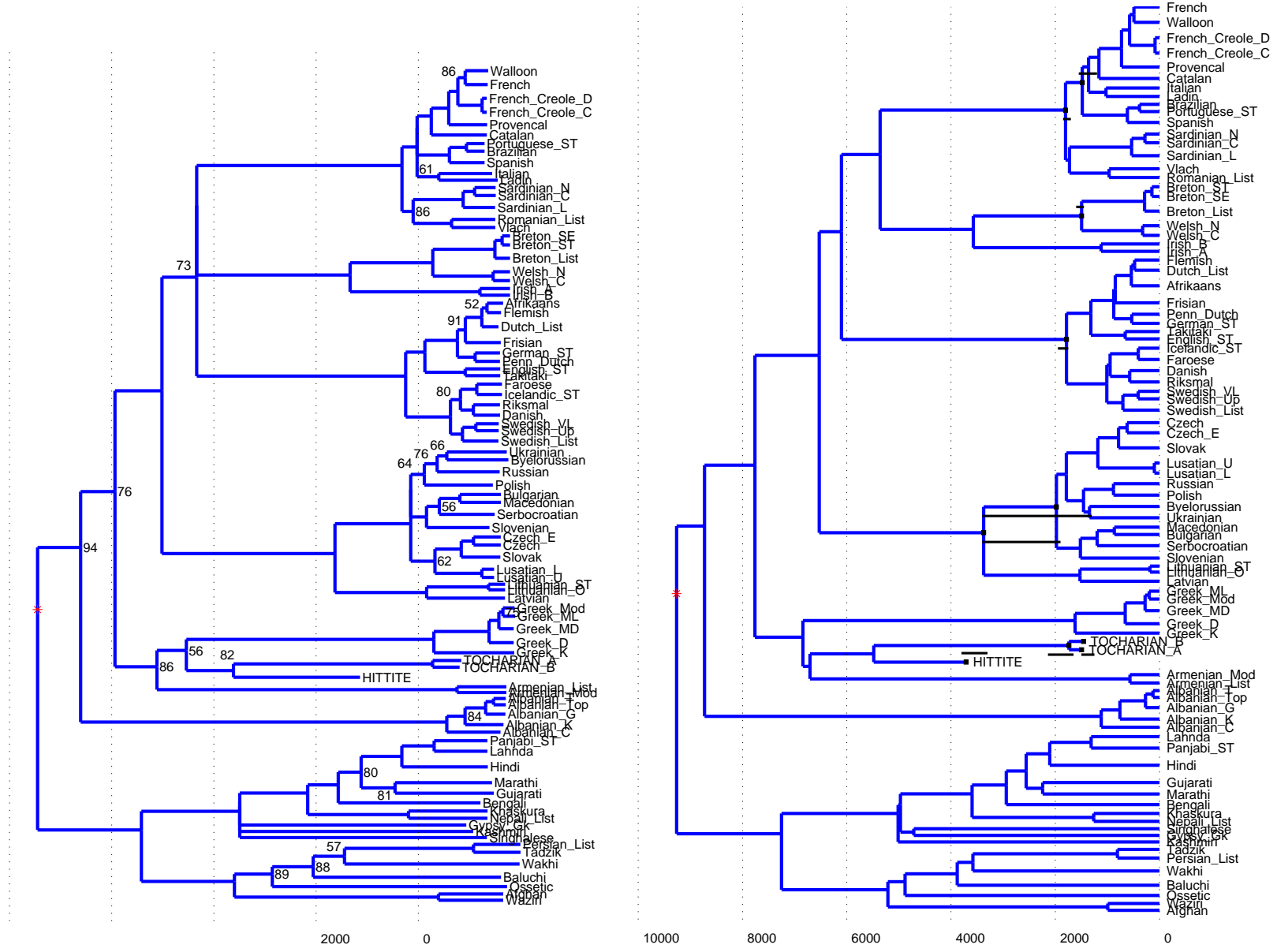


Fig. 16. Consensus tree (top) and a sampled state (bottom) illustrating the D_{YEN}/100/87 (bottom) posterior distribution. In the consensus tree, the posterior probability for the clade below each node is marked, as a percentage. It is omitted where it exceeds 95%. *Prior* uncertainties in clade root and leaf ages are indicated by vertical bars (bottom).

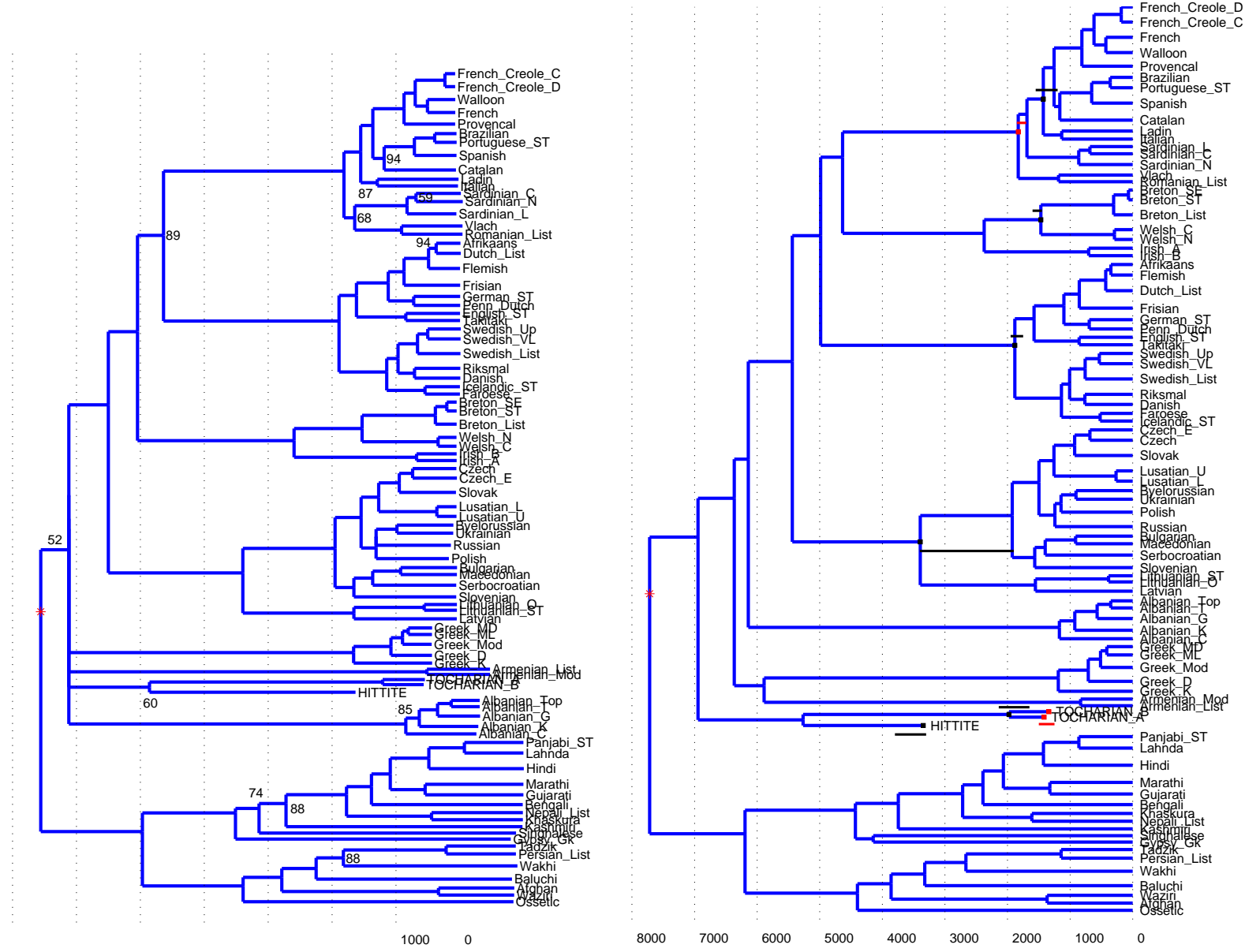


Fig. 17. Consensus tree (top) and a sampled state (bottom) illustrating the Dyen/200/87 (bottom) posterior distribution. Details as Supplement-Fig. 16.

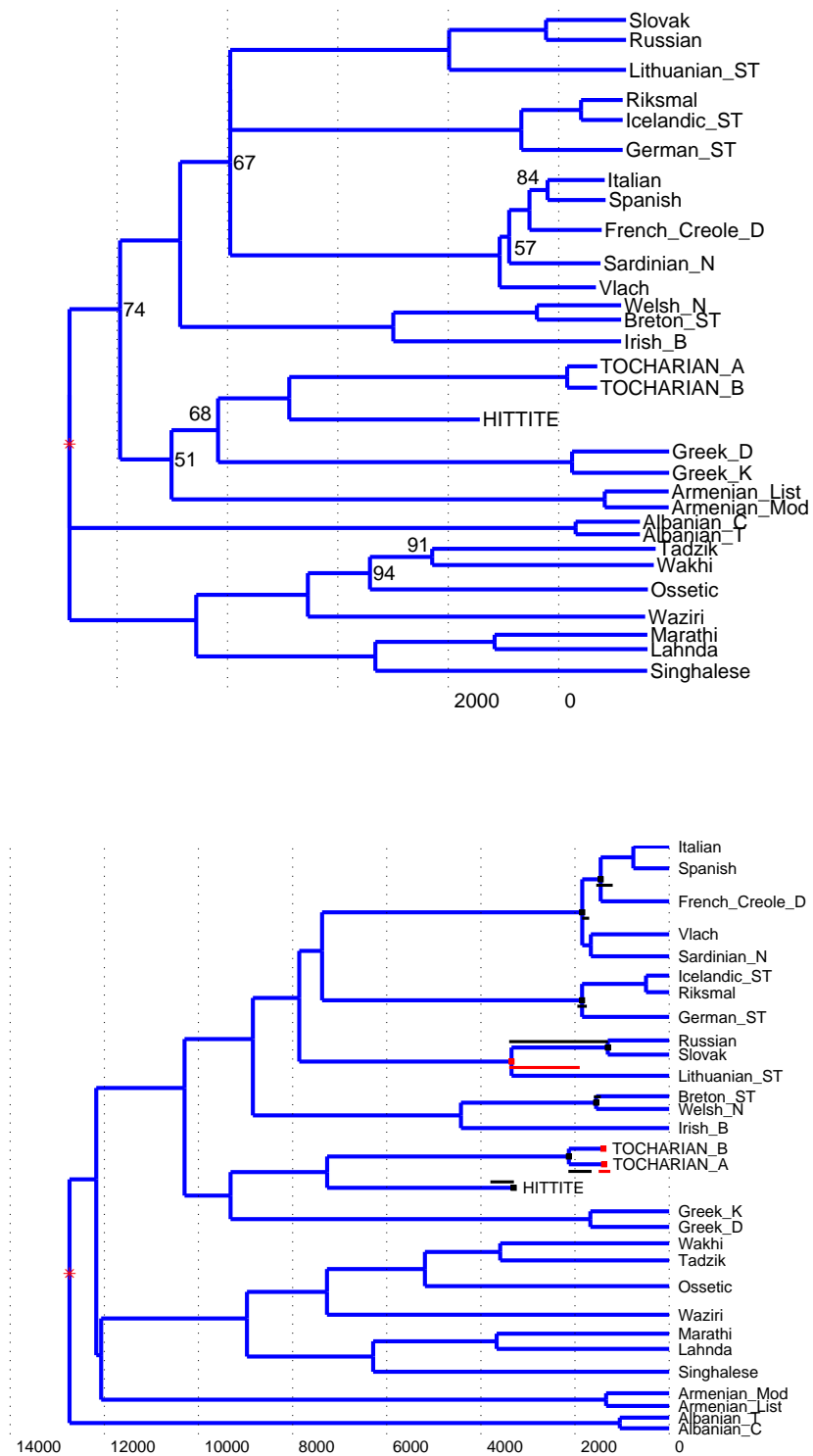


Fig. 18. Consensus tree (top) and a sampled state (bottom) illustrating the Dyen/100/30 (bottom) posterior distribution. Details as Supplement-Fig. 16.

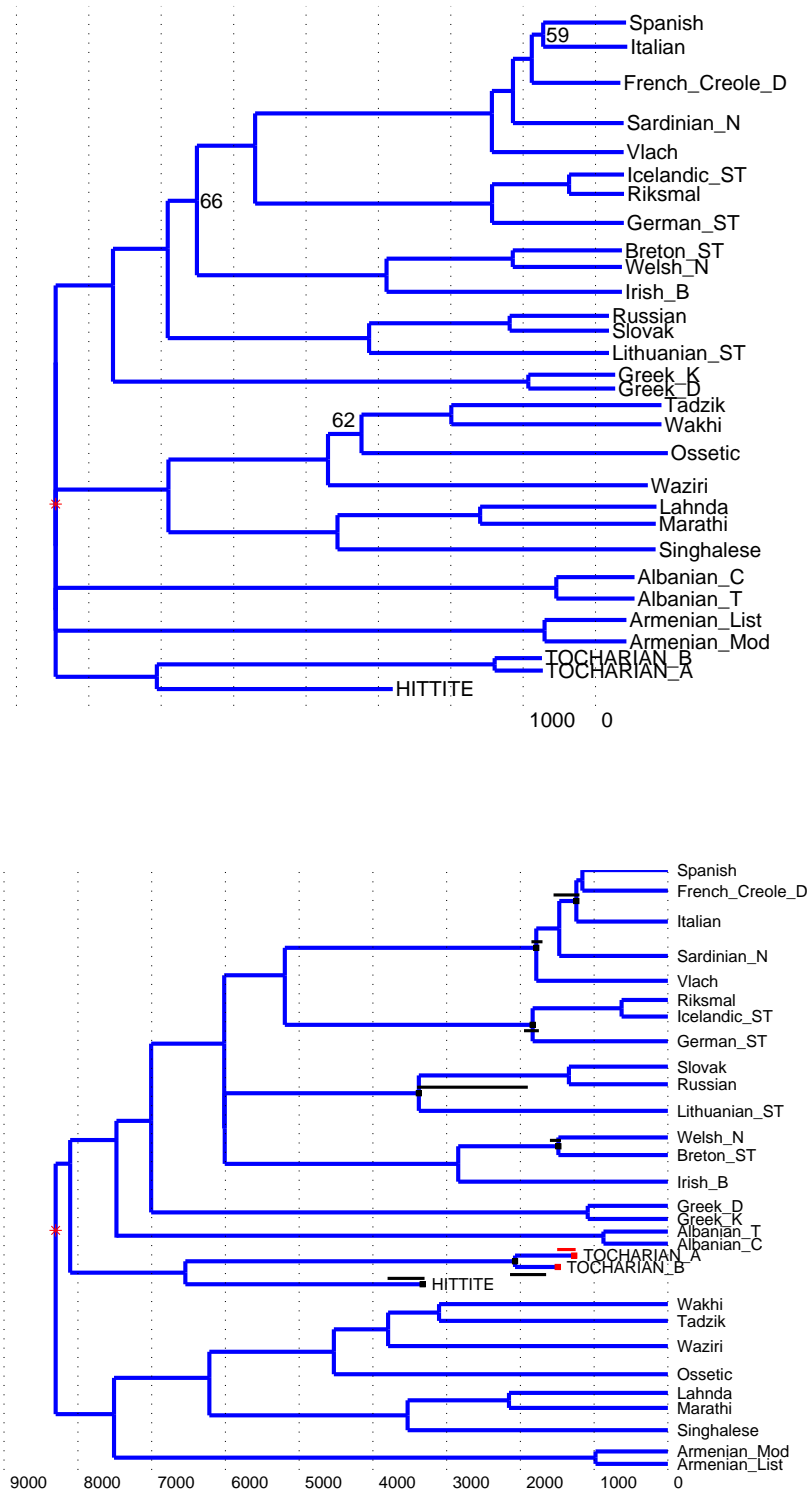


Fig. 19. Consensus tree (top) and a sampled state (bottom) illustrating the Dyen/200/30 (bottom) posterior distribution. Details as Supplement-Fig. 16.

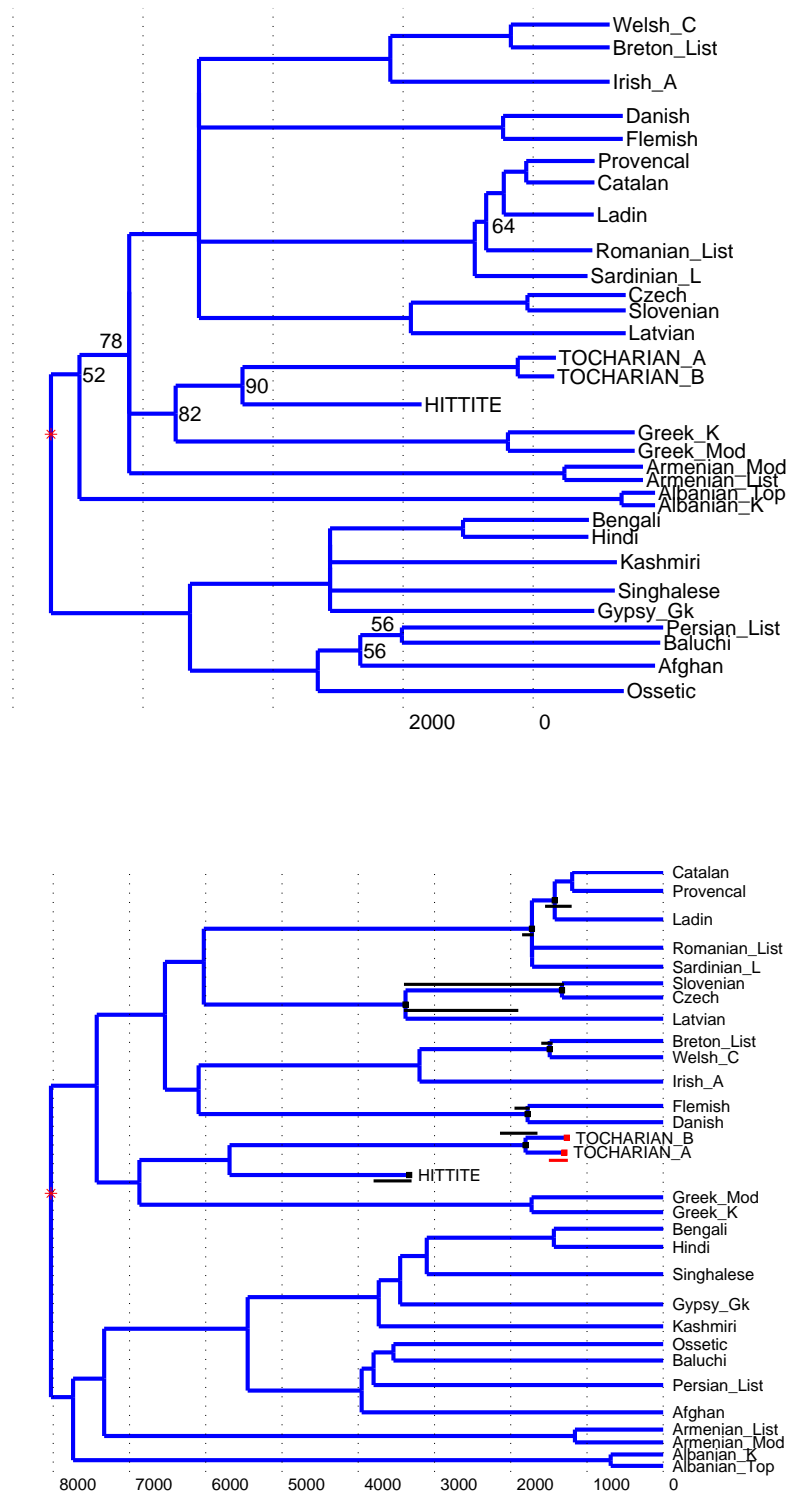


Fig. 20. Consensus tree (top) and a sampled state (bottom) illustrating the Dyen/100/31 (bottom) posterior distribution. Details as Supplement-Fig. 16.

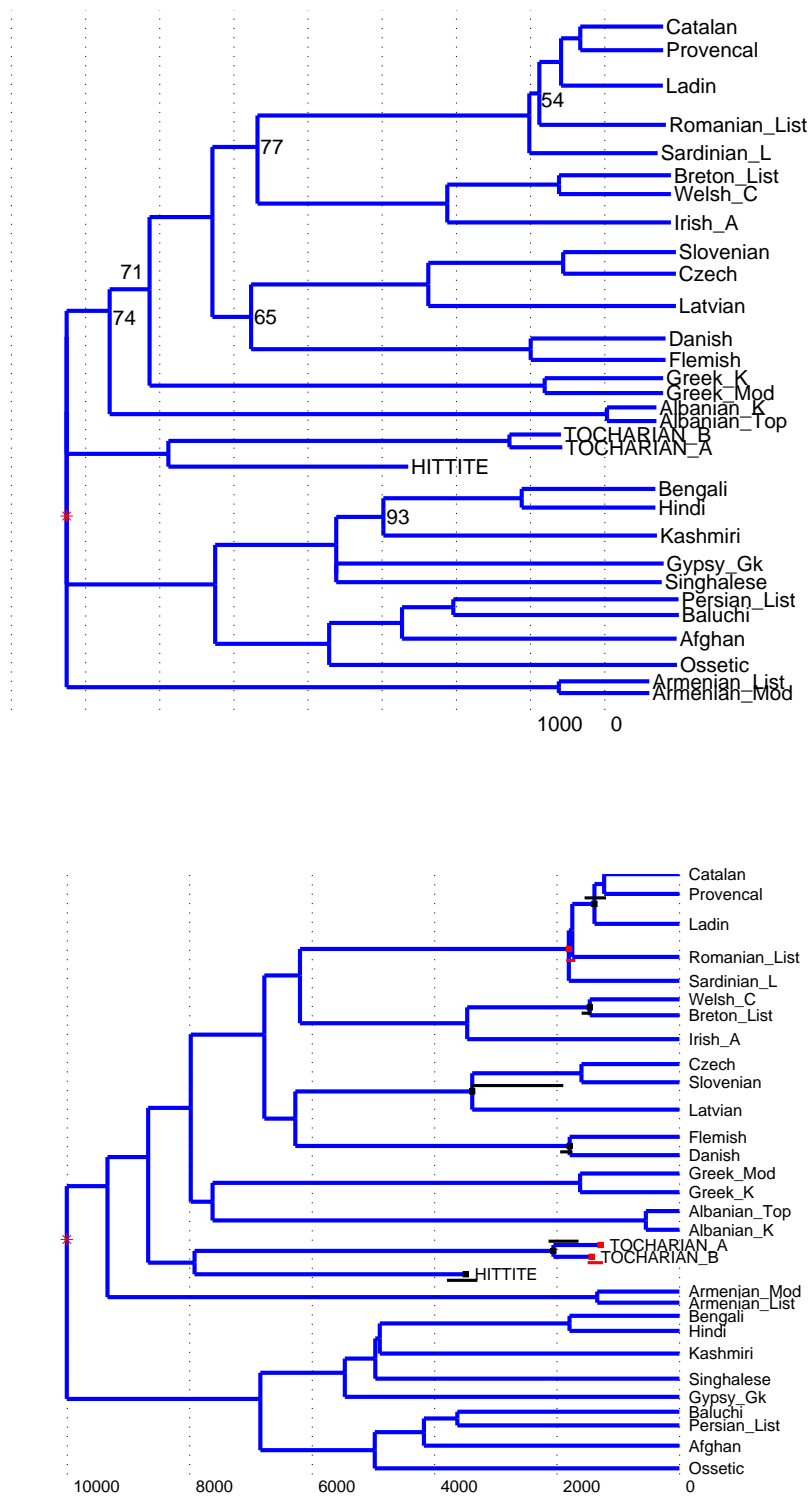


Fig. 21. Consensus tree (top) and a sampled state (bottom) illustrating the Dyen/200/31 (bottom) posterior distribution. Details as Supplement-Fig. 16.

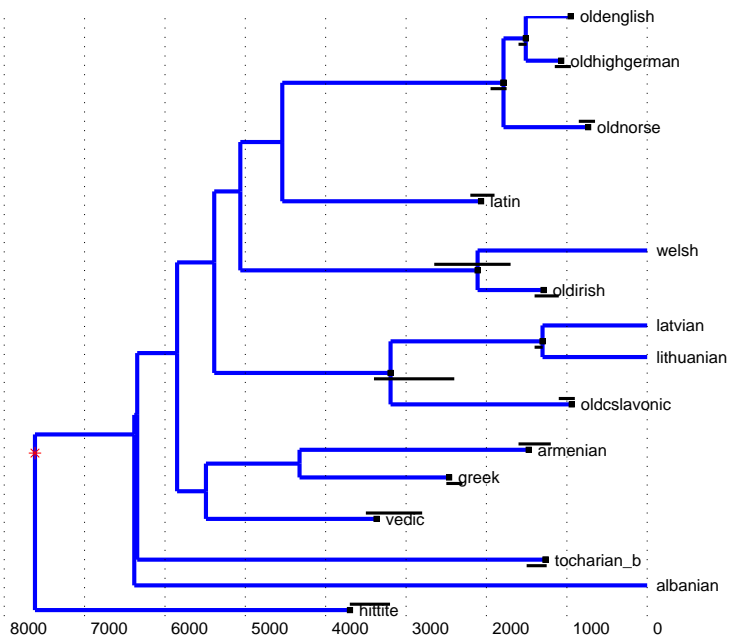
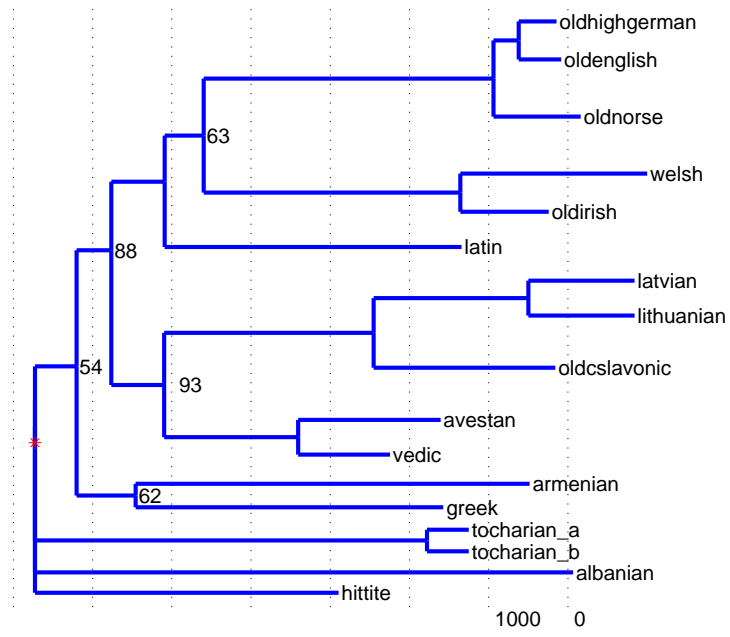


Fig. 22. Consensus tree (top) and a sampled state (bottom) illustrating the Ringe/100/17 (bottom) posterior distribution. Details as Supplement-Fig. 16.

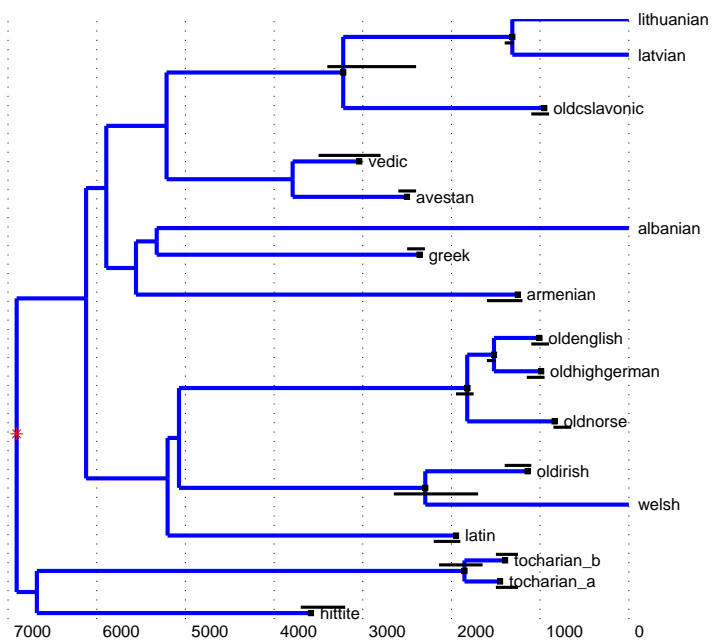
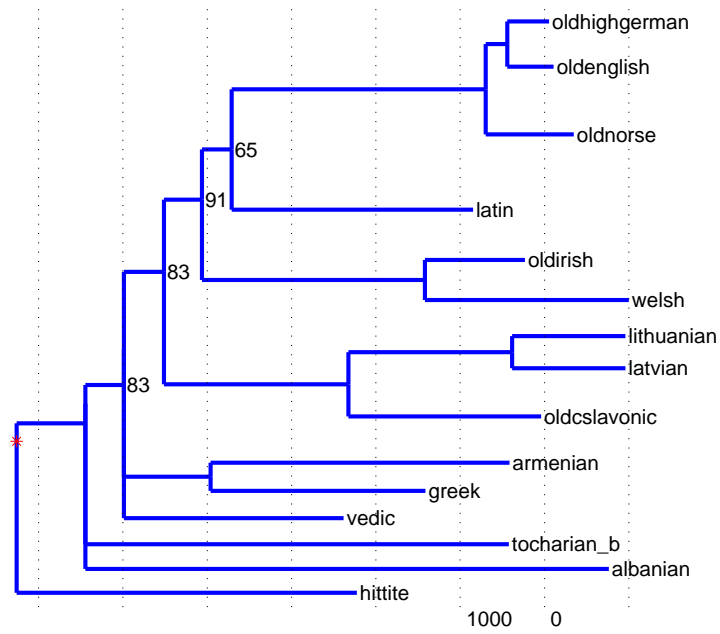


Fig. 23. Consensus tree (top) and a sampled state (bottom) illustrating the Ringe/200/15 (bottom) posterior distribution. Details as Supplement-Fig. 16.

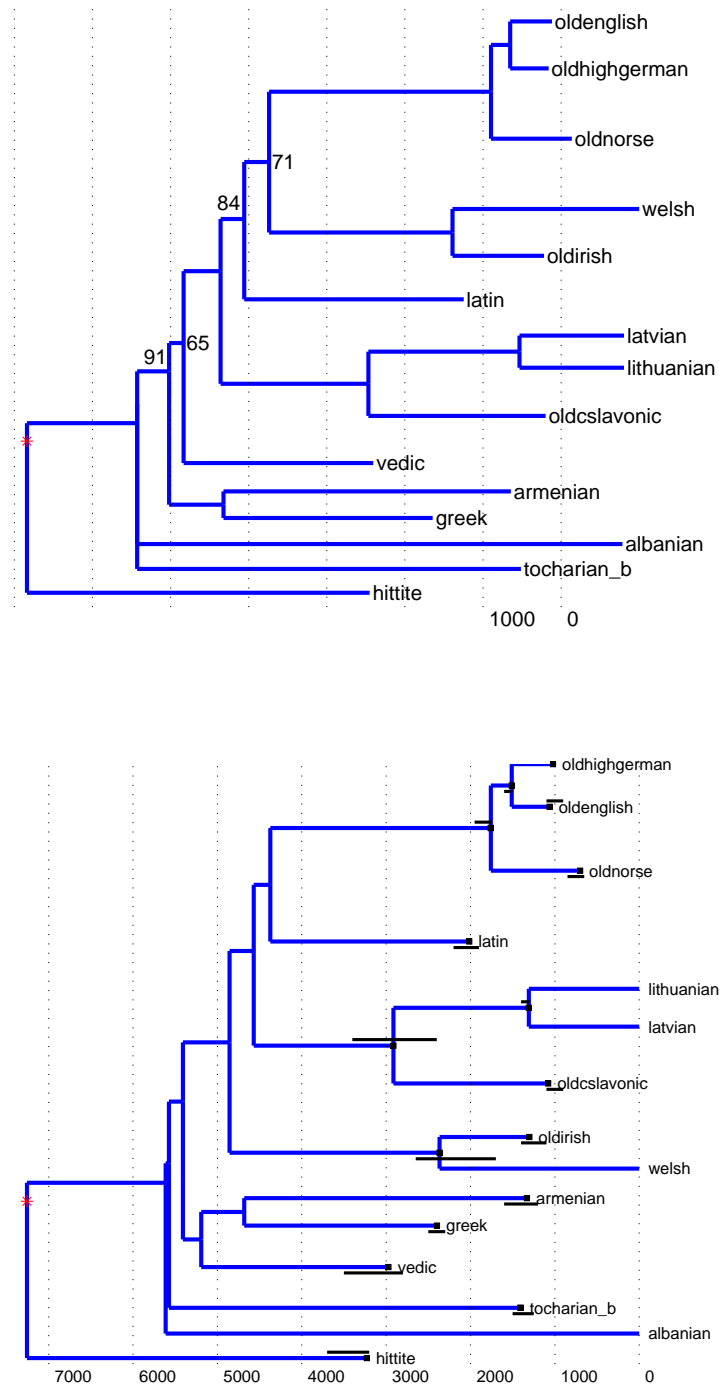


Fig. 24. Consensus tree (top) and a sampled state (bottom) illustrating the Ringe/328/15 (bottom) posterior distribution. Details as Supplement-Fig. 16.