Ancestry sampling for Indo-European phylogeny and dates.

Taraka Rama
Department of Linguistics
University of Tübingen
Email: taraka-rama.kasicheyanula@uni-tuebingen.de

***Abstract*—The date of the root of the Indo-European language family received much attention due to the application of Bayesian phylogenetic methods since the beginning of the last decade. The inferred root date of the family moved along with the development of new methods and better data. In this paper, I compare two dating techniques known as node-dating and total evidence dating for the Indo-European language family. I find that the total evidence dating based on a birth death tree infers age which is consistent with the Steppe hypothesis of spread of Indo-European languages.**

I. Introduction

The age of the Indo-European language family has been explored through the application of relaxed clock techniques that were developed for dating species divergence times (Atkinson et al., 2005; Bouckaert et al., 2012). These methods employ a binary trait matrix to infer a tree and then use internal calibration nodes to date the root of the inferred tree. The results of these methods have been shown to change if the *ancestory constraints* are imposed on the tree topology (Chang et al., 2015) search.

Chang et al. (2015) curate the IELex dataset developed by Michael Dunn and experiment on a wide range of model parameters and subsets of the original dataset. Their results converge on a date that seems to support the Indo-European Steppe hypothesis. However, Chang et al. do not employ a subset of languages that coincide with the dataset of ancient languages given by Ringe et al. (2002). In a preceding paper, Atkinson et al. (2005) employ the dataset of Ringe et al. (2002) to test the assumption of two different trait substitution matrices: Cognate loss-gain and Stochastic Dollo character evolution. Atkinson et al. (2005) find that the root date of the Indo-European tree support the Anatolian hypothesis.

Bouckaert et al. (2012) use a coalescent tree prior for inferring and dating the Indo-European tree. The authors use both tip dating as well node dating to infer the date of the root of the tree. The authors find that the date supports a Anatolian hypothesis of language expansion. Recently, Chang et al. (2015) raised the following questions:

1) Is the tree prior suitable for modeling language splitting events?
2) How to correct ascertainment bias of missing characters?
3) Are ancestry constraints justified?

The above questions can be answered in the framework of "Total Evidence Dating" (Ronquist et al., 2012a; Zhang et al., 2015). The rest of the paper is structured as followed. First, we discuss the different issues raised by the work of Chang et al. (2015) and then show how the dates can vary by employing birth-death process tree prior and a different relaxed clock.

II. Tree priors

There are at least two tree priors that are widely used in the Bayesian phylogenetics literature: Birth-death trees and Coalescent tree priors. The birth-death tree priors have been used for modeling the divergence dates of species whereas, the coalescent tree priors have been used for modeling the spread of virus across populations. Only recently have "Fossilized Birth-Death" tree priors (FBD) (Heath et al., 2014; Stadler, 2010) have been used for modeling speciation and extinction along with the observation of fossils for phylogenetics.

The FBD tree prior is a natural choice for dating the Indo-European language family since it allows the placement of fossil on a tree based on the prior information as well as the character data. The FBD tree priors do not solely depend on the character data of the fossil to date the tree. One of the main points raised by Chang et al. (2015) is that ancient languages such as Hittite are sparsely attested which might cause the tree inference program to move the root age further down the time scale to accommodate the fact there has been relatively less change on the branch leading to Hittite.

The FBD tree prior works with the following parameters:

- net speciation rate: $d = \lambda - \mu$.
- net turnover rate: $e = \mu/\lambda$.
- Species sampling probability: ρ.
- Fossil sampling proportion: $s = \psi/(\psi + \mu)$.

where, λ and μ are birth and death rates, ρ represents the probability of sampling extant languages, and ψ represents the fossil sampling probability.

In contrast, node dating works does not work directly with the fossils but works with the internal information derived from the fossils. Node dating places priors on the internal nodes that are derived from the secondary fossil information. Node dating does not require fossil sampling proportion that is required by FBD.

III. Experiments

In this section, I describe the two different experiments performed with node dating and total evidence dating. Both dating procedures assume that the extant species have been

randomly sampled. I set the value of ρ to 0.2 following the fact that there are about 400 doculects of Indo-European whereas, the dataset has 77 contemporary doculects and 20 extinct languages. The parameter settings for the total evidence dating is as followed: $d \sim Exp(1)$, $e \sim Beta(1,1)$, $s \sim Beta(1,1)$.

The root age of the tree was drawn from a uniform prior bounded between 4000 and 25000. All the fossil date priors were drawn from uniform distribution and are given in table I. I adopted topological constraints such that all the major subfamilies (table III) in Indo-European are always grouped together.

In both the experiments, the relaxed clock model is assumed to be a Independent Gamma rates model where the branch rates are drawn from a Gamma distribution (Lepage et al., 2007). The substitution model is a General Time Reversible model (GTR) where each site has a specific rate that is drawn from a Gamma distribution. The likelihood model also accounts for ascertainment bias where the coding is set to *variable* such that it accounts for the traits that are unobserved (unascertained)[1] in the data.[2]

Language	Priors	Language	Priors
Hittite	$3500-3600$	Old High German^{A}	$1000-1100$
Old Irish^{A}	$1100-1300$	Tocharian B	$1200-1500$
Classical Armenian^{A}	$1300-1600$	Tocharian A	$1200-1500$
Ancient Greek^{A}	$2400-2500$	Lycian	$2350-2450$
Luvian	$3275-3425$	Old Prussian	$500-600$
Vedic Sanskrit^{A}	$3000-3500$	Umbrian	$2100-2300$
Old English^{A}	$950-1050$	Avestan	$2450-2550$
Old Persian	$2375-2525$	Gothic	$1625-1675$
Latin^{A}	$2100-2200$	Old Norse^{A}	$750-850$
Oscan	$2100-2300$	Old Church Slavonic	$950-1050$

TABLE I
CALIBRATION DATES FOR THE ANCIENT LANGUAGES. ALL DATES ARE GIVEN AS BEFORE PRESENT (BP). ^{A} DENOTES THOSE LANGUAGES THAT ARE ASSUMED TO BE ANCESTORS OF CONTEMPORARY LANGUAGES BY CHANG ET AL. (2015).

The node dating priors are based on a mixture of information from historical and archaeological records. The internal node prior was modeled using truncated lognormal (LN) and exponential (Exp) distributions. The details of the internal node priors are given in table II.

Internal node	Node prior
Armenian	LN(1500,250,0.7)
Balto-Slavic	LN(950,500,1.0)
Germanic	LN(1635,400,0.4)
Indo-Aryan	Exp(3000,180)
Indo-Iranian	LN(3000,900,0.4)
Iranian	LN(2450,500,0.7)
Romance	Exp(2050,50)
Celtic	LN(1200,500,0.75)

TABLE II
THE LOGNORMAL DISTRIBUTION IS PARAMETRIZED AS OFFSET, μ, AND σ. THE EXPONENTIAL DISTRIBUTION IS PARAMETERIZED AS OFFSET AND λ.

[1]All characters that exhibit states 0 or 1 in all the languages are unobserved in the data.

[2]We use sites, characters, and traits interchangeably.

I show the mean, median, and 95% highest posterior density root ages of both the total evidence dating and the node dating experiments in the tables below. The consensus trees from both the experiments are given in the figures 1 and 2. The tree topology was constrained such that all established subfamilies come out separately in the tree. I did not specify that an ancient language is a ancestor of the modern languages but let the program infer if a ancient language is a ancestor or a coordinate branch to other languages in its subfamily. For example, the position of Old High German and Old English were not specified but were constrained to fall within the Germanic subfamily.

Armenian: Armenian List, Armenian Mod, Classical Armenian
Balto-Slavic: Bulgarian, Byelorussian, Czech, Latvian, Lithuanian ST, Lower Sorbian, Macedonian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian, Upper Sorbian, OLD CHURCH SLAVONIC, OLD PRUSSIAN
Germanic: Afrikaans, Danish, Dutch List, English, Faroese, Flemish, Frisian, German, Icelandic ST, Luxembourgish, Norwegian, Schwyzerdutsch, Swedish, GOTHIC, OLD ENGLISH, OLD HIGH GERMAN, OLD NORSE
Indo-Aryan: Assamese, Bengali, Bihari, Gujarati, Gypsy Gk, Hindi, Kashmiri, Lahnda, Marathi, Marwari, Nepali, Oriya, Panjabi ST, Sindhi, Singhalese, Urdu, VEDIC SANSKRIT
Iranian: Baluchi, Digor Ossetic, Kurdish, Pashto, Persian, Sariqoli, Shughni, Tadzik, Wakhi, Waziri, Zazaki, OLD PERSIAN, AVESTAN
Indo-Iranian: Iranian + Indo-Aryan
Romance: Catalan, French, Friulian, Italian, Ladin, Portuguese ST, Provencal, Romansh, Rumanian List, Sardinian C, Sardinian N, Spanish, Vlach, LATIN, OSCAN, UMBRIAN
Celtic: Breton ST, Irish B, Scots Gaelic, Welsh N, OLD IRISH
Italo-Celtic: Romance + Celtic
Nuclear-Indo-European: All languages other than Anatolian

TABLE III
TOPOLOGY CONSTRAINTS ASSUMED IN THIS PAPER. ANCIENT LANGUAGES ARE SHOWN IN SMALL CAPITALS.

The topology constraints were made to be general and not very fine-grained as Chang et al. (2015). The program automatically infers the positions of the ancient languages in their respective families based on the character data. For instance, the tree (cf. figure 1) shows Old English to be the ancestor of Modern English and this information need not be supplied to the program directly. The tree shows that Vedic Sanskrit was a common ancestor of all the modern Indo-Aryan languages; Ancient Greek is shown to be the ancestor of Modern Greek; Old Irish is placed as an ancestor of Irish B and Scots Gaelic. It has to be noted that these ancestry constraints were discovered by the method directly and need not be specified (Chang et al., 2015). The credibility interval for the root age of the total evidence dating is close to the intervals $(4860-7250)$ of Chang et al. (2015).[3]

Dating method	Mean	Median	HPD
Total Evidence Dating	6465	6434	$5186-7754$
Node dating	11265	11165	$8669-13920$

TABLE IV
AGE ESTIMATES OF THE ROOT IN "TOTAL EVIDENCE DATING" AND "NODE DATING".

[3]All our experiments were performed using MrBayes 3.2.5 (Ronquist et al., 2012b). The program was run for two different runs in both the experiments for 100 million iterations.

IV. Conclusion

I observe that the total evidence dating is largely successful at inferring the positions of ancestral languages in the Indo-European tree. I plan to experiment with different datasets under different parameter settings under the FBD tree prior to infer the root age distributions.

Acknowledgment

I thank Gerhard Jäger for the IELex dataset and the original python code.

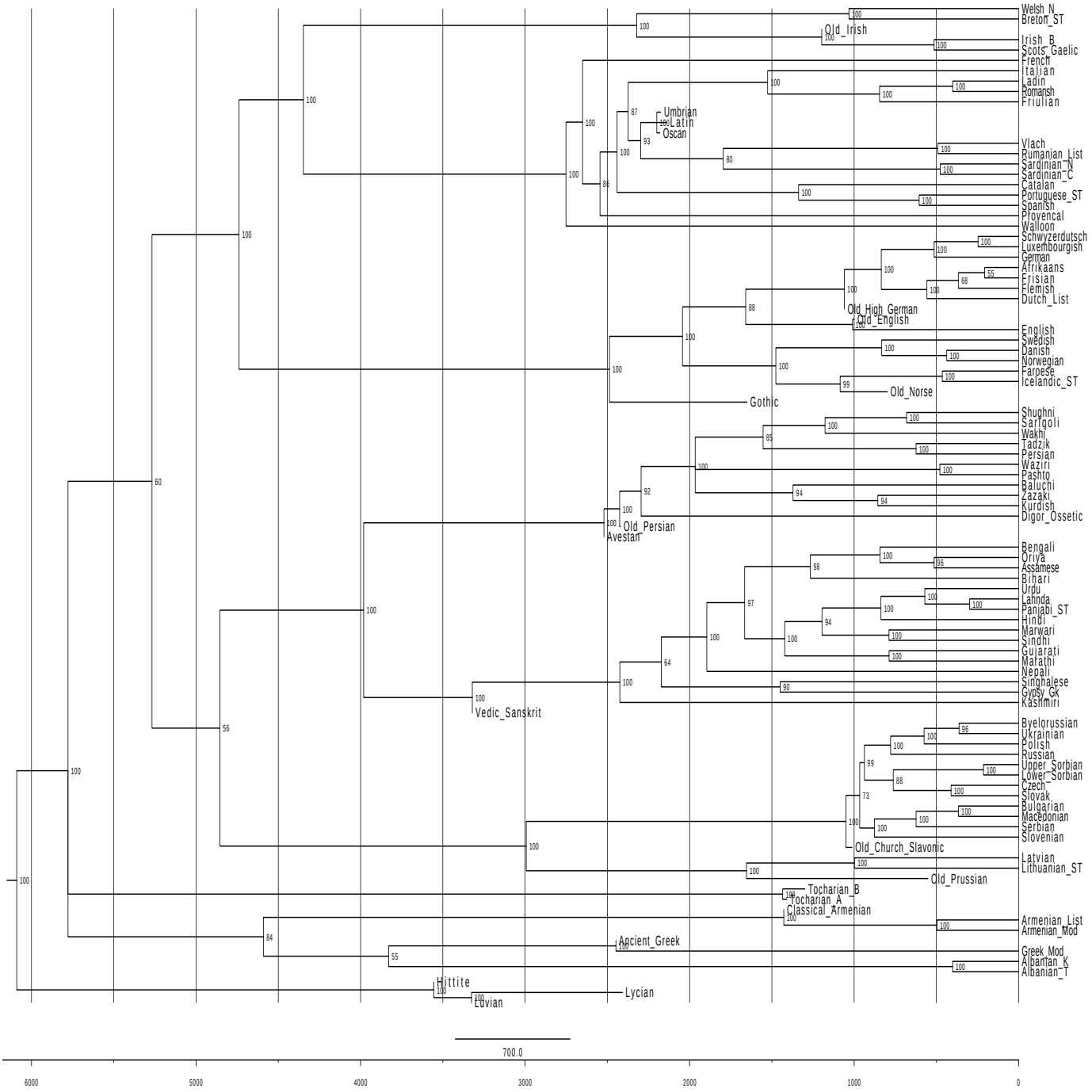

Fig. 1. The consensus tree from the total evidence dating procedure. The root age supports a Steppe hypothesis. All the internal splits of the tree show more than 50% support score.

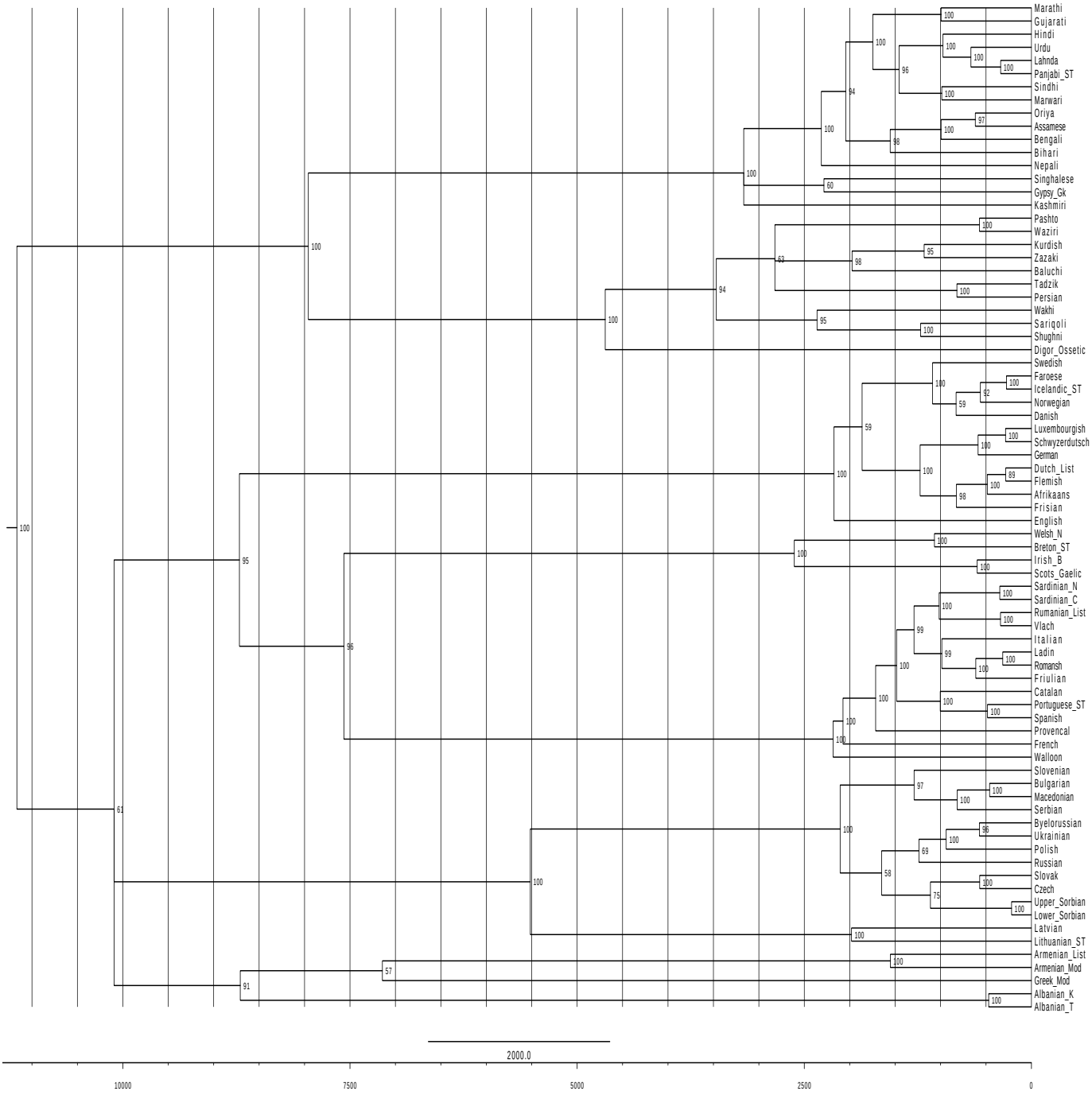

Fig. 2. The consensus tree from the node dating experiment. The root age supports a median date that falls beyond the time interval proposed by Anatolian hypothesis. All the internal splits of the tree show more than 50% support score. The method requires data from modern languages.

References

Q. Atkinson, G. Nicholls, D. Welch, and R. Gray, "From words to dates: water into wine, mathemagic or phylogenetic inference?" *Transactions of the Philological Society*, vol. 103, no. 2, pp. 193–219, 2005.

R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson, "Mapping the origins and expansion of the Indo-European language family," *Science*, vol. 337, no. 6097, pp. 957–960, 2012.

W. Chang, C. Cathcart, D. Hall, and A. Garrett, "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis," *Language*, vol. 91, no. 1, pp. 194–244, 2015.

D. Ringe, T. Warnow, and A. Taylor, "Indo-European and computational cladistics," *Transactions of the Philological Society*, vol. 100, no. 1, pp. 59–129, 2002.

F. Ronquist, S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn, "A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera," *Systematic Biology*, vol. 61, no. 6, pp. 973–999, 2012.

C. Zhang, T. Stadler, S. Klopfstein, T. A. Heath, and F. Ronquist, "Total-evidence dating under the fossilized birth–death process," *Systematic biology*, p. syv080, 2015.

T. A. Heath, J. P. Huelsenbeck, and T. Stadler, "The fossilized birth–death process for coherent calibration of divergence-time estimates," *Proceedings of the National Academy of Sciences*, vol. 111, no. 29, pp. E2957–E2966, 2014.

T. Stadler, "Sampling-through-time in birth–death trees," *Journal of Theoretical Biology*, vol. 267, no. 3, pp. 396–404, 2010.

T. Lepage, D. Bryant, H. Philippe, and N. Lartillot, "A general comparison of relaxed molecular clock models," *Molecular biology and evolution*, vol. 24, no. 12, pp. 2669–2680, 2007.

F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck, "Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space," *Systematic biology*, vol. 61, no. 3, pp. 539–542, 2012.