Bioinformatics, 2024, **40(4)**, btae162
https://doi.org/10.1093/bioinformatics/btae162
Advance Access Publication Date: 21 March 2024
Applications Note

Phylogenetics

AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss

Benoit Morel [1,2], **Tom A. Williams**[3], **Alexandros Stamatakis** [1,2,4], **Gergely J. Szöllősi**[5,6,7,*]

[1]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg 69118, Germany
[2]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany
[3]School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom
[4]Institute of Computer Science, Biodiversity Computing Group, Heraklion GR-70013, Greece
[5]ELTE-MTA "Lendület", Evolutionary Genomics Research Group, Budapest H-1117, Hungary
[6]Institute of Evolution, HUN-REN Centre for Ecological Research, Budapest H-1121, Hungary
[7]Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan

*Corresponding author. Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna, Kunigami District, Okinawa 904-0412, Japan. E-mail: Gergely.Szollosi@oist.jp (G.J.S.)

Associate Editor: Russell Schwartz

Abstract

Motivation: Genomes are a rich source of information on the pattern and process of evolution across biological scales. How best to make use of that information is an active area of research in phylogenetics. Ideally, phylogenetic methods should not only model substitutions along gene trees, which explain differences between homologous gene sequences, but also the processes that generate the gene trees themselves along a shared species tree. To conduct accurate inferences, one needs to account for uncertainty at both levels, that is, in gene trees estimated from inherently short sequences and in their diverse evolutionary histories along a shared species tree.

Results: We present AleRax, a software that can infer reconciled gene trees together with a shared species tree using a simple, yet powerful, probabilistic model of gene duplication, transfer, and loss. A key feature of AleRax is its ability to account for uncertainty in the gene tree and its reconciliation by using an efficient approximation to calculate the joint phylogenetic—reconciliation likelihood and sample reconciled gene trees accordingly. Simulations and analyses of empirical data show that AleRax is one order of magnitude faster than competing gene tree inference tools while attaining the same accuracy. It is consistently more robust than species tree inference methods such as SpeciesRax and ASTRAL-Pro 2 under gene tree uncertainty. Finally, AleRax can process multiple gene families in parallel thereby allowing users to compare competing phylogenetic hypotheses and estimate model parameters, such as duplication, transfer, and loss probabilities for genome-scale datasets with hundreds of taxa.

Availability and implementation: GNU GPL at https://github.com/BenoitMorel/AleRax and data are made available at https://cme.h-its.org/exelixis/material/alerax_data.tar.gz.

1 Introduction

Simultaneously inferring gene trees and the species tree is challenging. Genomes contain abundant information about evolutionary history, but single gene sequences are often too short to reliably resolve gene trees (Haag *et al.* 2022). Moreover, gene trees are not species trees, but each is the unique result of series of evolutionary events. Starting from an individual site in a genome up to the species level, a hierarchy of evolutionary processes generate genomic sequences, with each level of the hierarchy contributing to the phylogenetic signal that can induce differences between reconstructed gene trees (Szöllősi *et al.* 2015). Segregating mutations that cross speciation events (a process called incomplete lineage sorting) leave topological signatures in gene trees. Duplications, transfers, and losses of genes (DTL) lead to substantial differences in the size and phylogenetic distribution of families of homologous genes, and at the same time induce patent phylogenetic discord.

While most species tree inference methods take single-copy genes as input that are assumed *a priori* to be orthologous (i.e. to have originated from a common ancestral gene solely through speciation), more recent methods are able to handle more general homologous gene families (i.e. genes that originated from a common ancestor through speciation, gene duplication and possibly transfer events) including multi-copy ones. DupTree (Wehe *et al.* 2008) searches for the species tree with the most parsimonious reconciliation cost, measured as the number of duplication events. STAG (Emms and Kelly 2018) infers a species tree by applying a distance method to each gene family that covers *all* species, and subsequently builds a consensus tree from all these distance-based trees. FastMulRFS (Molloy and Warnow 2020) extends the definition of the

Received: 11 October 2023; **Revised:** 30 January 2024; **Editorial Decision:** 4 February 2024; **Accepted:** 19 March 2024

Robinson-Foulds (RF) distance to multi-copy gene trees and strives to minimize this distance between the species tree and all input gene trees. ASTRAL-Pro 2 (Zhang and Mirarab 2022) infers a species tree from multi-copy gene trees by maximizing a duplication-aware quartet score. We recently developed SpeciesRax (Morel *et al.* 2022), a tool that models gene DTL events and that estimates a maximum likelihood (ML) rooted species tree under the so-called UndatedDTL model. However, all those methods only consider a single *estimated* gene tree per gene family as input, and are thus sensitive to gene tree uncertainty. Alternative methods such as Phyldog (Boussau *et al.* 2013) jointly estimate the species tree *and* the gene trees that evolved along it via speciation, duplication, and loss. However, Phyldog does not scale to large datasets, does not model gene transfer, and is less accurate than SpeciesRax and Astral-Pro on simulations, even in the absence of gene transfers (Morel *et al.* 2022).

Another challenge is to infer a reconciled gene tree: that is, a gene tree together with a reconciliation, or series of DTL and speciation events that trace its evolution along the species tree. To this end, we previously developed GeneRax (Morel *et al.* 2020), a probabilistic method that searches for the reconciled gene trees that maximize the joint likelihood, defined as the product of the phylogenetic likelihood (the probability of the multiple sequence alignment (MSA) given the gene tree) and the reconciliation likelihood (the probability of the gene tree given the species tree under UndatedDTL). However, GeneRax does not provide a confidence measure for the inferred gene trees. Amalgamated Likelihood Estimation (ALE) (Szöllősi *et al.* 2013) is an alternative two-step approach that sums the joint likelihood over all reconciled gene trees. The probabilities of the gene trees are estimated based on the conditional clade probabilities of their constituent clades (Höhna and Drummond 2012, Larget 2013). Confidence in a particular DTL event can then be estimated by counting how frequently it appears in the different reconciled gene trees. However, ALE is not able to simultaneously process multiple gene families, or to infer the species tree.

Here we present AleRax, a novel probabilistic method for phylogenetic tree inference that can perform both species tree inference *and* reconciled gene tree inference from a sample of gene trees. We first present the method and the key features of AleRax. Then, we present the results of our benchmarks: we found that AleRax is on par with ALE in terms of gene tree reconstruction accuracy, but an order of magnitude faster and robust to numerical underflow that sometimes causes ALE to fail. Based on simulations, AleRax is 25% more accurate than SpeciesRax and twice as accurate as ASTRAL-Pro 2 for species tree inference, but at least one order of magnitude slower than those two methods.

2 Method

2.1 Likelihood definition and computation

Let S be the species tree, and A be an alignment of homologous genes, that is, a gene family alignment. The probability of observing A given S is obtained by summing over all possible rooted gene trees G for alignment A:

$$P(A|S) = \sum_G P(A|G)P(G|S) \quad (1)$$

(Felsenstein 1988, Szöllősi *et al.* 2013).

AleRax approximates $P(A|S)$ by using the ALE (Szöllősi *et al.* 2013) algorithm, which takes as input a sample of gene trees for each gene family, ideally computed via a Bayesian inference tool such as Phylobayes (Lartillot *et al.* 2013) or MrBayes (Ronquist *et al.* 2012) prior to running AleRax. The gene trees in the samples can be rooted or unrooted. Instead of naïvely computing the sum $\sum_G P(A|G)P(G|S)$, AleRax uses the sample of gene trees provided to estimate a nonzero probability for all gene tree topologies that can be "amalgamated" (David and Alm 2011) from clades present in the sample using conditional clade probabilities (CCPs) (Höhna and Drummond 2012, Larget 2013). This corresponds to the maximum entropy distribution given marginal split frequencies in the sample of gene trees provided as input (Szöllősi *et al.* 2013). Using CCPs allows AleRax to exploit a joint dynamic programming recursion that efficiently and accurately approximates Equation (1) (Szöllősi *et al.* 2013). An in-depth description of this algorithm is provided in the Supplementary Material. Importantly, unlike ALE, AleRax uses arbitrary-precision floating point values when necessary, allowing it to process datasets that cause ALE to fail due to numerical underflow. AleRax can also process multiple gene families in parallel given a sample of genes for each family allowing it to infer the species tree as well as more realistic DTL models, e.g. branch-wise DTL parameters on the species tree, by maximizing the likelihood of the species tree, which can be written as the product over all per-family alignment probabilities:

$$\mathcal{L}(S|A) = \prod_i P(A_i|S). \quad (2)$$

2.2 Species tree inference

AleRax searches for the rooted ML species tree. It implements the same tree search strategy as SpeciesRax (Morel *et al.* 2022), which applies hill-climbing subtree prune and regraft (SPR) moves until it cannot find a move that yields a tree with a better likelihood. However, unlike SpeciesRax, which relies on a single gene tree per gene family, AleRax uses a sample of per family gene trees to integrate over gene tree uncertainty by approximating $P(A_i|S)$. AleRax recomputes the model parameters (described in the next subsection) after each round of SPR moves. The starting species tree can be generated at random, estimated with MiniNJ (Morel *et al.* 2022) (a duplication-aware distance method), or specified by the user.

2.3 Model parameter estimation

In general, the reconciliation model used by AleRax, the so called UndatedDTL model (Morel *et al.* 2020) can have three free parameters per gene family k and per species tree branch e: δ_e^k, τ_e^k, and λ_e^k, representing the duplication, transfer, and loss probabilities, respectively. However, estimating such a large number of free parameters (proportional to the product of the number of species in S multiplied by the number of gene families) can lead to over-parameterization. Hence, AleRax implements three distinct approaches to parameter estimation that can be chosen by the user. In the *global parameters* mode, all families and all species share the same set of parameters $\{\delta, \tau, \lambda\}$, resulting in only three free parameters. In the *per-family parameters* mode, AleRax optimizes a different set of parameters $\{\delta_k, \tau_k, \lambda_k\}$ for each gene family k. This corresponds to $3K$ free parameters, where K is the number of gene families. In the *per-species parameter* mode, the

user can provide a list of species groupings that share the same set of parameters. In this mode, the model has $3n$ free parameters, where n is the number of groupings defined. In each mode, AleRax optimizes the model parameter values with respect to the likelihood via gradient descent. Our experiments suggest that AleRax performs similarly in terms of gene tree and species tree accuracy under the different modes.

2.4 Sampling reconciled gene trees

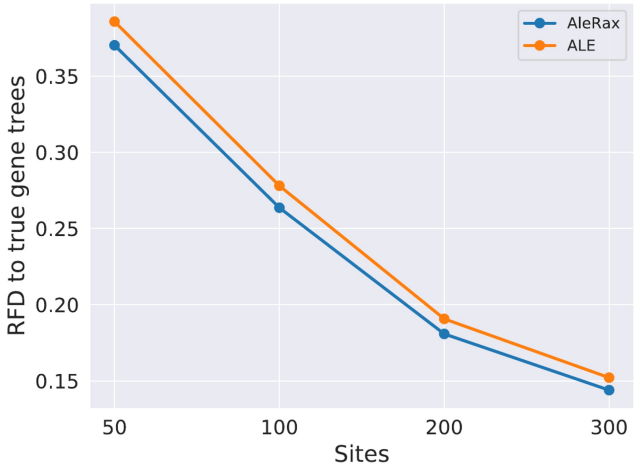
AleRax samples, per gene family, r reconciled gene trees proportional to their joint likelihood using stochastic backtracking following the dynamic programming recursion (Szöllősi *et al.* 2013) used to approximate Equation (1). We describe the algorithm in the Supplementary material. The number of gene trees r to sample is specified by the user. For each gene family, AleRax outputs the r reconciled gene trees, the rooted majority-rule consensus gene tree, and the number of gene events per species, averaged over all output samples for this specific gene family. The reconciliations are stored in RecPhyloXML format (Duchemin *et al.* 2018) and can be converted into SVG figures using Thirdkind (Penel *et al.* 2022). AleRax also outputs the species pairs involved in horizontal gene transfers, sorted by the number of times those transfers have been sampled (per family and summed over all families). In addition, for each species, it returns the number of ancestral gene copies and the number of DTL events. We also provide scripts to extract lists of families that were involved in transfers between specific species pairs, or families that experienced a specific event for a given species (e.g., all families with a gene duplication in *Arabidopsis thaliana*).

3 Results

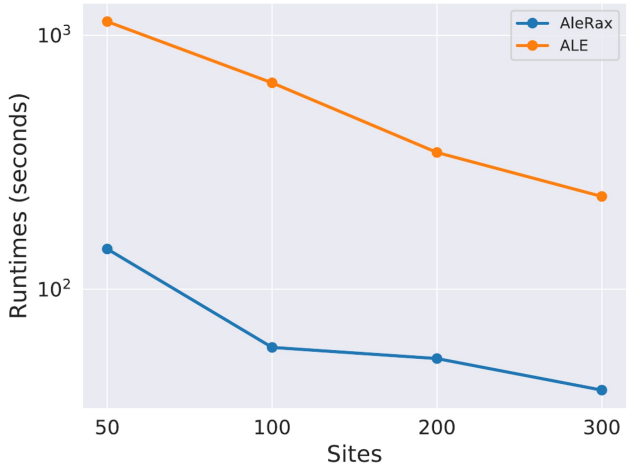
Here we only provide a summary of the experiments and their results. We describe them in detail in the Supplementary material.

3.1 Gene tree reconciliation

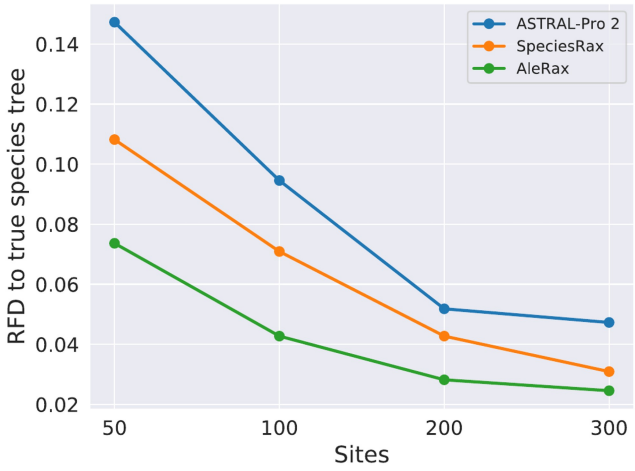
We ran both ALE and AleRax on the simulated datasets used to benchmark SpeciesRax (Morel *et al.* 2022). As expected, ALE and AleRax perform analogously in terms of accuracy (Fig. 1). We also executed both tools on a real dataset derived from the HOGENOM (Penel *et al.* 2009) database's HOGENOM-CORE subdataset, which contains 666



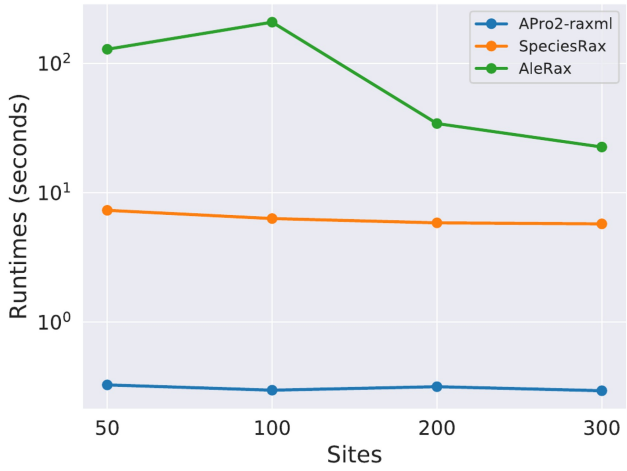
(a) Gene tree inference accuracy



(b) Gene tree inference runtime



(c) Species tree inference accuracy



(d) Species tree inference runtime

Figure 1. Results on simulated datasets for varying average gene sequence lengths (sites), with 100 gene families and 25 species. When inferring gene trees, the true species tree is assumed to be known.

representative genomes spanning the diversity of cellular life. ALE failed for 25 families (out of 12 408) because of numerical underflow. AleRax in contrast successfully processed all families. In terms of runtime, AleRax was on average an order of magnitude faster than ALE, both on simulated and empirical datasets (Fig. 1).

3.2 Species tree inference

We investigated the impact of using gene tree distributions instead of single ML gene trees by comparing AleRax, SpeciesRax, and ASTRAL-Pro 2 on simulated datasets. We used the same benchmarks that were employed for assessing SpeciesRax (Morel *et al.* 2022), that is, we varied different simulation parameters individually while keeping the rest fixed, and measured the distance between the true and the inferred species trees. We ran AleRax on gene tree distributions inferred with MrBayes (Ronquist *et al.* 2012). We found that, over all conditions tested, AleRax is on average 25% more accurate than SpeciesRax and twice as accurate as ASTRAL-Pro 2. For instance, Fig. 1 compares the accuracy of the different methods when varying the average gene sequence lengths. Our explanation is that using fixed gene tree estimates does not capture all the information contained in the gene sequences. Treating gene trees as latent variables allows AleRax to avoid this information loss, and thus to be more robust to phylogenetic uncertainty. When using gene tree distributions with 1000 gene trees per gene family, AleRax is from one to two orders of magnitude slower than SpeciesRax (Fig. 1d).

To investigate the performance of rooted species tree inference with AleRax on real data, we returned to an earlier phylogenetic analysis of the archaeal domain of life (Williams *et al.* 2017, Davín *et al.* 2018). The Archaea constitute one of the two prokaryotic domains of life, but their early evolution and the position of the root remains poorly understood. Previous studies, including a 2017 analysis by some of us, have placed the root between the genome-reduced DPANN Archaea and the other lineages (Williams *et al.* 2017, Dombrowski *et al.* 2020, Aouad *et al.* 2022). In the 2017 study, we used ALE to infer the optimal root position based on a fixed unrooted topology estimating using concatenation of a set of 45 vertically-evolving marker genes. This hybrid strategy was employed because ALE cannot search the space of rooted species trees. We performed AleRax, SpeciesRax, and ASTRAL-Pro-2 analyses on the set of 5379 gene families, using the same MCMC gene tree samples to represent gene tree uncertainty as in the original study in the case of AleRax. On a 40-cores machine, Astral-Pro 2 finished after 15 s, SpeciesRax after 167 s, and AleRax after 9215 s. AleRax inferred a rooted species tree closely similar to the original study and to other recent analyses (Dombrowski *et al.* 2020, Aouad *et al.* 2022), with a root between DPANN Archaea and the rest, and recovering the monophyly of major archaeal groups including DPANN, TACK Archaea, and most of the Euryarchaeota (see Fig. 2), although two incomplete genomes (SAG I15 and Thaumarchaeota E09) were likely misplaced. The SpeciesRax tree was heterodox, recovering a mixture of TACK Archaea, Euryarchaeota and DPANN on both sides of the root (Fig. 2). The unrooted species tree inferred by ASTRAL-Pro-2 was very similar to the topology of the AleRax tree (Fig. 2). If current views about archaeal phylogeny are broadly correct, this analysis suggests that AleRax can be substantially more accurate than SpeciesRax when inferring the root and topology of species trees from real data, likely because of its ability to model gene tree uncertainty. The analysis also demonstrates that reconciliation methods can capture information about the root and topology of species trees, and that in the case of the Archaea this signal is congruent with that from traditional concatenation-based phylogenetics. In this case study, AleRax and ASTRAL-Pro-2 both recovered species trees that are consistent with current views. While ASTRAL-Pro-2 was much faster, the AleRax analysis provided an estimate of the root of the species tree, reconciliation scenarios for each gene family, and gene content reconstructions for each internal node of the tree. These comparative genomic outputs are an important part of the use case for AleRax and similar reconciliation tools, and can be used (for example) to propose hypotheses about ancestral gene repertoires and organismal characteristics, including metabolisms (Williams *et al.* 2017).

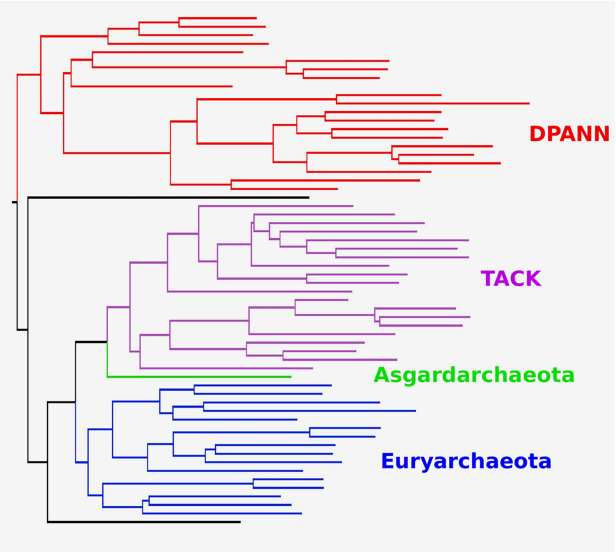


Figure 2. Archaeal tree inferred with AleRax, with the four main groups: DPANN, Euryarchaeota, TACK, and Asgardarchaeota.

4 Conclusion and future work

AleRax is an efficient and user-friendly tool for species tree inference and gene tree-species tree reconciliation that can be applied to datasets from across the tree of life. It is particularly useful in biological applications where not only the species tree, but also the histories individual of gene families are of interest. We showed that AleRax is on par with ALE in terms of reconciled gene tree accuracy, while being one order of magnitude faster and more robust to numerical errors. Furthermore, it infers more accurate species trees than SpeciesRax and ASTRAL-Pro 2, because it can accommodate gene tree uncertainty. In the future, we plan to improve our model to accommodate DTL rate heterogeneity over species, incomplete lineage sorting, and horizontal gene transfer time constraints.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Funding

B.M. and A.S. are financially supported by the Klaus Tschira Foundation, by DFG grant STA 860/6–2, and by the European

Union (EU) under Grant Agreement No 101087081 (CompBiodiv-GR). G.S. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 714774 and the grant GINOP-2.3.2.–15-2016–00057. This work was funded by the Gordon and Betty Moore Foundation through grant GBMF9741 to T.A.W. and G.S.

Data availability

All data are incorporated into the article and its online supplementary material.

References

Aouad M, Flandrois J-P, Jauffrit F *et al.* A divide-and-conquer phylogenomic approach based on character supermatrices resolves early steps in the evolution of the archaea. *BMC Ecol Evol* 2022;**22**:1–12.

Boussau B, Szöllosi GJ, Duret L *et al.* Genome-scale coestimation of species and gene trees. *Genome Res* 2013;**23**:323–30.

David LA, Alm EJ. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 2011;**469**:93–6.

Davín AA, Tannier E, Williams TA *et al.* Gene transfers can date the tree of life. *Nat Ecol Evol* 2018;**2**:904–9.

Dombrowski N, Williams TA, Sun J *et al.* Undinarchaeota illuminate dpann phylogeny and the impact of gene transfer on archaeal evolution. *Nat Commun* 2020;**11**:3939.

Duchemin W, Gence G, Arigon Chifolleau A-M *et al.* RecPhyloXML: a format for reconciled gene trees. *Bioinformatics* 2018;**34**:3646–52.

Emms D, Kelly S. Stag: species tree inference from all genes. *bioRxiv* 2018.

Felsenstein J. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet* 1988;**22**:521–65.

Haag J, Höhler D, Bettisworth B *et al.* From easy to hopeless—predicting the difficulty of phylogenetic analyses. *Mol Biol Evol* 2022;**39**:msac254.

Höhna S, Drummond AJ. Guided tree topology proposals for bayesian phylogenetic inference. *Syst Biol* 2012;**61**:1–11.

Larget B. The estimation of tree posterior probabilities using conditional clade probability distributions. *Syst Biol* 2013;**62**:501–11.

Lartillot N, Rodrigue N, Stubbs D *et al.* Phylobayes mpi: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* 2013;**62**:611–5.

Molloy EK, Warnow T. FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics* 2020;**36**:i57–i65.

Morel B, Kozlov AM, Stamatakis A *et al.* GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol Biol Evol* 2020;**37**:2763–74.

Morel B, Schade P, Lutteropp S *et al.* Speciesrax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *Mol Biol Evol* 2022;**39**:msab365.

Penel S, Arigon A-M, Dufayard J-F *et al.* Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 2009;**10**:S3–13.

Penel S, Menet H, Tricou T *et al.* Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation. *Bioinformatics* 2022;**38**:2350–2.

Ronquist F, Teslenko M, van der Mark P *et al.* MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012;**61**:539–42.

Szöllősi GJ, Rosikiewicz W, Boussau B *et al.* Efficient exploration of the space of reconciled gene trees. *Syst Biol* 2013;**62**:901–12.

Szöllősi GJ, Tannier E, Daubin V *et al.* The inference of gene trees with species trees. *Syst Biol* 2015;**64**:e42–e62.

Wehe A, Bansal MS, Burleigh JG *et al.* DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* 2008;**24**:1540–1.

Williams TA, Szöllősi GJ, Spang A *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci U S A* 2017;**114**:E4602–E4611.

Zhang C, Mirarab S. Astral-pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* 2022;**38**:4949–50.

Bioinformatics, 2024, **40(4)**, btae162
https://doi.org/10.1093/bioinformatics/btae162
Advance Access Publication Date: 21 March 2024
Applications Note