SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss

Benoit Morel, Paul Schade, Sarah Lutteropp, Tom A. Williams, Gergely J. Szöllösi, and Alexandros Stamatakis

1Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
2Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
3School of Biological Sciences, University of Bristol, Bristol, United Kingdom
4ELTE-MTA “Lendület” Evolutionary Genomics Research Group, Budapest, Hungary
5Department of Biological Physics, Eötvös University, Budapest, Hungary
6Institute of Evolution, Centre for Ecological Research, Budapest, Hungary

*Corresponding author: E-mail: benoit.morel@h-its.org.

Abstract

Species tree inference from gene family trees is becoming increasingly popular because it can account for discordance between the species tree and the corresponding gene family trees. In particular, methods that can account for multicycle gene families exhibit potential to leverage paralogy as informative signal. At present, there does not exist any widely adopted inference method for this purpose. Here, we present SpeciesRax, the first maximum likelihood method that can infer a rooted species tree from a set of gene family trees and can account for gene duplication, loss, and transfer events. By explicitly modeling events by which gene trees can depart from the species tree, SpeciesRax leverages the phylogenetic rooting signal in gene trees. SpeciesRax infers species tree branch lengths in units of expected substitutions per site and branch support values via paralogy-aware quartets extracted from the gene family trees. Using both empirical and simulated data sets we show that SpeciesRax is at least as accurate as the best competing methods while being one order of magnitude faster on large data sets at the same time. We used SpeciesRax to infer a biologically plausible rooted phylogeny of the vertebrates comprising 188 species from 31,612 gene families in 1 h using 40 cores. SpeciesRax is available under GNU GPL at https://github.com/BenoitMorel/GeneRax and on BioConda.

Key words: species tree inference, gene family tree, maximum likelihood, gene duplication, horizontal gene transfer, gene loss.

Introduction

Species tree inference is a challenging fundamental problem in evolutionary biology. Species trees provide a framework for interpreting biological diversity, determining evolutionary relationships, studying the evolution of key traits, and for understanding the underlying processes of evolution. But existing methods for species tree inference have widely appreciated limitations (Dagan and Martin 2006; Bryant and Hahn 2020). New methods that can make better use of genome-scale data, including multicopy gene families, resolve long-standing debates about challenging nodes in the tree of life and the processes of genome evolution that underpin major transitions in evolution.

Perhaps, the most widely used approach for species tree inference is concatenation (also known as the supermatrix approach). Here, per-gene sequences are first aligned into per-gene multiple sequence alignments (MSAs) and subsequently concatenated into a single, large supermatrix. Then, statistical tree inference methods (maximum likelihood [Kozlov et al. 2019; Minh et al. 2020] or Bayesian inference [Aberer et al. 2014; Lartillot et al. 2009; Ronquist et al. 2012]) are applied to infer a tree on these supermatrices. The concatenation approach heavily depends on accurate orthology inference, which still constitutes a challenging problem (Altenhof et al. 2019). In addition, concatenation methods were shown to be statically inconsistent under the multispecies coalescent model (Kubatko and Degnan 2007; Mendes and Hahn 2018) because of potential incomplete lineage sorting (ILS). Furthermore, supermatrix analyses can be misled in unpredictable ways if horizontal gene transfers (HGTs) are included in the concatenated supermatrix, for instance when analyzing microbial data (Williams and Emsley 2014; Dombrowski et al. 2020). Hybridization (Elworth et al. 2019; Lutteropp et al. 2021), and recombination (Posada 2000) constitute other
sources of incongruence between the species tree and the
per-gene sequences that might cause the supermatrix ap-
proach to fail.

As gene family tree (GFT) methods can alleviate some of
the pitfalls of the supermatrix approach, they are becoming
increasingly popular. GFT methods can take into account
that the evolutionary histories of the gene trees and the
species tree are discordant due to biological phenomena such
as ILS, gene duplication, gene loss, and HGT. In addition, some
GFT methods can estimate the root of the species tree with-
out the need for an outgroup. Character evolution tracing or
phylogenetic post analysis methods such as dating or species
delimitation require a rooted tree as input, to identify ances-
tor–descendant relationships. But inferring the root can be
difficult using the standard outgroup method when the out-
group is distant or when no obvious outgroup is available
(Shavit et al. 2007).

At present, the most widely used GFT tools only model ILS
(Bouckaert et al. 2014; Zhang et al. 2018) and are limited to
single-copy gene families. Several tools that can handle multi-
copy gene families have been developed (Wehe et al. 2008;
Boussau et al. 2012; Zhang et al. 2020; Molloy and Warnow
2020), but, despite the potential of these approaches, they
have not yet been widely adopted. This may be due to the
unfamiliarity of the software and, for some tools, the lack of
an efficient, scalable, and user-friendly approach. This situa-
tion is beginning to change: ASTRAL-Pro can deal with dupli-
cations and losses, and SpeciesRax—the tool we introduce
here—can account for duplications, losses and transfers,
which may be particularly important in the analysis of micro-
bial genomes.

One class of existing methods to infer species trees from
multiple-copy gene families attempts to simultaneously esti-
mate the GFTs and the species tree (Boussau et al. 2012; de
Oliveira Martins and Posada 2017). However, these methods
are computationally demanding and are limited to small data
sets comprising less than 100 species.

Another class of existing methods handles the GFT infer-
ence and the species tree inference steps separately. As input
they require a set of given, fixed GFTs and do not attempt to
correct the GFTs during the species tree inference step. DupTree
(Wehe et al. 2008) and DynaDUP (Bayzid et al. 2013) search
for the species tree with the most parsimonious reconcilia-
tion cost, measured as the number of duplication events in
DupTree, and the sum of duplication and loss events in DynaDUP.
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 consen-
sus tree from all these distance-based trees. However,
STAG ignores a substantial fraction of signal by discarding
gene families that do not cover all species. FastMulRFS
(Molloy and Warnow 2020) extends the definition of the
Robinson–Foulds (RF) distance to multiple-copy GFTs and
strives to minimize this distance between the species tree and
all input GFTs. ASTRAL-Pro (Zhang et al. 2020) is a promising
improvement of ASTRAL that can handle multiple-copy
GFTs. ASTRAL-Pro uses dynamic programming to infer the
species tree that maximizes a novel measure of quartet
similarity that accounts for orthology as well as paralogy.
All of the above methods are nonparametric and do not
deploy a probabilistic model of evolution. In addition, none
of them explicitly models HGT.

Here, we present SpeciesRax, the first maximum likelihood
method for inferring a rooted species tree from a set of GFTs
in the presence of gene duplication, gene loss, and HGT. We
implemented it in the GeneRax framework, our recently pub-
lished species-tree-aware GFT correction tool (Morel et al.
2020). SpeciesRax takes as input a set of MSAs and/or a set
of GFTs. If MSAs are provided, SpeciesRax will infer one max-
imum likelihood GFT tree per gene family using RaxML-NG
(Kozlov et al. 2019). Thereafter, SpeciesRax first generates an
initial, reasonable (i.e., nonrandom) species tree by applying
MiniNJ (which we also introduce in this paper), our novel
distance-based method for species tree inference from GFTs
in the presence of paralogy. MiniNJ shows similar accuracy as
other nonparametric methods while being at least two orders
of magnitude faster on large data sets. Finally, SpeciesRax
executes a maximum likelihood tree search heuristic under
an explicit statistical gene loss, gene duplication, and HGT
model starting from the MiniNJ species tree. When the spe-
cies tree search terminates, SpeciesRax calculates approxi-
mate branch lengths in units of mean expected substi-
tutions per site. Furthermore, it quantifies the recon-
struction uncertainty by computing novel quartet-based
branch support scores on the species tree. Because we imple-
mented all of these new methods in our GeneRax software,
users can now perform GFT inference, species tree inference,
GFT correction, and GFT reconciliation with the species tree
using a single tool. We show that SpeciesRax is fast and at
least as accurate as the best competing species tree inference
tools. In particular, SpeciesRax is twice as accurate (in terms
of relative RF distance to the true species trees) than all other
tested methods on simulations with large numbers of paral-
ologous genes.

Results

Accuracy on Simphy Simulations

We plot the accuracy of the different species tree reconstruc-
tion methods under varying parameters for the D TLSIM and
DLSIM experiments in figures 1 and 2, respectively. We ex-
cluded DupTree and NJst from the D TLSIM plots and NJst
from the DLSIM plots for the sake of an improved visual
representation of the results because of their very high error
rate. In addition, we represent the average relative RF distance
of all tested methods over all simulated data sets in table 1.
For species trees with 25 taxa, one incorrect split corresponds
to a relative RF distance of 0.045.

We also assessed the accuracy of the reconstructed GFTs
under different simulation parameter configurations. Under
the default parameters, the average relative RF distance be-
tween the true GFTs and the GFTs inferred with RaxML-NG
is 0.43 for the D TLSIM and DLSIM experiments. When vary-
ing the number of sites, the average relative RF distance
ranges between 0.26 for 300 sites and up to 0.54 for 50 sites
(see fig. 3a). When varying the GFT branch length scaler, the
average RF distance varies from 0.42 without scaling to 0.91 for the smallest scaling value of 0.01 and 0.94 for the largest scaling value of 100,000 (see fig. 3). The remaining simulation parameters do not have a substantial impact on the GFT inference accuracy.

Table 1 shows that SpeciesRax is on average the most accurate method, followed by MiniNJ, ASTRAL-Pro, and FastMulRFS. DupTree and NJst are substantially less accurate than SpeciesRax (by a factor of 4 up to 5 in presence of HGT) and all aforementioned methods. As expected, all methods perform better when the degree of informative data (number of sites, number of families) increases and perform worse when the discordance between the GFTs and the species tree (ILS level, DTL rates) increases. We do not observe a clear correlation between the number of species and the reconstruction accuracy.

Compared with the other methods, SpeciesRax reconstruction accuracy seems to be less affected by increasing DTL rates and almost unaffected by increasing DL rates. We hypothesize that high DL rates are more likely to introduce adversarial (or hidden) paralogy. In other words, paralogous genes are likely to be labeled as orthologous genes by approaches such as ASTRAL-Pro or FastMulRFS that only infer paralogy from the GFTs (see also Molloy and Warnow 2020). For instance, FastMulRFS has only been proven to be consistent when no adversarial paralogy occurs (Molloy and Warnow 2020). In addition, ASTRAL-Pro and FastMulRFS do not attempt to handle HGT events. SpeciesRax explicitly models both, DL, and HGT events, which might explain why it is more robust to both, adversarial paralogy, and HGT.

Large ancestral population sizes are associated with higher rates of ILS. As SpeciesRax does not explicitly model ILS, it might be expected to perform poorly as the rate of ILS increases. In our simulations, all methods performed worse as the ancestral population size was increased, but the reduction in accuracy was similar for SpeciesRax and for methods that account for ILS.

For instance, in the DLSIM experiment, for a moderate population size ($10^7$ individuals), SpeciesRax, MiniNJ, and ASTRAL-Pro exhibit analogous accuracy ($r_{RF} = 0.057$, $r_{RF} = 0.052$, and $r_{RF} = 0.061$, respectively). Under the largest simulated population size ($10^9$ individuals), SpeciesRax ($r_{RF} = 0.183$) outperforms MiniNJ ($r_{RF} = 0.219$), ASTRAL-Pro ($r_{RF} = 0.264$), DupTree ($r_{RF} = 0.266$), and FastMulRFS ($r_{RF} = 0.379$).

Finally, increasing gene conversion (GC) rates does not seem to substantially affect the accuracy of the competing methods.

**Accuracy on Empirical Data Sets**

Here, we describe the results of species tree inference on empirical data sets with ASTRAL-Pro, DupTree, FastMulRFS, MiniNJ, and SpeciesRax. We excluded NJst from this analysis because it performed poorly on most empirical data sets. Initially, we only compare unrooted topologies and defer the root placement analysis to a separate subsection. We provide a graphical representation of the species trees inferred with SpeciesRax as well as the relative pairwise RF distances between all pairs of inferred trees in Supplementary Material online.

**Vertebrates188 Data Set**

All tested methods inferred a different species tree. We first counted the number of splits that differ between the inferred trees and the multifurcating NCBI taxonomy (Federhen 2012) tree. The SpeciesRax, ASTRAL-Pro, and FastMulRFS tools
disagree on five splits, MiniNJ on six splits, and DupTree disagrees on 20 splits.

Then, we focused on the five splits on which SpeciesRax disagrees with the NCBI taxonomy tree that we downloaded from the Ensembl Compara database. Among those discordant splits, the SpeciesRax tree seems to clearly violate only one well established phylogenetic relationship (Venkatesh et al. 2014): The elephant shark, which is a Chondrichtian, is placed by SpeciesRax within Ostecichthians, as sister to Sarcopterygii. Note that all tested methods (ASTRAL-Pro, FastMulRFS, DupTree, and MiniNJ) agree with SpeciesRax.

In the following we analyze the remaining four disagreements. First, SpeciesRax (as well as all other competing tools) places Cichliformes as sister to Ambassidae whereas the
taxonomy places Pomacentridae as sister to Ambassidae. Most studies we have found support the taxonomy (Betancur-R et al. 2013; Near et al. 2013; Hughes et al. 2018) but other studies are undecided about the resolution of these clades and present trees inferred using different inference methods that support the three alternative resolutions (Eytan et al. 2015).

Another discordance with the taxonomy occurs within the avian subtree, between the Estrildidae, Fringillidae, and Passerellidae clades: the taxonomy groups the Estrildidae and Fringillidae together, whereas SpeciesRax, ASTRAL-Pro, and FastMulRFS group Fringillidae and Passerellidae together. A recently published 363 taxon bird phylogeny (Feng et al. 2020) agrees with SpeciesRax on this split and perfectly matches the remaining 24 taxon avian subtree we inferred.

In addition, all tested tools place Bos mutus (yak) and Bison bison closer to each other than to Bos taurus, whereas the taxonomy places B. mutus next to B. taurus. To our knowledge, the literature agrees with our resolution (Decker et al. 2009; Kumar et al. 2018).

The last inconsistency between the taxonomy and the SpeciesRax tree occurs among the Platyrhini (monkey sub-order) when placing Aotidae, Cebus/Saimiri, and Callitrichidae. This split is perhaps more interesting because SpeciesRax disagrees with the competing methods: the taxonomy places Cebus/Saimiri and Callitrichidae together, whereas SpeciesRax places Aotidate and Callitrichidae together. The ASTRAL-Pro, FastMulRFS, MiniNJ, and DupTree tools all group Aotidate with Cebus/Saimiri. There exist studies that agree with the SpeciesRax (Perelman et al. 2011; Springer et al. 2012) but also the ASTRAL-Pro (Fabre et al. 2009) resolutions of these clades.

**Plants83 Data Set**

Both SpeciesRax and ASTRAL-Pro species trees disagree with the literature by placing the Malvales as sister to Malpighiales (instead of sister to Brassicales [Albert et al. 2013; Garcia-Mas et al. 2012]). The SpeciesRax species-driven quartet support scores positively support our resolution, suggesting a potentially misleading signal from the GFTs. When investigating the GFTs, we observed that the Brassicales genes often diverge much earlier than they should and that they are often placed outside of the Rosids clade to which they should belong. A hypothesis for this misleading signal is the apparent overestimation of the gene family sizes during the gene family clustering performed in the original study (Garcia-Mas et al. 2012) as many gene families contain 150 genes (the maximum family size cutoff used in the respective gene family clustering procedure). In addition, the GFTs exhibit clear clusters of genes covering all species separated by extremely long branches. We note, however, that DupTree and FastMulRFS correctly inferred the entire species tree.

**Plants83 Data Set**

The unrooted topologies of the SpeciesRax (fig. 4), ASTRAL-Pro, and FastMulRFS trees are in very good agreement with current biological opinion on the Viridiplantae phylogeny, recovering Setaphyta, and the monophyly of bryophytes (Puttick et al. 2018; Leebens-Mack et al. 2019; Harris et al. 2020). The SpeciesRax tree further agrees with several recent analyses (Leebens-Mack et al. 2019; Harris et al. 2020) in placing the Coleochoetales algae as the closest relatives of Zygnematophyceae and Embryophyta (land plants). The tree inferred by DupTree features a number of disagreements with conventional views of plant relationships.

**Fungi60 Data Set**

All tools found a species tree that disagrees with the literature: They placed the clade formed by Chytridiomycota and Zygomycota between Basidiomycota and Ascomycota, which are typically grouped together (Lutzoni et al. 2004; Marcec-Houben and Gabaldón 2009). The positive extended quartipartition internode certainty (EQPIC) score computed with SpeciesRax along the relevant path shows that the quartets of the GFTs do support this incorrect split. We conclude that the GFTs contain a misleading signal around this split. One possible explanation is that Encephalitozoon cuniculi is evolutionarily very distant from the remaining species, potentially causing a long branch attraction effect (Bergsten 2005). Apart from this split, SpeciesRax, ASTRAL-Pro, and FastMulRFS inferred the same tree, which agrees with the original species tree obtained via concatenation (Marcet-Houben and Gabaldón 2009). The tree inferred with DupTree differs from the SpeciesRax tree in one split.

**Primates13, Cyanobacteria36, Vertebrates22, and Fungi16 Data Sets**

All tools inferred the same species trees for the Primates13, Cyanobacteria36, and Fungi16 (fig. 5) data sets and do not violate any well-established phylogenetic relationship. On the Vertebrates22 data set, all tested methods inferred trees that agree with the multifurcating NCBI taxonomy, but the inferred bifurcating trees are nonetheless different: ASTRAL-Pro, MiniNJ, and SpeciesRax inferred the same tree, which differs from the FastMulRFS tree by one split and theDupTree tree by two splits.

**Archaee364**

The original authors (Dombrowski et al. 2020) suggested that one reason for the difficulty in resolving the archaeal tree was the presence of host–symbiont gene transfers in broadly conserved marker genes, in which members of the host-associated DPANN Archaea sometimes grouped with their hosts in single gene phylogenies. Using the full set of

| Method     | TLSIM | DLSIM |
|------------|-------|-------|
| SpeciesRax | 0.110 | 0.092 |
| MiniNJ     | 0.132 | 0.099 |
| ASTRAL-Pro | 0.172 | 0.121 |
| FastMulRFS | 0.202 | 0.133 |
| DupTree    | 0.518 | 0.212 |
| Njst       | 0.455 | 0.346 |
marker genes, the SpeciesRax tree recovered a clan (Wilkinson et al. 2007) of DPANN; that is, all DPANN Archaea clustered together on the tree. The unrooted SpeciesRax topology is congruent with several recent analyses of the archaeal tree (Raymann et al. 2015; Williams et al. 2017; Dombrowski et al. 2020).

**Fig. 3.** Average relative average distance between true and inferred GFTs on simulated data sets (SIMDL experiment). (a) Number of sites and (b) GFT branch length scaler.
corresponding branch. The values range between $-1$ and 1, with positive values indicating positive support for a branch from the GFTs. For instance, the branch that separates the seed plants from the other plants has a support of 0.691 (excellent support) and the branch that separates the bryophytes from the other plants has a support of 0.005 (low but positive support). Branches with a negative support are represented in red.

In this subsection, we compare the root placement accuracy of SpeciesRax and DupTree with respect to the root split score we introduced in the Experiments section. A lower root split score corresponds to a higher accuracy. On average, over all DLSIM experiments, SpeciesRax and DupTree with respect to the root split score plots in Supplementary Material online.

**Rooting Accuracy on the Empirical Data Sets**

In the following, we conduct an in-depth assessment of the accuracy of the species tree root inference with SpeciesRax on the tested empirical data sets.

We first discuss the data sets on which SpeciesRax inferred a species tree root that agrees with the current literature. On the priates13 data set, SpeciesRax correctly places the species tree root between the Strepsirrhini and Haplorhini clades (Chatterjee et al. 2009). On the Fungi16 data set, the root inferred with SpeciesRax correctly separates the Candida and Saccharomyces clades (Butler et al. 2009). The root we inferred on the Vertebrates22 species tree correctly separates the Actinopterygii and Sarcopterygii clades (Meyer and Zardoya 2003). On the Plants23 data set, our species tree root correctly separates the Chlorophyta and Streptophyta clades (Leliaert et al. 2012). On the Fungi60 data set, we correctly find that E. cuniculi (Microsporidia clade) diverged earlier than the other clades contained in the data set (Nagy and Szöllösi 2017). On the Vertebrates188 data set, SpeciesRax infers a root that groups lampreys and hagfishes, on one side, and cartilaginous fishes, bony fishes, and tetrapods on the other side. The position of the vertebrate root is still controversial (Takezaki et al. 2003; Miyashita et al. 2019) and our resolution complies with some of the plausible
scenarios discussed in the literature (Meyer and Zardoya 2003; Takezaki et al. 2003; Miyashita et al. 2019).

On the Plants83 data set, SpeciesRax agrees with the literature in placing Embryophyta (land plants) within the Streptophyte algae. However, the inferred root is three branches away from the consensus position, in the common ancestor of the Chlorophyta (Volvox, Chlamydomonas, and Uronema).

On the Cyanobacteria36 data set, the root placement inferred by SpeciesRax is one branch away from one of the three plausible roots inferred in a recent study (Szöllősi et al. 2012).

The Archaea364 data set only contained single-copy gene families, and thus no gene duplications. As a result, the position of the root was uncertain. However, the 95% confidence set of possible root placements obtained via the approximately unbiased (AU) test (Shimodaira 2002) was compatible with several recent suggestions in the literature, including a root between DPANN and all other Archaea (Williams et al. 2017; Dombrowski et al. 2020) and a root within the Euryarchaeota (Raymann et al. 2015), among a range of other positions within and between the major archaeal lineages.

The root inferred by SpeciesRax on the Life92 data set is biologically implausible as it is located between Viridiplantae...
and all other taxa. One possibility is that root inference for these data is affected by large differences in gene content among the included taxa. For example, the Viridiplantae (and other Archaeplastida) have chloroplasts, and so possess an additional source of bacterial-origin genes compared with other Eukaryotes and Archaea. To evaluate the impact of major gene content differences, we performed another speciesRax analysis in which the gene families covering less than half of the species were removed. In this second analysis, the root was inferred to lie between the Eukaryotes and Archaea. This root position is compatible with a three-domain tree of life hypothesis. However, this should be interpreted with caution, because the branch separating Eukaryotes and Archaea is one along which major gene content changes occurred, including (but not limited to) the acquisition of a bacterial genome’s worth of genes in the form of the mitochondrial endosymbiont (Roger et al. 2017).

Runtime

Before comparing runtimes, we emphasize again that we executed the experiments on a 40-core machine and that only speciesRax and ASTRAL-Pro provide a parallel implementation. Although this choice might appear to favor speciesRax and ASTRAL-Pro, we argue that the absence of parallelization constitutes a substantial limitation of the remaining tools as completing an analysis in less than 1 day on a parallel system instead of having to wait for several weeks represents a strong advantage.

We also emphasize that speciesRax is the only tested tool that can be executed across several compute nodes with distributed memory in contrast to ASTRAL-Pro that can only run on a single shared memory node. All tools, with the exception of MiniNJ, required huge amounts of memory for the largest data set (>200 GB on vertebrates188) and can therefore not be executed on most common servers. The speciesRax MPI implementation allows to distribute the memory footprint over different compute nodes, which is not feasible with the other tools.

We show the runtimes for an increasing number of species and an increasing number of families for the simulated data sets in figure 7. Our MiniNJ method requires <0.1 s for all parameter combinations and is the fastest method we tested. The runtimes of DupTree and FastMulRFS grow faster with increasing number of gene families, and DupTree runtime quickly raises with the number of species. The speciesRax and ASTRAL-Pro runtimes are less affected by these parameters.

On almost all empirical data sets, MiniNJ and speciesRax are the fastest methods. On the two largest data sets (Plants83 and Vertebrates188), MiniNJ is at least one order of magnitude faster than speciesRax and speciesRax is at least one order of magnitude faster than all other methods. In particular, speciesRax only requires 1 h on 40 cores to infer the 188 vertebrate species tree with 188 species and 31,612 gene families.

Discussion

A Fast and Accurate Approach

We introduced two new methods for species tree inference from GFTs in the presence of paralogy. Our MiniNJ tool is a distance-based method that is faster than all tested methods while being at least as accurate as all other nonparametric methods for the majority of our simulated data experiments. In particular, MiniNJ inferred a species tree with 188 species in <1 min from more than 30,000 gene families. SpeciesRax, a novel maximum likelihood tree search method that explicitly accounts for gene duplication, gene loss, and HGT events. Our speciesRax tool infers rooted species trees with branch lengths in units of mean expected substitutions per site. Further, to assess the confidence of the inferred species tree, we introduce several quartet-based support measures.

In terms of accuracy, speciesRax is more accurate than its competitors on simulated data sets, and up to twice as accurate under high duplication, loss, and HGT rates. On empirical data sets, speciesRax is on par with or more accurate than its competitors. In addition, among the tested tools, speciesRax, PHYLDOG, and DupTree are the only methods that can infer rooted species trees. On simulated data sets, speciesRax outperforms both PHYLDOG and DupTree in terms of root placement accuracy. SpeciesRax inferred the correct (biologically well-established) roots on six out of ten empirical species trees, and found roots that are close to the plausible roots in three out of the remaining four data sets (Plants83, Archaea364, and Cyanobacteria). For the most challenging-to-root data set (Life92), we managed to infer a plausible root by removing those gene families that only covered less than half of the species.

Despite being a compute-intensive maximum likelihood based tree search method, speciesRax is faster than all tested methods (except MiniNJ) on large empirical data sets. This is due to our fast method MiniNJ for inferring a reasonable starting tree and to our efficient reconciliation-aware search strategy. In addition, speciesRax provides a parallel implementation and can be run on distributed memory cluster systems. Thereby it facilitates conducting large-scale analyses (table 2).

Further, speciesRax has been integrated into our GeneRax tool that is available via Github and BioConda (Grüning et al. 2018). With GeneRax, users can execute the following (optional) steps in one single run: GFT inference from the gene alignments, rooted species tree inference with speciesRax
from the GFTs, species-tree aware GFT correction, and GFT reconciliation with the rooted species tree. Alternatively, SpeciesRax can be used to infer the root of a user-specified species tree (typically obtained from concatenation methods) before inferring reconciliations. Thus, GeneRax has become a versatile, one-stop shop for executing likelihood-based analyses on multiple-copy gene families.

Future Work

Despite our encouraging results, SpeciesRax still faces some challenges.

First, SpeciesRax cannot currently take into account GFT reconstruction error/uncertainty. This issue will become more prevalent with increasing taxon numbers and the associated increase in reconstruction uncertainty. Therefore, we intend to explore several ideas to overcome this limitation. A first idea consists in contracting the low-support branches of the GFTs and adapting our reconciliation model to multifurcating GFTs. Alternatively, we will explore if coestimating the species tree and the GFTs is feasible, as conducted by Phyldog (Boussau et al. 2012), for instance. Finally, we could take as input a distribution of GFTs for each gene family (instead of just one maximum likelihood GFTs) and integrate over this distribution of per gene family GFTs to compute the likelihood score. Such a GFTs distribution could obtained from Bayesian inference tools (Ronquist et al. 2012), from bootstrap trees (Felsenstein 1985), or from a set of plausible GFTs (Morel et al. 2021).

Secondly, we plan to implement more complex models of GFT evolution in SpeciesRax. First, the UndatedDTL model could be improved by modeling variation in DTL intensities among different branches of the species tree. This could, for instance, result in a better placement for microsporidia in the fungi60 tree, because this species is known to have a higher rate of losses than the other fungi. Furthermore, the UndatedDTL model implemented in SpeciesRax allows for HGTs between any pair of species, even if such HGTs are impossible timewise. Some models (Szöllösi et al. 2012) extract time information from the species tree to prohibit non-contemporary HGTs, that is, HGTs between species that have not coexisted and thus could not have exchanged genes. Finally, some promising work (Chan et al. 2017; Rasmussen and Kellis 2012; Li et al. 2020) has been conducted to account for both DTL events and ILS in a single model. We hope that models that better reflect the complexity of gene family evolution will yield more reliable species tree inference.

Finally, there now exists a range of quite distinct methods for species tree inference, and choosing the best method for analyzing a given data set is difficult. The available methods differ in the extent to which they consider ILS, HGT, phylogenetic uncertainty, and the underlying complexities of molecular sequence evolution. Some methods are also more scaleable.
than others. Because no method provides a full treatment of all of the complexities of genome evolution, prior biological opinion about the extent to which ILS, HGT, GFT error, or other signals might contribute in specific cases may help to guide the decision about which method to use. In addition to being able to infer species trees from multiple-copy gene families, one potential advantage of SpeciesRax is that it might be less sensitive to errors in orthology assignment, or—as observed in the archaeal data set—gene transfers of marker genes, because the DTL model can account for duplication-loss or transfer events. Benchmarking of the available tools on both empirical and simulated data sets will provide further insight into their relative performance and will help to develop practical guidelines for users interested in species relationships and genome evolution across the tree of life.

Materials and Methods

We recently (Morel et al. 2020) introduced the undatedDTL model, which is a phenomenological discrete time model that describes the evolution of a GFT along a species tree through gene duplication, gene loss, speciation, and HGT events. In addition, we described an algorithm for computing the corresponding reconciliation likelihood, that is, the probability of observing a set of GFTs \( G = (G_1, \ldots, G_n) \) given a rooted species tree \( S \) and the set \( \Theta \) of duplication, loss, and HGT intensities describing the expected frequency of those events relative to the frequency of the speciation events:

\[
L(S, \Theta|G) = \prod_{k=1}^n P(G_k|S, \Theta).
\]

As already mentioned, SpeciesRax takes a set of unrooted GFTs as input. It starts its computations from an initial species tree \( S \) and the model parameters \( N \) that maximize the reconciliation likelihood \( L(S, \Theta|G) \). At the end of the search, it also calculates support values for the inner branches of the inferred species tree from the GFTs. Finally, we also describe the adaptation of our likelihood score to better account for missing data and inaccurate assignment of sequences to gene family clusters.

Computing a Reasonable Initial Species Tree with MiniNJ

Here, we introduce MiniNJ (Minimum internode distance neighbor joining [NJ]), our novel distance-based method for inferring an unrooted species tree in the presence of paralogy. MiniNJ is fast, that is, it is well-suited for generating an initial species tree for the subsequent maximum likelihood optimization. MiniNJ is inspired by NJst (Liu and Yu 2011), a distance-based method that performs well in the absence of paralogy. Initially, we briefly outline the NJst algorithm, and subsequently describe our modifications.

NJst initially computes a distance matrix from the unrooted GFTs and then applies NJ to reconstruct the species tree. NJst defines the gene internode distance \( D_g(x, y) \) is the number of internal nodes between the terminal nodes \( x \) and \( y \) in a GFT. NJst computes the distance between two species as the average over the internode distances between all pairs of gene copies mapped to those two species.

More formally, let \( a \) and \( b \) be two species. Let \( K \) be the number of GFTs. Let \( M_k(a) \) be the number of terminal nodes from the GFT \( k \) mapped to species \( a \). Let \( x_{ik}(a) \) be the \( i \)th terminal node from the GFT \( k \) mapped to species \( a \). NJst computes each element of the distance matrix \( D_{NJst} \) by first summing over all gene distances from all gene families, and then by dividing by the number of pairwise gene distances involved in this sum:

\[
D_{NJst}(a, b) = \frac{\sum_{k=1}^K \sum_{i=1}^{M_k(a)} \sum_{j=1}^{M_k(b)} D_g(x_{ik}(a), x_{jk}(b))}{\sum_{k=1}^K M_k(a) M_k(b)}.
\]  

(2)

NJst has two drawbacks. First, it accounts for all pairs of gene copies, including paralogous gene copies that do not contain information about speciation events (see fig. 8). Secondly, it assigns very high (quadratic) weights to gene families comprising a high number of gene copies: for instance, a gene family \( k \) with five gene copies in both species \( a \) and \( b \) will contribute 25 times to the distance between \( a \) and \( b \), whereas a single-copy family \( k_2 \) will only contribute once. For instance, \( \sum_{i=1}^1 \sum_{j=1}^1 D_g(x_{ik}(a), x_{jk}(b)) \) is the sum over 25 gene internode distances for family \( k_1 \), and of only one gene internode distance for family \( k_2 \). Since the normalization by the number of gene internode distances is conducted after summing over all these quantities (with the denominator in eq. 2), the contributions of families \( k_1 \) and \( k_2 \) are unbalanced.

MiniNJ adapts equation (2) to address these two issues. It assumes, that on average, two orthologous genes from two given, distinct species are closer to each other than two paralogous genes from the same two species (see for instance fig. 8). Note that this assumption does not always hold, for instance in the presence of HGT. MiniNJ attempts to discard pairs of paralogous gene copies by only considering the two closest GFT terminal nodes mapped to a pair of species for each family, according to the internode distance. Let \( \delta_{abk} \) be equal to 1 if gene family \( k \) contains at least one gene copy mapped to \( a \) and one gene copy mapped to \( b \), and 0 otherwise. We define \( D_{MiniNJ} \):

\[
D_{MiniNJ}(a, b) = \frac{\sum_{k=1}^K \min_i \min_j \min_{i'} \min_{j'} D_g(x_{ik}(a), x_{jk}(b))}{\sum_{k=1}^K \delta_{abk}}.
\]  

(3)

Note that for any two species \( a \) and \( b \), all gene families that cover \( a \) and \( b \) contribute equally to \( D_{MiniNJ}(a, b) \).

MiniNJ then infers an unrooted species tree from this distance matrix using the NJ algorithm (Saitou and Nei 1987).

The distance matrix computation has time complexity \( O(\sum_{k=1}^K |g_k|^2) \) where \( |g_k| \) is the number of gene sequences in the
The NJ algorithm has time complexity $O(|S|^3)$ where $|S|$ is the number of species. The overall time complexity of MiniNJ is thus $O(|S|^3 + \sum_{k=1}^{K} |g_k|^2)$.

### Maximum Likelihood Rooted Species Tree Search

Given a set $\mathcal{G}$ of unrooted GFTs, SpeciesRax implements a hill-climbing algorithm to search for the rooted species tree $S$ and optimize the set of model parameters $\Theta$ (duplication, loss, and HGT intensities) that maximize the reconciliation likelihood $L(S, \Theta|\mathcal{G})$. In this section, we only briefly explain our search algorithm. We provide a more detailed description of the different steps and the order by which we apply them in Supplementary Material online.

The search starts from an initial species tree $S$ and a default or user-specified set of initial model parameters $\Theta_0$. Then, we alternate between optimizing the species tree root position, $\Theta$, and the species tree topology until we cannot find a configuration with a better likelihood. We optimize the root position by evaluating the likelihood of the neighbors of the current candidate root position and repeat this process until we do not encounter a neighboring candidate root position with a higher likelihood. We optimize $\Theta$ via a gradient descent approach. To optimize $S$, we alternate between two complementary tree search strategies that both rely on subtree prune and regraft (SPR) moves: The transfer-guided SPR search assumes that the SPR moves that reduce the number of required HGT events necessary to reconcile the GFT with the species tree are more likely to improve the reconciliation likelihood. In Supplementary Material online, we detail how we identify and apply such SPR moves. The local SPR search tries all possible SPR moves within a user-specified radius (one by default). In both search strategies, when SpeciesRax finds a species tree $S'$ with a better likelihood than $S$, it replaces $S$ by $S'$.

When applying the final root position search, SpeciesRax outputs the per-GFT likelihood scores for all tested root positions. The file with these per-GFT likelihoods can then be further analyzed with the Consel tool (Shimodaira and Hasegawa 2001) to perform a range of statistical significance tests (e.g., the AU test [Shimodaira 2002]) to generate a confident set of root placements.

Calculating the reconciliation likelihood under the UndatedDTL model represents the major computational bottleneck. To reduce its computational cost, we introduce several approximations. The first approximation is to ignore certain combinations of events that are unlikely to happen (e.g., a gene being transferred twice in a row), and which therefore exhibit a negligible contribution to the overall likelihood score. The second approximation is to dynamically estimate the most likely root position for each GFT during the species tree search, thereby avoiding the need to evaluate all possible GFT root positions. These approximations are described in more detail in Supplementary Material online.

### Support Value Assessment

Here, we describe how SpeciesRax calculates branch support values on the species tree from a set of unrooted GFTs $\mathcal{G}$. We first revisit the definition of a speciation-driven quartet (SQ). Then, we explain how we use the SQ frequency to assess branch support values. Finally, we describe two alternative SQ-based scores, namely the quadripartition internode certainty (QPIC) and the EQPIC scores.

We first briefly revisit the definition of an SQ (Zhang et al. 2020). Let $\mathcal{G}$ be a set of rooted GFTs with internal nodes either tagged by “duplication” or “speciation” events as estimated from $\mathcal{G}$. A quartet from $\mathcal{G} \in \mathcal{G}$ only contains information
about the speciation events, if it includes four distinct species and if the lowest common ancestor of any three out of the four taxa of this quartet is a speciation node. Such a quartet is called SQ. We refer to Zhang et al. (2020) for a more formal definition of the SQ count and for its computation from a set of unrooted and unlabeled GFTs.

We now introduce several notations in order to define the SQ frequency of a pair of internal nodes in the species tree. Let $S$ be an unrooted species tree. Let $(u, v)$ be a pair of distinct internal nodes in $S$. The nodes $u$ and $v$ define a metaquartet $M_{u,v} = (A, B, C, D)$, where $A$ and $B$ (respectively, $C$ and $D$) are the leaf sets under the left and right children of $u$ (respectively, $v$) with $S$ rooted at $v$ (respectively, $u$). Let $z = (z_1, z_2, z_3)$ such that $z_1$ (respectively, $z_2$ and $z_3$) is the SQ count in $S$ corresponding to the metaquartet topology $AB|CD$ (respectively, $AC|BD$ and $AD|BC$). Note that $z_1$ corresponds to the metaquartet topology that agrees with $S$ ((A, B|C, D)) and that $z_2$ and $z_3$ correspond to the two possible alternative distinct metaquartet topologies ($AC|BD$ and $AD|BC$). Let $z = (z_1, z_2, z_3)$ such that $z_i = \frac{z_i}{z_1 + z_2 + z_3}$ for $i \in \{1, 2, 3\}$. We define the SQ frequency of $(u, v)$ in $S$ given $G'$ as $\text{SQF}_{G'}(u, v) = \tilde{z}_1$.

The SQ frequency represents how many SQs around $u$ and $v$ support the species tree topology. However, it does not always reflect if $(AB|CD)$ is the best supported of the three possible metaquartet topologies, in particular when $(1/3) < z_1 < (1/2)$. For instance, $\tilde{z} = (0.4, 0.3, 0.3)$ suggests that $(AB|CD)$ is the correct topology, but $\tilde{z} = (0.4, 0.6, 0.0)$ suggests that the alternative topology $(AC|BD)$ is better supported. Thus, the value of $z_1$ alone is not sufficiently informative to assess our confidence in a branch defined by $u$ and $v$.

To overcome this limitation, we therefore also compute the QPIC and EQPIC scores introduced in Zhou et al. (2020). Note that these scores were initially defined for single-copy gene families. Because SpeciesRax operates on multiple-copy families, we adapt the scores by only counting SQs instead of counting all quartets. Let $(u, v)$ be two distinct nodes of $S$.

$$\text{qpic}'(u, v) = 1 + \tilde{z}_1 \log(\tilde{z}_1) + \tilde{z}_2 \log(\tilde{z}_2) + \tilde{z}_3 \log(\tilde{z}_3)$$

$$\text{QPIC}(u, v) = \begin{cases} 0 & \text{if } z_1 = z_2 = z_3 = 0 \\ \text{qpic}'(u, v) & \text{if } z_1 = \max(z_1, z_2, z_3) \\ -\text{qpic}'(u, v) & \text{otherwise} \end{cases} \quad (4)$$

In particular, if $u$ and $v$ are neighbors, we define the QPIC of the branch $e$ between $u$ and $v$ as $\text{QPIC}(e) = \text{QPIC}(u, v)$. One limitation of the QPIC score is that it discards all SQs defined by nodes $u$ and $v$ that are not neighbors. Zhou et al. (2020) extends the QPIC score by defining the EQPIC score of a branch $e$:

$$\text{EQPIC}(e) = \min_{(u, v) \in N(e)}(\text{QPIC}(u, v))$$

where $N(e)$ is the set of node pairs $(u, v)$ such that the branch $e$ belongs to the unique path between $u$ and $v$.

Note that both QPIC and EQPIC scores range between –1 and 1. They take positive values when they support the relevant metaquartet topologies of the species tree $S$ and negative values otherwise.

We also note that support values are not calculated under the probabilistic UndatedDTL model that SpeciesRax employs to infer the species tree. Instead, they provide a measure of the topological conflict between the GFTs and each branch of the species tree based on a nonprobabilistic heuristic.

### Accounting for Missing Data

We refer to missing data as gene copies that are absent from a gene family to which they should belong. This can occur, for instance, when some gene sequences have not been sampled or when the gene family clustering is inaccurate. Missing data is problematic for species tree estimation, in particular, when the missing data pattern distribution is nonrandom (Xi et al. 2016). In particular, reconciliation methods like SpeciesRax can be affected by missing gene copies: for instance, if sequences for a subset of the species under study have not been sampled for several families, the statistical reconciliation model will attempt to explain these missing gene copies via additional, yet incorrect loss events. Thus, a candidate species tree that groups such a subset of species into one subtree will typically exhibit a better reconciliation likelihood score than the “true” species tree. This is the case, because only one loss event per family would be necessary to explain all missing gene copies. We alleviate this problem to a certain extent by deploying a species tree pruning mode: Let $G$ be a GFT and $S$ a species tree. We replace the reconciliation likelihood term $L(S, G)$ by $L(S', G)$, where $S'$ is obtained from $S$ by pruning all species that are not covered by $G$ and by removing internal nodes of degree 1 until the tree is bifurcating. Thus, if a species is not present in a family, the reconciliation likelihood of this family does not depend on the position of this species in the species tree.

A downside of this approach is that it can disregard some true gene loss events. In addition, it is currently unclear how the pruned mode affects the properties of the reconciliation likelihood model. The experiments presented here suggest that the pruned mode neither decreases reconstruction accuracy nor increases runtimes. A mathematically more appealing solution would directly incorporate missing data in the UndatedDTL model.

### Parallelization

We parallelized SpeciesRax with MPI (Message Passing Interface) which allows to execute it using several compute nodes with distributed memory (e.g., compute clusters). We distribute the gene families among the available cores to parallelize the reconciliation likelihood computation.

### Experiments

#### Tested Tools

In the following, we describe the settings we used for executing all tools (summarized in table 3) in our experiments. We ran DupTree, FastMULRFS, and MiniNJ with default
parameters. Among the four outputs that FastMulRFS provides, we discarded the outputs that may contain multifurcating trees (“majority” and “strict”). Among the two remaining outputs (“greedy” and “single”), we selected “single” because it performed slightly better in our experiments.

We used our own (re-)implementation of Njst (available in GeneRax) because the existing implementation written in R was too slow for completing our tests in a reasonable time.

We executed ASTRAL-Pro using all available memory (“-Xms700G -Xmx700G”) and a fixed random number seed (“-seed 692”).

We executed PHYLDOG using all available cores on the server (“mpiexec -np 40”) and conducted a nearest-neighbor interchange-based (NNI) tree search for the GFT optimization (“rearangement.gene.tree = nni”). We used MiniNJ to generate the starting species tree.

We executed SpeciesRax starting from a MiniNJ tree, with the UndatedDTL model, with per-family duplication, HGT, and loss (DTL) rates. We also disabled all irrelevant steps such as gene tree optimization (“-s MiniNJ -optimize-species-tree -do-not-optimize-gene-trees -rec-model UndatedDTL -per-family-rates -skip-family-filtering -do-not-reconcile”). For the experiments on empirical data sets, we added the SpeciesRax option “-prune-species-tree” described in Account for Missing Data. To analyze the empirical data set that do not contain any multiple-copy gene families (Archaea364), we disabled the gene duplication events in the UndatedDTL model (option “-no-dup”).

**Hardware Environment**

We executed all experiments on the same machine with 40 physical cores, 80 virtual cores, and 750GB RAM. Note that DupTree, Njst, and MiniNJ only offer a sequential implementation. In contrast, SpeciesRax, FastMulRFS, ASTRAL-Pro, and PHYLDOG provide a parallel implementation (although FastMulRFS has at least one step that seems to be run sequentially) and were run using all available cores. We discuss the implications of this choice in Results.

**Simulated Data Sets**

We generated simulated data sets with SimPhy (Mallo et al. 2016) to assess the influence of the simulation parameters on the reconstruction accuracy of the methods.

The parameters we studied are the average number of sites per gene family MSA, the number of families, the size of the species tree, the average DTL rates, the GC rate, the GFT branch length scaler, and the population size. For each parameter we studied, we varied its value while keeping all other parameters fixed. In addition, we varied the DL rates while keeping the T rate fixed, and also varied the T rate while keeping the DL rates fixed. We generated 100 replicates for each set of parameter values. We executed the entire experiment twice, once including HGTs (DTLSIM experiment) and once excluding HGTs (DLSIM experiment).

We reused the default parameters of the S25 (25 species per species tree) experiment of (Zhang et al. 2020) with some modifications that we list in the following. By default, we do not simulate ILS, because SpeciesRax does not model ILS. Therefore, the species tree inference is easier than in the original S25 experiment. To make the reconstruction more challenging and to reduce the computational cost of the entire experiment, we reduced the number of families from 1,000 to 100. To increase the heterogeneity among gene families, we used a log-normal distribution for the sequence length (with, by default, 100 sites per sequence on average) and for the DTL rates. In the DTLSIM experiment, we simulated under the distance-independent HGT model (i.e., the receiving species is uniformly sampled from all contemporary species) and we set the default average HGT rate equal to the default average duplication rates. We provide a detailed list of the SimPhy parameters in Supplementary Material online.

We inferred the GFTs with ParGenes (Morel et al. 2018), performing one RAxML-NG search on a single random starting tree per gene family under the general time reversible model of nucleotide substitution with four discrete gamma rates (GTR + G4) (Tavaré 1986; Yang 1993). Then, we inferred the species trees from the inferred GFTs with every tool listed in table 3.

Finally, for each data set, we assessed the species tree reconstruction accuracy by computing the average relative RF distance between each inferred species tree and the corresponding true species tree using the ETE Toolkit (Huerta-Cepas et al. 2016). We assessed the root placement accuracy of SpeciesRax, DupTree, and PHYLDOG by introducing the root split score. It is defined as the proportion of inferred rooted species trees whose root incorrectly splits the species into two groups compared with the split induced by the correct, true root. The remaining methods we tested do not infer rooted species trees.

Due to excessive runtimes, we only executed PHYLDOG on the 100 replicates of the default parameter set of the DLSIM experiment. PHYLDOG was more than two orders of magnitude slower than all the other methods and less accurate than SpeciesRax (see Supplementary Material online).
Empirical Data Sets

We used empirical data sets from various sources to cover a wide range of organisms including plants, fungi, vertebrates, bacteria, and archaea. We describe these data sets in Table 4. When the data sets included outgroups, we excluded them from the analysis, because SpeciesRax does not need any outgroup to root the species trees. For data sets where we pruned outgroups and for which alignments were available, we reinferred the GFTs from the alignments. This was done to avoid any potential bias in the tree reconstruction that could be caused by the outgroup (Holland et al. 2003). When the alignments were available, we used GeneRax to correct the GFTs and to estimate the DTL intensities on those empirical data sets, to confirm that they fall in the same range as the DTL intensities estimated from the simulated data sets (see Supplementary Material online). In the following we describe in detail how we assembled each empirical data set.

**Primates13 and Vertebrates188 Data Sets.** We extracted the alignments comprising 199 species from the Ensembl Compara database (Zerbino et al. 2018). We removed five nonvertebrate species to obtain the Vertebrates188 data set. Further, we extracted 13 primate species to obtain the Primates13 data set. For both data sets, we inferred the GFTs with ParGenes under the GTR + G4 model with one random starting tree per RAxML-NG maximum likelihood search.

**Cyanobacteria36 Data Set.** We reused the protein alignments of a previous study (Szöllősi et al. 2013) covering 36 cyanobacteria species to generate the Cyanobacteria36 data set. We inferred the GFTs with ParGenes under the same substitution model used in the original study (LG + G4 + I) with one random starting tree per RAxML-NG search.

**Fungi60, Plants83, and Vertebrates22 Data Sets.** We extracted data sets from three different phylomes of the PhylomeDB (Huerta-Cepas et al. 2014) database: vertebrates (phyllome ID = 200), fungi (phyllome ID = 3), and plants (phyllome ID = 84). We removed the two outgroup species (Arabidopsis thaliana and Human) from the fungi phylome to generate the Fungi60 data set. We removed the five outgroup species (outgroups: human, Drosophila, Caenorhabditis elegans, Saccharomyces cerevisiae, and Plasmodium falciparum) from the plant phylome to generate the plants21 data set. We reinferred the GFTs of both, the fungi and plants data sets using ParGenes with best-fit model selection enabled (-m option) and one random starting tree per RAxML-NG search. We generated the Vertebrates22 data set from the vertebrates phyloome. Here, we did not remove any outgroup and did therefore not reinfer the corresponding GFTs.

**Life92 Data Set.** To compare to the supertrees inferred in the original study (Williams et al. 2020), we extracted the original GFTs covering 92 species from the Eukaryote and Archaea domains. To take advantage of the signal from duplications and transfers, we also inferred new homologous gene families from the genomes used in that study. To do so, we performed all-versus-all Diamond (Buchfink et al. 2015) searches, then clustered gene families using mcl (Enright 2002) with an inflation parameter value of 1.4. As in the original study, sequences were aligned using MAFFT (Katoh and Standley 2013) and poorly aligning positions removed using BLOSUM 1.12 (Cricuolo and Gribaldo 2010) with the BLOSUM30 matrix.

**Archaea364 Data Set.** We downloaded the MSAs of the marker proteins from the original study (Dombrowski et al. 2020). We inferred the GFTs with ParGenes using the LG + G4 substitution model.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

This work was financially supported by the Klaus Tschira Foundation and by DFG grant STA 860/6-2. G.J.S. received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program under grant agreement no. 714774 and the grant GINOP-2.3.2.-15-2016-00057. T.A.W. was supported by a
Royal Society University Fellowship and NERC grant NE/P00251X/1. This work was funded by the Gordon and Betty Moore Foundation through grant GBMF9741 to T.A.W. and G.J.S.

Data Availability

The code is available at https://github.com/BenoitMorel/GenetRax and data are made available at https://cmr.h-its.org/exelixis/material/speciesrax_data.tar.gz.

References

Aberer A, Koberst K, Stamatakis A. 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol Biol Evol. 31(10):2533–2536.

Altenhoff AM, Glover NM, Desimoz C. 2019. Inferring orthology and paralogy. New York: Springer. p. 149–175.

Albert VA, Babazuk WB, dePamphilis CW, Der JP, Leebens-Mack J, Ma H, Palmer JD, Rounsley S, Sankoff D, Schuster SC, et al. Amborella Genome Project. 2013. The Amborella genome and the evolution of flowering plants. Science 342(665):1241089.

Bayzid M, Mirarab S, Warnow T. 2013. Inferring optimal species trees under gene duplication and loss. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 18th Pacific Symposium on Biocomputing, PSB 2013. Conference date: 03-01-2013 through 07-01-2013. p. 250–261.

Bergsten J. 2005. A review of long-branch attraction. Mol Phylogenet Evol. 33(3):566–573.

Butler G, Rasmussen M, Lin M, Sakthikumar S, Munro C, Rheinbay E, Butler GJ, Ansorge W, Andrews ZW, et al. 2012. Genome-scale coestimation of species and gene trees. Genome Res. 23(2):323–330.

Bryant D, Hahn MW. 2020. The concatenation question. In: Scornavacca C, ed. Advances in Phylogenetics. Springer. p. 317–360.

Cavender J, Hsu L, Hu S, et al. 2018. Complete mitogenome reveals genetic divergence and phylogenetic relationships among Indian cattle (Bos indicus) breeds. Anim Biotechnol. 30:219–232.
Morel B, Kozlov AM, Stamatakis A. 2018. ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* 25(17):2286–2288.

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574(7780):679–685.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, Clerck OD. 2012. Phylogeny and molecular evolution of the green algae. *Crit Rev Plant Sci.* 31(1):1–46.

Li Q, Scornavacca C, Galtier N, Chan Y-B. 2020. The multicus multi-species coalescent: a flexible new model of gene family evolution. *Syst Biol.* 70(4):822–837.

Liu L, Yu L. 2011. Estimating species trees from unrooted gene trees. *Syst Biol.* 60(5):661–667.

Lutteropp S, Scornavacca C, Kozlov AM, Morel B, Stamatakis A. 2021. Nextrax: Accurate and fast maximum likelihood phylogenetic network inference. *bioRxiv*.

Lutzeri F, Kauff F, Cox C, McLaughlin D, Celio G, Dentinger B, Padamsee M, Hibbett D, James T, Baloch E et al. 2004. Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. *Ann. J. Bot.* 91(10):1446–1480.

Mallo D, De Oliveira Martins L, Posada D. 2016. SimPhy: phylogenomic simulation of gene, locus, and species trees. *Syst Biol.* 65(2):334–344.

Marcet-Houben M, Gahbaldón T. 2009. The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One* 4(2):e3457.

Mendes FK, Hahn MW. 2018. Why concatenation fails near the anamorph zone. *Syst Biol.* 67(1):158–169.

Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* 23(5):733–736.e2.

Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22(4):755–765.

Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the archaea. *Proc Natl Acad Sci U S A.* 112(21):6670–6675.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4(4):406–425.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51(1):492–508.

Shimodaira H, Hasegawa M. 2001. CONSEL for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.

Saitou N, Crow SW, Anise HN, Horvath JE, Moreira OSI GJ. 2017. Chapter Two - Fungal phylogeny in the age of genomic data sets. In: Townsend J, Wang Z, editors. *Fungal phylogenetics and the molecular evolution of DNA sequences*. Lecture Notes in Life Sciences 34. Berlin, Heidelberg: Springer. p. 49–72.

Sjööblom A, Fusetani N, Kjelleberg S, Welinder L. 2008. The two-domain tree of life: progress, classification, and evolution of subcellular traits. *Mol Biol Evol.* 25(5):1177–1179.

Sonnheim TL, Pine A, Tishkoff SA, Wellman CH, Schneider H, Pisani D et al. 2018. The interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* 28(5):733–745.e2.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*. 10(6):1396–1401.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaita L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541(7637):353–358.

Zerbino DR, Achuthan P, Akanni W, Amode M, Barrell D, Bhai J, Bills K, Cummins C, Gall A, Giron CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res*. 46(D1):D754–D761.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19(S6).

Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. Astral-pro: quartet-based species tree inference despite paralogy. *Mol Biol Evol*. 37(11):3292–3307.

Zhou X, Lutteropp S, Czech L, Stamatakis A, Looz MV, Rokas A. 2020. Quartet-based computations of internode certainty provide robust measures of phylogenetic incongruence. *Syst Biol*. 69(2):308–324.