PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Accurate, scalable, and fully automated inference of species trees from raw genome assemblies using ROADIES

Anshu Gupta[a], Siavash Mirarab[b], and Yatish Turakhia[b,1]

Affiliations are included on p. 9.

Edited by Marcus Feldman, Stanford University, Stanford, CA; received January 9, 2025; accepted March 31, 2025

Current genome sequencing initiatives across a wide range of life forms offer significant potential to enhance our understanding of evolutionary relationships and support transformative biological and medical applications. Species trees play a central role in many of these applications; however, despite the widespread availability of genome assemblies, accurate inference of species trees remains challenging due to the limited automation, substantial domain expertise, and computational resources required by conventional methods. To address this limitation, we present ROADIES, a fully automated pipeline to infer species trees starting from raw genome assemblies. In contrast to the prominent approach, ROADIES incorporates a unique strategy of randomly sampling segments of the input genomes to generate gene trees. This eliminates the need for predefining a set of loci, limiting the analyses to a fixed number of genes, and performing the cumbersome gene annotation and/or whole genome alignment steps. ROADIES also eliminates the need to infer orthology by leveraging existing discordance-aware methods that allow multicopy genes. Using the genomic datasets from large-scale sequencing efforts across four diverse life forms (placental mammals, pomace flies, birds, and budding yeasts), we show that ROADIES infers species trees that are comparable in quality to the state-of-the-art studies but in a fraction of the time and effort, including on challenging datasets with rampant gene tree discordance and complex polyploidy. With its speed, accuracy, and automation, ROADIES has the potential to vastly simplify species tree inference, making it accessible to a broader range of scientists and applications.

phylogenetics | species tree inference | bioinformatics

The rapid progress in genome sequencing technologies and assembly methods has ushered in an era wherein accurate and complete genome assemblies of diverse species are being produced at an unprecedented rate. For example, tens of thousands to millions of eukaryotic species are expected to be sequenced and assembled in the next decade through various ongoing projects (1–3). These genomes hold the promise to resolve long-standing questions surrounding the evolutionary relationships of species (species trees) and illuminate differences in evolutionary history across the genome (gene trees) (4–6). Species trees are central to many comparative and evolutionary studies (7–11); however, while genome assemblies are widely accessible, species tree inference pipelines remain challenging to use, especially for nonexperts, due to their complexity and computational demands. In order to fully tap into the potential of available data and support a broad range of scientific studies, there is an urgent need for fully automated, scalable, and robust phylogenomic pipelines capable of inferring accurate species trees directly from raw genome assemblies, even those lacking prior annotations.

Despite this need, automation of accurate species tree inference from genome assemblies has been an enduring challenge. While there is no universally agreed-upon method for accurately and reliably inferring species trees, modern phylogenetic pipelines are increasingly adopting methods that account for gene tree discordance and are statistically consistent under various models of genome evolution (12–15). Most often, these pipelines define "genes" by selecting and annotating loci (which are often but not always functional genes) in exemplar species and identifying orthologous regions corresponding to those genes in other species (9, 16–21) (Fig. 1*B*). However, precise gene annotations and orthology inference are not only computationally slow but can require domain expertise for tuning parameters and curating results when reference annotations are unavailable (8, 18, 22, 23). Some studies (5, 8) extract orthologous loci directly from multiple whole-genome alignments (m-WGA) (Fig. 1*B*), such as from Progressive Cactus (24), as input. However, these multiple whole-genome aligners adopt a progressive alignment strategy which is even more computationally intensive and also require high-quality guide trees to maintain

Significance

Understanding how different species are evolutionarily related is key to addressing fundamental questions in biology and medicine. Scientists use species trees to describe these relationships, and they are typically inferred from genome sequences in modern research using methods that account for the mosaic of histories across the genome. However, while genome assemblies are easily available, species tree inference remains challenging for many researchers because the analysis pipeline demands considerable computing power and domain expertise. To address these barriers, we developed ROADIES—an automated, scalable, and user-friendly tool that infers species trees directly from genome assemblies. ROADIES enables researchers with broad expertise to infer accurate evolutionary trees with ease and efficiency, thus democratizing species tree inference.

Author contributions: Y.T. designed research; A.G. performed research; A.G., S.M., and Y.T. analyzed data; S.M. reviewed and edited the manuscript; and A.G. and Y.T. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: yturakhia@ucsd.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2500553122/-/DCSupplemental.

Published May 2, 2025.

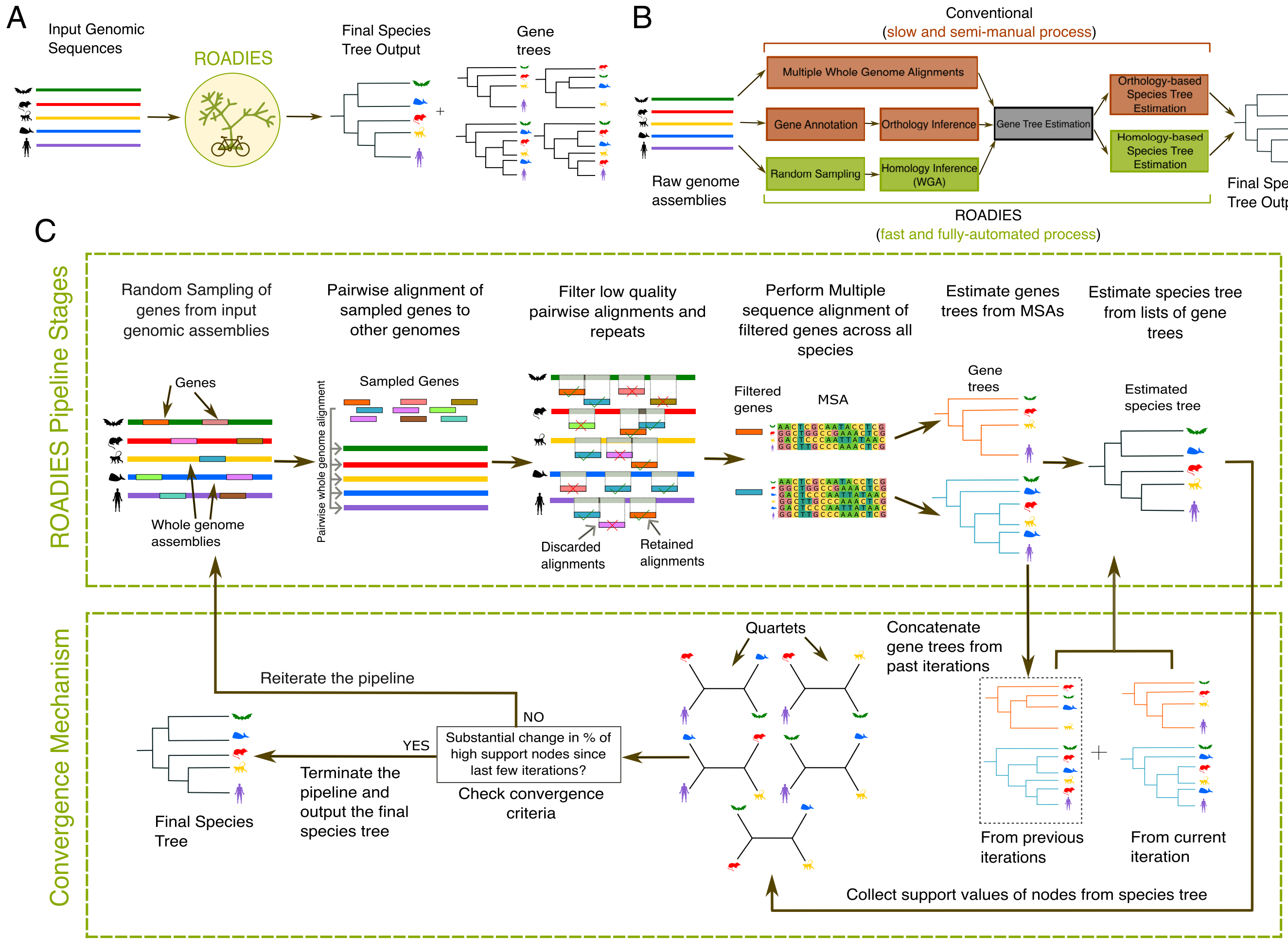
A
Input Genomic Sequences
ROADIES
Final Species Tree Output
Gene trees
B
Conventional
(slow and semi-manual process)
Multiple Whole Genome Alignments
Gene Annotation
Orthology Inference
Random Sampling
Homology Inference (WGA)
Gene Tree Estimation
Orthology-based Species Tree Estimation
Homology-based Species Tree Estimation
Raw genome assemblies
Final Species Tree Output
ROADIES
(fast and fully-automated process)
C
ROADIES Pipeline Stages
Random Sampling of genes from input genomic assemblies
Pairwise alignment of sampled genes to other genomes
Filter low quality pairwise alignments and repeats
Perform Multiple sequence alignment of filtered genes across all species
Estimate genes trees from MSAs
Estimate species tree from lists of gene trees
Genes
Whole genome assemblies
Sampled Genes
Pairwise whole genome alignment
Discarded alignments
Retained alignments
Filtered genes
MSA
Gene trees
Estimated species tree
Convergence Mechanism
Reiterate the pipeline
NO
YES
Substantial change in % of high support nodes since last few iterations?
Check convergence criteria
Terminate the pipeline and output the final species tree
Final Species Tree
Quartets
Concatenate gene trees from past iterations
From previous iterations
From current iteration
Collect support values of nodes from species tree

Fig. 1. An overview of the ROADIES pipeline. (*A*) ROADIES input and output (*B*) A comparison of the different steps involved in the species tree inference in the conventional approaches and ROADIES. (*C*) A detailed view of the ROADIES pipeline's various stages and the convergence mechanism.

the alignment accuracy (24), leading to a need for alignment-free methods to infer guide trees.

As a result of these challenges, full automation of species tree inference has rarely been attempted. While there have been efforts to automate parts of the phylogenetic workflow [e.g., for using Benchmarking Universal Single-Copy Orthologs (BUSCO) genes (25) or ultraconserved elements (UCEs) (26)], the only existing fully automated tools focus on simpler methods that involve fewer parameters and algorithmic stages, such as distance-based methods (27), but have considerable limitations in terms of the accuracy, reliability, and interpretability (*Methods*). There have also been efforts to build assembly free pipelines that infer phylogenetic trees from raw sequencing reads (28–30). While these approaches could be suitable in low sequencing coverage scenarios when reads are not assembled in genome sequences, they may not provide sufficient accuracy, as confirmed by our results. Moreover, in some cases, their reliance on incomplete external databases of preannotated orthologous gene markers (31) poses further challenges.

In this paper, we attempt to automate accurate, reliable, fast, and scalable inference of species trees starting from raw genome assemblies. Our automated tool is called ROADIES, which is an abbreviation for "*Reference-free, Orthology-free, Annotation-free, Discordance-aware Estimation of Species Trees*," as it includes the following key features:

1. Reference-free: ROADIES does not require a reference species, gene set, or genome annotations. It also does not depend on a single genome as a reference for alignments or gene selection and is, therefore, free from reference bias (32, 33).
2. Orthology-free: ROADIES uses multicopy gene trees (inferred from homologous regions) and does not require orthology to be determined prior to gene tree inference. Note that we use the term "gene" to refer to putative coalescent genes (c-genes), i.e., short regions of the genome that are ideally recombination-free (34, 35), and not the more common usage referring to functional genes.
3. Annotation-free: Unlike traditional methods, ROADIES does not require any annotations or m-WGA to be performed. Instead, ROADIES randomly samples loci of fixed user-configurable length from the entire input genomes to generate gene trees. This strategy is not susceptible to poor annotation quality (36, 37), allows sampling of an arbitrary number of loci until a desired confidence level is reached, and can also generate guide trees for annotation pipelines that rely on m-WGA (38, 39).
4. Discordance-aware: ROADIES uses a state-of-the-art and statistically consistent discordance-aware method to combine gene trees into a species tree.

ROADIES has been evaluated for large-scale phylogenetic inference on four diverse datasets with a wide range of evolutionary timescales–placental mammals, birds, pomace flies, and budding yeasts. Additionally, ROADIES has been evaluated on a smaller bamboo dataset that features complex polyploidy. We found that the phylogenetic trees generated by ROADIES were largely concordant with the findings of expert-led studies on these datasets that employed state-of-the-art practices for deducing phylogenetic relationships (5, 8, 9, 40, 41). ROADIES is highly parallelized, scalable, and configurable. We estimate that ROADIES requires a fraction of the runtime and no human intervention compared to the state-of-the-art species tree estimation approaches that include either gene annotation or m-WGA and orthology inference. We also found ROADIES inferences to be more accurate than multiple phylogenomic pipelines available for generating species trees, namely MashTree (27), Read2Tree (28), and a BUSCO-based pipeline (25).

With its speed, accuracy, scalability, and usability, we expect ROADIES to find broad applications. These range from constructing guide trees for multiple whole-genome aligners (24), which are increasingly adopted in comparative and functional genomics research (8, 38, 39, 42–45), to contributing to various evolutionary biology studies and aiding in the resolution of the Tree of Life.

Results

ROADIES Overview. ROADIES is a fully automated pipeline to produce species and gene trees from raw genome assemblies without gene annotations (Fig. 1*A* and *SI Appendix*). There are two ways in which ROADIES differs from the dominant discordance-aware coalescent-based analyses for species tree inference (Fig. 1*B*). *First*, many (though not all) conventional analyses restrict themselves to single-copy gene trees, ideally representing sets of orthologous genes. However, separating orthologs from paralogs is challenging and error-prone, especially for nonexperts (22, 46, 47). Hence, there has been recent focus on using multicopy gene trees to estimate species trees (13, 48–50). By allowing multicopy genes, ROADIES frees itself from the need for orthology inference. It performs species tree inference using ASTRAL-Pro3 (49), a discordance-aware summary method that takes as input multicopy gene trees without needing to separate orthologs from paralogs. In effect, ROADIES leaves teasing out orthology and paralogy to ASTRAL-Pro3, which can perform this step more easily based on the gene trees. *Second*, most conventional pipelines rely on prior annotation, typically of protein-coding genes (9, 17, 51–53), or use m-WGA to select evenly sampled loci from single-copy regions (5, 8) (Fig. 1*B*). Annotation tends to be computationally expensive and requires significant care to avoid many pitfalls with wrong annotation (35–37). Besides, some studies have noted issues in using protein-coding genes as input to phylogenomic tools as they violate the assumptions in multispecies coalescent (MSC) and standard models of sequence evolution used in maximum likelihood methods (5, 18, 54–56). The use of m-WGA is also complicated by the difficulty of inferring such alignments, which itself requires a guide tree (24) and finding single-copy regions. Instead of relying on protein-coding gene annotations or m-WGA, ROADIES samples random sequences from different input genomes and masks out the highly repetitive regions. This proposed approach in ROADIES offers multiple benefits: i) it eliminates the need to perform annotations of protein-coding genes or other regions on input genomes, ii) it eliminates reference bias, as instead of a single genome, genes are sampled randomly from all input genomes with a uniform distribution, iii) by not restricting the genes from coding regions of the genome, this strategy enables the inclusion of intergenic regions, which are more likely to adhere to the assumptions of commonly used model of sequence evolution, and iv) it does not require costly m-WGA and a starting guide tree. These differences with respect to conventional methods greatly simplify the computational process in ROADIES, thus achieving full automation and high speed-up without sacrificing accuracy.

ROADIES is an iterative approach (Fig. 1*C*) and provides configurability in parameters with three convenient modes of operation: accurate (default), balanced, and fast (*SI Appendix*). All modes keep iterating with an increasing gene count till a confident and stable tree is achieved. It starts with 250 randomly sampled genes in the first iteration, each of 500 bp, selected from randomly sampled input genomes. It then finds homologous regions corresponding to these genes, which are identified across genomes using LASTZ (57). Next, in accurate mode, for each gene, a multiple-sequence alignment (MSA) of all its homologs in all genomes is inferred using PASTA (58), followed by multicopy gene tree inference using RAxML-NG (59). Balanced mode uses FastTree (60) for faster gene tree inference. Fast mode eliminates the MSA step and produces gene trees using MashTree (27), a neighbor-joining-based multicopy gene tree of Mash distances inferred from homologs. All modes combine multicopy gene trees into a single species tree using ASTRAL-Pro3 (49), which reports substitution branch lengths (61) and local posterior probability (localPP) (62) confidence scores for each branch in the species tree. ROADIES doubles the gene count at each iteration until the number of high-confidence branches (i.e., localPP ≥ 0.95) changes by less than 1% from the previous iteration. ROADIES also provides an additional setting for deep phylogenies and uses appropriate filtering strategies between stages to ensure the quality of final results (see *SI Appendix* for full details).

ROADIES Estimates an Accurate Phylogenetic Tree of 240 Placental Mammals. We evaluated the performance of ROADIES in accurate mode using the genome assemblies of 240 placental mammals provided by the Zoonomia consortium (8) (Dataset S1). The assemblies include representative species from all 20 orders of the placental mammalian phylogeny (Dataset S1). For accuracy comparison, we used the tree topology provided by Zoonomia (8) as a representative of the state-of-the-art scientific literature. We found that the ROADIES-inferred phylogeny is largely in agreement with the Zoonomia phylogeny (Fig. 2), with a species-level normalized Robinson–Foulds distance (normRF) (63) of 0.038. All species are correctly assigned to their phylogenetic orders, though minor differences occur in the arrangement of orders (Fig. 2*B*). At the order level, only two relationships changed (nodes 1 and 2, Fig. 2*B*), resulting in a normRF distance of 0.11. These contested relationships are resolved with low confidence (localPP < 0.78) compared to other highly confident branches (localPP $\geqq 0.95$). ROADIES samples the number of genes stochastically, achieving near-uniform representation of all species in gene trees. However, some species, such as rodents, have comparatively fewer aligned genes (Fig. 2*A*), likely because of their significantly faster evolution rate compared to other neighboring orders (64, 65), which makes it challenging to infer homology. Despite this, there is substantial representation of the noncoding regions in ROADIES analysis (Dataset S6). Next, we compare the two phylogenies (reference and ROADIES) with a specific focus on historically contested branches:

Agreement on Atlantogenata. Historically, there has been a debate over the correct placement of the superorders Afrotheria, Xenarthra, and Boreoeutheria. ROADIES supports the Atlantogenata hypothesis (Afrotheria + Xenarthra) with full localPP support (Fig. 2*B*), aligning with the Zoonomia and prevailing scientific consensus (66–72), although alternative hypotheses of Epitheria (Afrotheria + Boreoeutheria) (73) and Exafroplacentalia (Xenarthra + Boreoeutheria) (74) have also been proposed.

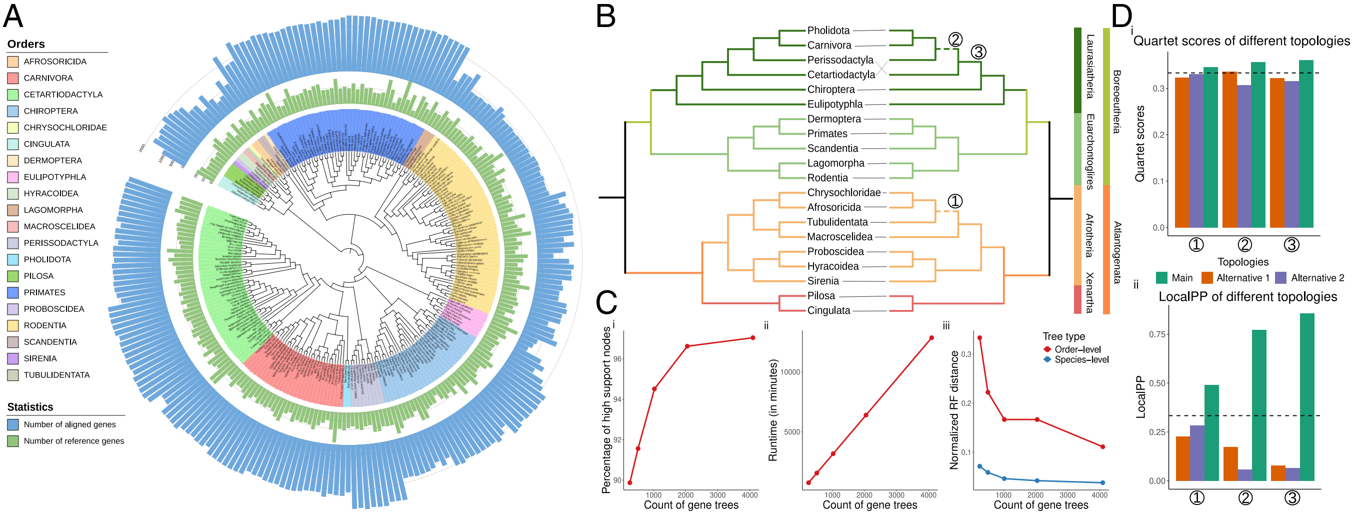

Fig. 2. ROADIES results evaluated on the dataset of 240 placental mammals (in the accurate mode). (*A*) The species-level phylogenetic tree of 240 placental mammals estimated by ROADIES. The number of genes aligned to each species (blue) and the count of genes sampled from each species (green) (8) are also shown. (*B*) Order-level trees of 240 placental mammals estimated by ROADIES (on the *Right*) and the reference tree from the Zoonomia consortium (8) (on the *Left*). Dashed branches show the differences between the two trees. (*C*) ROADIES convergence in accurate mode. As the number of gene trees increases, we show the percentage of highly supported species tree nodes with localPP ≧ 0.95 (plot i), the linear increase in runtime (ii), and the normRF of the final species tree to the reference tree (iii). (*D*) Quartet scores i) and localPP branch support ii) of all three topologies around three branches (marked in *B*), which had low support in the final tree.

Agreement on the placement of Sirenia and Scandentia. The placement of Sirenia within the clade Paenungulata (Hyracoidea + Proboscidea + Sirenia) (73, 75) and Scandentia within the superorder Euarchontoglires have been a subject of ongoing debate (71, 73, 75, 76), but ROADIES and Zoonomia tree topologies are in agreement here (Fig. 2*B*), consistent with recent studies (71, 75, 76).

Disagreement within the Laurasiatheria phylogeny. ROADIES differs from the Zoonomia tree in placing Perissodactyla and Cetartiodactyla within the superorder Laurasiatheria. Zoonomia places Perissodactyla as the sister group to Carnivora + Pholidota (clade Zooamata). ROADIES places it as the sister group of the clade [Cetartiodactyla, (Carnivora, Pholidota)] with a modest localPP of 0.77 (node 2, Fig. 2*B*). This is also the only order-level branch where the ROADIES tree deviates from the inference of CASTER (77), a recent site-based method utilizing multiple whole-genome alignments. There is substantial gene tree discordance around this branch, with two alternatives getting quartet support above ⅓, a pattern that points to possible violations of the MSC model (78) due to incomplete lineage sorting (ILS) (Fig. 2*D*). Since the correct placement of orders within Laurasiatheria remains contentious, with different studies suggesting Perissodactyla as the sister group to Chiroptera (79), Carnivora (8, 68, 80), or Cetartiodactyla (71, 81), the inference made by ROADIES offers a plausible hypothesis supported by some prior studies.

Disagreement in the placement of Macroscelidea and Tubulidentata. ROADIES differs from the Zoonomia tree in placing two Afrotherian orders, Macroscelidea and Tubulidentata, which contributes to the unresolved phylogeny of early placental mammals. While Zoonomia tree placed Macroscelidea as the sister of Tubulidentata (8, 82), O'Leary et al. (73) inferred Tubulidentata as the sister group of Paenungulata, whereas Meredith et al. (75) combined Tubulidentata with Afrosoricida + Macroscelidea. ROADIES places Macroscelidea as a sister with the clade of Tubulidentata and Afrosoricida + Chrysochloridae, albeit with a low localPP support of 0.49 (topology of node 1, Fig. 2*B*) and backed by only 34% of gene tree quartets (Fig. 2*D*). This placement by ROADIES, although supported by some prior studies (71, 83, 84), remains unresolved due to its low confidence. Even the Zoonomia authors reported low confidence for these Afrotherian orders in their analysis (4), highlighting the need for further investigation.

ROADIES Generalizes Well on Different Datasets with a Range of Evolutionary Timescales. To evaluate the generalizability of ROADIES, we tested it on three additional large-scale datasets: i) 100 drosophilid (pomace fly) genomes (40) (Dataset S2), ii) 363 avian (bird) genomes (42) (Dataset S3), and iii) 332 budding yeast genomes (subphylum Saccharomycotina) (Dataset S4). The three datasets vary widely in evolutionary timescales, spanning over 400 My for budding yeast species (9), and in complexity, with the avian phylogeny being among the most confounding cases in prior studies (5, 18, 42, 85, 86). The reference trees for these datasets are taken from the studies of Kim et al. (40), Stiller et al. (5), and Shen et al. (9), respectively (*Methods*).

Drosophilid dataset. On the dataset of 100 drosophilid genomes, ROADIES converged in 1,105 min or 18 h 25 m (<1 d) with 1,627 gene trees (*SI Appendix*, Fig. S9*C*). The final phylogeny had 94% high-support branches (*SI Appendix*, Fig. S9 *C*, *i*). ROADIES accurately identified all the group-level relationships without any discrepancies with the reference (group-level normRF is 0) (Fig. 3*B* and *SI Appendix*, Fig. S9 *C*, *iii*), which also matches with most studies (87–89). ROADIES also matched the reference for the relationships within groups, apart from a few debatable relationships within the group melanogaster as follows (species-level normRF is 0.062).

Kim et al. (40) infer *Drosophila mauritiana* and *Drosophila simulans* as sisters with a lower support value of 0.55 (localPP used by Kim et al. (40) is reported by ASTRAL) (node 3 in red, *SI Appendix*, Fig. S9*A*), while ROADIES groups *Drosophila sechellia* and *Drosophila simulans* as sisters with a support value of 0.548 (node 3, *SI Appendix*, Fig. S9 *B*, *ii*). This reflects high gene tree discordance, which is also evident from the relatively close Quartet scores for the different alternatives (*SI Appendix*, Fig. S9 *B*, *i*). Moreover, Kim et al. (40) places *Drosophila kurseongensis* with the clade of (*Drosophila carrolli*, *Drosophila rhopaloa*) with a support value of 0.99 (node 2 in red, *SI Appendix*, Fig. S9*A*), whereas ROADIES infers *Drosophila kurseongensis* as a sister with the clade of [*Drosophila fuyamai*, (*Drosophila carrolli*, *Drosophila rhopaloa*)] with localPP of 0.549

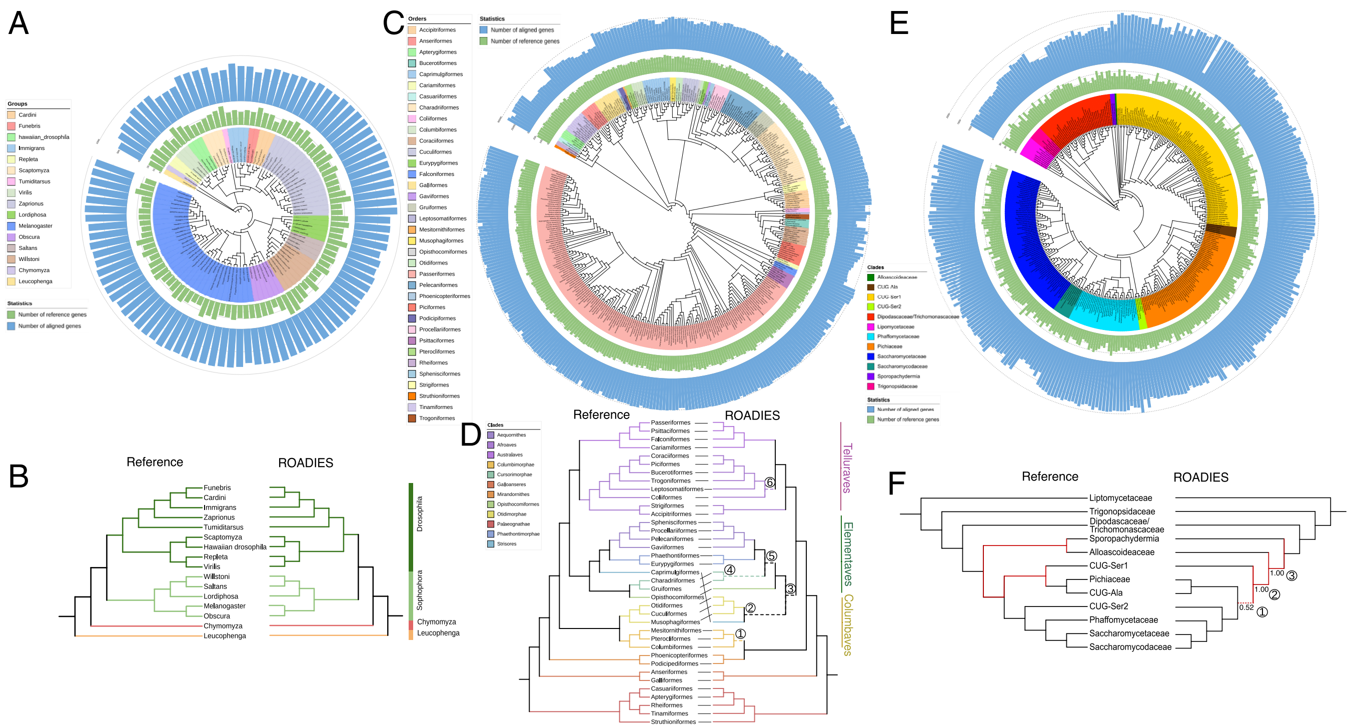

Fig. 3. ROADIES results evaluated on the dataset of (*A* and *B*) 100 drosophilid species, (*C* and *D*) 363 aves, and (*E* and *F*) 332 budding yeasts (in the accurate mode). (*A*, *C*, and *E*) The species-level phylogenetic tree of (*A*) 100 drosophilid species, (*C*) 363 avian species, and (*E*) 332 budding yeasts estimated by ROADIES. All trees were estimated in the accurate mode of ROADIES, with the budding yeast tree estimated with deep setting additionally enabled. The number of genes aligned to each species (blue) and the count of genes sampled from each species (green) are also shown. (*B*, *D*, and *F*) Cophylogenetic plots comparing the reference tree (on the *Left*) with the tree estimated by ROADIES (on the *Right*) shown at (*B*) group-level for 100 drosophilid species, (*D*) order-level for 363 avian species, and (*F*) clade-level for 332 budding yeasts. Reference and ROADIES trees match exactly at the group-level for drosophilid species. Dashed branches in the ROADIES trees show the differences with the reference trees in the remaining two cases (*D* and *F*).

(node 2, *SI Appendix*, Fig. S9 *B*, *ii*). This difference may stem from sampling or statistical inconsistency since similar ROADIES experiments (refer to the section "ROADIES is scalable and stable") inferred the exact topology with the reference tree with a localPP of 1. Finally, the placement of *Drosophila ficusphila* by ROADIES is supported by localPP of 0.876 (main topology of node 1, *SI Appendix*, Fig. S9 *B*, *ii*), which differs from the reference tree with the confidence score of 0.51 (node 1 in red, *SI Appendix*, Fig. S9*A*).

Avian dataset. ROADIES is estimated to require 2,811 instance hours (117 d) on a single 16-core Amazon Web Services (AWS) EC2 instance (r6a.4×large) to reach convergence for 363 avian genomes dataset (we parallelized the pipeline using 16 such instances to ensure it completed within approximately 8 d). The final species tree consisted of 99% highly supported nodes but required seven iterations and 62,413 gene trees for generating a stable converged tree (*SI Appendix*, Fig. S9 *E*, *i*). The much slower convergence of ROADIES compared to other datasets matches the understanding that avian phylogeny is among the most challenging. Remarkably, the gene count in ROADIES is also close to the 63,430 gene trees Stiller et al. (5) used in their analyses.

The ROADIES phylogeny achieved a species-level normRF score of 0.027 compared to the reference tree (*SI Appendix*, Fig. S9 *E*, *iii*), suggesting a reasonably high congruence between the two phylogenies, with differences mostly in low-support branches (Fig. 3 *C* and *D*). At the order level, ROADIES and the reference tree are aligned i) in capturing the relationship of orders within Australaves and Aequornithes within Neoaves and ii) in the relationship of orders within core waterbirds, both of which have been involved in considerable debate, and have been even suggested by some authors to represent a hard polytomy (90). ROADIES also recovered the so-called magnificent seven groups that have been identified in recent genome-wide analyses (91).

Six ordinal-level branches were different in ROADIES compared to the reference (order-level normRF = 0.228), and all of those are among those that are contentious, and in every case, these had quartet frequencies that were close to ⅓, which signifies a lack of resolution (Fig. 3*D* and *SI Appendix*, Fig. S9*E*). This level of discordance is not unusual for the avian phylogeny, and similar or higher levels of discordance are observed between the four expert-led authoritative studies on this subject (5, 18, 85, 86) (*SI Appendix*, Fig. S4*E*). Specifically, ROADIES differs from the reference tree in the placements of Tinamiformes and Rheiformes in Palaeognathae, but past studies have not reached a consensus in this case, and some studies support ROADIES' findings in their main analyses (92–94). There are also two differences in early neoavian diversification. i) ROADIES infers Columbimorphae and Phoenicopterimorphae as sister groups, recovering the Columbea group proposed by Jarvis et al. (18) and Yuri et al. (95) but differs from our reference tree. As Mirarab et al. (96) have recently shown, this result is sensitive to the choice of loci–they reported a recombination-depleted region in chromosome 4 of chicken which has an unusually strong signal for Columbea. ii) ROADIES infers Otidimorphae as sisters to Caprimulgiformes, breaking the Elementaves group proposed by Stiller et al. (5). However, the support for this relationship in ROADIES is low (localPP = 0.69) (*SI Appendix*, Fig. S9*E*), and the ROADIES tree is compatible with Elementaves if we contract branches with localPP < 0.95. The remaining two differences, in which ROADIES places Accipitrimorphae as sister to all other land birds (Telluraves) and moves Opisthocomiformes within Elementaves, are possibilities that Stiller et al. (5) have not fully ruled out as they had low confidence and consistency for these relationships in their results.

Yeast dataset. We also tested ROADIES with 332 budding yeast datasets from subphylum Saccharomycotina, collected from the data published by Shen et al. (9) (Dataset S4), with the estimated divergence time of over 400 My, making it a challenging phylogeny to accurately estimate. With this input dataset, ROADIES consumed around 3.52 d to generate the final species tree on 64-core AWS EC2 instances, achieving a normRF of 0.170 with 2,813 gene trees and 94.52% of high-support nodes (*SI Appendix*, Fig. S9*G*).

Similar to the reference tree, ROADIES was able to accurately capture the grouping of species within their respective clades (Fig. 3 *E* and *F*). It successfully captured stable and consistent relationships within Saccharomycetes, i.e., between clades Saccharomycetaceae, Saccharomycodaceae, Phaffomycetaceae, and CUG-Ser2 (97, 98), and among early diverging clades such as Lipomycetaceae, Trigonospidaceae, and Dipodascaceae/Trichomonascaceae. Notably, the clade Pichiaceae, which had been reported as nonmonophyletic in earlier studies (97), was recovered by ROADIES.

However, some differences were observed between the ROADIES tree and the reference tree, particularly in the placements of the clades Sporopachydermia, Alloascoideaceae, and CUG-Ser1 (Fig. 3*F*), resulting in a clade-level normRF of 0.22. The placement of CUG-Ser1, which consists of families Debaryomycetaceae and Metschnikowiaceae, has historically been a subject of debate (98, 99), and ROADIES was not confident about this placement (node 1, localPP = 0.52, *SI Appendix*, Fig. S9 *F*, *ii*). Likewise, the reference placed Sporopachydermia and Alloascoideaceae as sister clades, which is consistent with most recent studies (100), but some earlier studies suggested an alternative placement consistent with ROADIES (98).

ROADIES Estimates Accurate Species Tree Even in the Presence of Complex Polyploidy. Polyploidy is known to present challenges to phylogenetic inference (101, 102). To test ROADIES' accuracy in the presence of polyploidy, we used a dataset of 11 bamboo species (family: Poaceae, subfamily: Bambusoideae). As discussed by Ma et al. (41), these bamboo genomes have undergone recurrent hybridization between diploid ancestors of woody lineages, followed by polyploidization and introgression between ancestral woody and herbaceous lineages, resulting in various ploidy levels (diploid, tetraploid, and hexaploid) (Dataset S7). Both the dataset and the reference tree for this comparison were obtained from Ma et al. (41) (*Methods*).

ROADIES converged on this dataset in two iterations, inferring 263 gene trees and requiring 51 min of runtime. The resulting species tree exhibited high support for all branches and its topology matched exactly with the reference tree (normRF = 0, *SI Appendix*, Fig. S10 *A* and *B*). Notably, ROADIES accurately grouped multiple species of herbaceous and woody bamboos while confidently resolving the relationships among the four major bamboo lineages, consistent with previous studies (41, 103). We observed that the diploid genomes were the smallest (626 Mb, on average) and contributed fewer copies per gene tree, as expected (*SI Appendix*, Fig. S10*C*). However, compared to tetraploids, hexaploids did not show an increase in gene copies despite higher ploidy. This could be related to the tendency of genome assemblers (104) to collapse polyploid regions with lower heterozygosity and the fact that due to the accumulation of transposable elements (105), tetraploid assemblies are larger than hexaploid ones (1.63 Gb vs. 1.12 Gb). Nonetheless, ROADIES inferred an accurate phylogeny, demonstrating that its multicopy gene tree-based inference approach based on ASTRAL-Pro3 is robust even under complex scenarios.

ROADIES Outperforms State-of-the-Art Species Tree Estimation Pipelines. We compared ROADIES with several state-of-the-art species tree estimation pipelines, including MashTree (27), Read2Tree (28), and the concatenation- and coalescent-based pipelines using BUSCO (25) genes (Table 1 and *SI Appendix*, Fig. S8). These pipelines were evaluated for performance and accuracy using a smaller 48-bird dataset meant to challenge methods [taken from Jarvis et al. (18)—Dataset S5] on a 64-core AWS EC2 machine, with Stiller et al.'s tree (5) serving as the reference.

Among the baseline tools, MashTree stands out as the only fully automated pipeline that, like ROADIES, requires no other input apart from the raw genomic assemblies. However, MashTree employs a simple workflow based on Mash distances (106) and a Neighbor-Joining algorithm, as implemented in QuickTree (107). As expected, MashTree was the fastest of all methods tested, completing analyses in significantly less time than ROADIES (Table 1). However, it also recovered only around one third of branches from the reference phylogeny (normRF: 0.622), indicating inaccuracy. For example, MashTree inaccurately inferred most clade relationships within Neoaves (*SI Appendix*, Fig. S8).

Read2Tree, while similar to MashTree and ROADIES in its goal of generating species trees in a scalable manner, uses a semi-automated approach that differs from ROADIES in two fundamental ways. *First*, Read2Tree is an assembly-free approach which relies on raw sequencing reads as input, making its accuracy sensitive to sequencing coverage. In our experiments, Read2Tree failed to infer phylogeny with low-coverage reads for certain species (*Taeniopygia guttata*). *Second*, Read2Tree also requires the users to provide orthologous gene markers downloaded from the OMA database (108) as input. However, the OMA database is incomplete; for example, it contained data for only 7 of the 48 bird species we evaluated. We observed that when the number of reference species increased from 3 to 7, the normRF improved only slightly from 0.711 to 0.666 (Table 1). In summary, even though the runtime of Read2Tree was comparable to ROADIES, the accuracy was far worse, and like MashTree, it grouped most clades within Neoaves inaccurately (*SI Appendix*, Fig. S8).

We also evaluated pipelines based on BUSCO genes, using both concatenation and coalescent approaches (*Methods* and *SI Appendix*, Fig. S8). These pipelines, unlike ROADIES, rely on the OrthoDB reference database (31), which provides single-copy, complete protein-coding genes as markers. Although BUSCO has automated the download step to a large extent, the availability and completeness of the reference database still pose a concern. For example, only 8,338 BUSCO genes are available for birds (*aves_odb10* database). In our evaluations, ROADIES outperformed BUSCO-based pipelines in terms of accuracy, achieving

Table 1. Runtime and accuracy comparison of ROADIES with state-of-the-art species tree estimation pipelines using 48 avian species

Pipeline	ROADIES	MashTree	Read2Tree (three references)	Read2Tree (seven references)	BUSCO (concatenation)	BUSCO (coalescent)
NormRF with reference	0.133	0.622	0.711	0.666	0.522	0.533
Runtime	109 h	2 min	48 h	98 h	170 h	178 h

lower normRF distances relative to the reference tree, and also in terms of running time (Table 1). In particular, ROADIES performed far better in resolving relationships within Neoaves. Previous studies have shown that coding regions are not optimal for Neoaves (18, 54, 55), which possibly explains the poor accuracy of the BUSCO pipeline.

In summary, existing tools are inaccurate relative to ROADIES for challenging phylogenies, and some are slower too. ROADIES stands out as the first fully automated species tree inference method with high accuracy for challenging data and independence from external databases.

ROADIES Offers Flexibility in Balancing Speed and Accuracy. Species tree inference often involves a trade-off between the speed and accuracy of the final results. In most biological applications, accuracy is paramount, though speed is desirable and often a constraint. We have implemented ROADIES as a Snakemake workflow (109), which allows modular implementation and exposes numerous parameters of individual tools to end users to provide flexibility (*SI Appendix*). We anticipate that most users will not need to adjust the ROADIES pipeline, staying with the default mode, which prioritizes accuracy. However, we have provided two convenient additional modes in ROADIES, fast and balanced, to cater to users who might be resource-constrained or might be working with datasets or applications that might be more tolerant of errors (see *SI Appendix* for details).

We evaluated ROADIES' balanced and fast modes of operation for mammals, pomace flies, and birds datasets (*SI Appendix*, Fig. S1). For mammals, ROADIES in balanced mode used 3,738 gene trees, converging in 160 h 55 m (6.7 d) with 96% high-support nodes on a 16-core AWS EC2 instance (r6a.4×large) (*SI Appendix*, Fig. S1 *E*, *i*). Fast mode required 14,929 gene trees and took 98 h 21 m (4.1 d), achieving 97% high-support nodes (*SI Appendix*, Fig. S1 *F*, *i*). Both modes were 1.33-fold and 2.17-fold faster than the accurate mode, respectively, resolving the once-highly debated relationship between Afrotheria, Xenarthra, and Boreoeutheria (66–72), (*SI Appendix*, Fig. S1 *A* and *B*). The balanced mode aligns more closely with the reference tree (order-level normRF: 0.05), differing only in the placement of Perissodactyla and Cetartiodactyla (*SI Appendix*, Fig. S1*C*), while the fast mode showed additional differences which were previously debated, including the placements of Afrosoricida and Tubulidentata (71, 73, 75, 84), Scandentia in Euarchontoglires (71, 73, 75, 76), Perissodactyla and Cetartiodactyla in Laurasiatheria (68, 71, 73, 79–81), and Proboscidea in the clade Paenungulata (73, 75) (order-level normRF: 0.33, *SI Appendix*, Fig. S1*D*).

For pomace flies, balanced mode required 8,192 gene trees and 16 h 17 m to get a converged tree with 100% highly supported nodes (group-level normRF: 0.071, *SI Appendix*, Fig. S2 *E*, *i*), while fast mode required 4,096 gene trees and 1 h 59 m, with 95% high-support branches (group-level normRF: 0.214, *SI Appendix*, Fig. S2 *F*, *i*). Both modes captured stable orders well (87–89), though fast mode showed some placement differences, such as for Tumiditarsus and Zaprionus (*SI Appendix*, Fig. S2 *C* and *D*). For birds, the balanced mode required 62,441 gene trees and 1,066 h (44.4 d), while the fast mode needed 31,692 gene trees and 101 h (4.2 d) runtime to converge with 98% and 96% support, respectively (*SI Appendix*, Fig. S3). However, even though the balanced and fast modes provided 2.6-fold and 27.8-fold speed-up over the accurate mode, they both struggled with accuracy, resulting in order-level normRFs of 0.6 and 0.743, respectively.

In summary, the balanced and fast modes demonstrate good speed-ups with reasonable accuracy and might be suitable for certain resource-constrained applications. However, like our baseline methods (Table 1), even these modes struggle with accuracy when presented with a challenging phylogeny, such as birds.

ROADIES Is Scalable and Stable. We have designed ROADIES to support large-scale phylogenetic analyses on high-performance computing environments, leveraging Snakemake (109) for efficient parallelization. Specifically, it exploits thread and process-level parallelism, supporting large datasets with scalable workflows. To characterize the strong scaling efficiency, we tested ROADIES with varying numbers of system cores on AWS EC2 instance for 100 drosophilid datasets in accurate mode (*Methods*), fixing the gene count to 4,000. We observed roughly linear speed-up with the number of cores—the runtime decreased by 7.3-fold (from 3,147 to 430 min) as core count increased from 8 to 128, achieving a scaling efficiency of 57.7% (*SI Appendix*, Fig. S5*A*). A slight runtime increase with 256 cores is primarily due to the communication overheads, which we will address in the future releases. Runtime also scaled well with the number of input species, increasing 19.8-fold (from 90 to 1,778 min) as species count grew 6.7-fold (from 15 to 100), indicating a superlinear but subquadratic scaling (*SI Appendix*, Fig. S5*B*).

Given that ROADIES uses random sampling of genes from input genomes, it could be susceptible to sampling noise. To evaluate ROADIES' consistency despite random gene sampling, we conducted four separate trials on the drosophilid dataset, each using a different set of randomly selected genes. While one trial (experiment ID 3) converged in four iterations compared to five in others, the resulting phylogenies were nearly identical, with a maximum species-level normRF distance of 0.04 between them (*SI Appendix*, Fig. S5 *C–F*). Moreover, the percentage of highly supported nodes in the final converged species tree varied minimally across trials and followed similar trends with increasing gene tree count (*SI Appendix*, Fig. S5*E*).

Slight differences in species-level normRF distances between trials result from a few debated drosophilid phylogenies: i) *Drosophila fuyamai* and *Drosophila kurseongensis*, where experiments alternated between matching or diverging from the reference, both with low support (localPP = 0.587 and 0.652); ii) *Drosophila mauritiana*, *Drosophila simulans*, and *Drosophila sechellia*, where all trials consistently differed from the reference, which also had low support (localPP = 0.55); and iii) *Drosophila quadrilineata* and *Drosophila repletoides*, where 3 of 4 trials matched the reference. At the order level, normRF distances were consistent except for Tumiditarsus (*Drosophila repletoides*), which caused occasional jumps from 0 to 0.07. Overall, ROADIES produced stable phylogenies with negligible dependence on specific sampled genes, with differences only limited to a few debatable and low-confidence topologies.

Discussion

ROADIES can perform discordance-aware species tree construction directly from raw (unannotated) genome assemblies without relying on a single reference genome, input alignments, gene annotations, and predefined orthology. With genomic data continuing to grow exponentially, and with telomere-to-telomere assemblies now possible for complex genomes at reasonable costs (8, 42, 110), ROADIES is addressing a critical problem of automating species tree inference from genomes with high accuracy. ROADIES achieves these results based on three design decisions. *First*, instead of predefined genomic regions with specific characteristics, such as protein-coding genes, UCEs (111), or transposons (80, 112, 113), ROADIES is based on a random sampling of loci from input genomes. This eliminates the need for genome annotations and the bias caused by using a single reference genome. *Second*, the summary method that ROADIES uses for species tree inference can work with multicopy gene trees that account for both orthology and paralogy (49). This eliminates the need for strict orthology inference from genomes—a longstanding problem in automating reliable pipelines. While good progress has been made on this problem in recent years, such as the development of TOGA (114), orthology inference typically requires intensive algorithms to accurately distinguish orthologs from paralogs and pseudogenes. On the other hand, homology (a combination of orthology and paralogy) is not just easier to infer; it is also computationally less expensive. Once gene trees are inferred, many cases of paralogy become immediately clear, a fact that ASTRAL-Pro exploits. *Third*, choosing the correct number of gene trees is a challenge when genomes are available, as it heavily depends on the complexity of the evolutionary histories of the input genomes, which is often unknown in advance. ROADIES tackles this issue through a convergence algorithm. As expected, ROADIES requires a far greater number of gene trees to converge for the bird phylogeny, the most complicated case in our datasets, compared to the phylogeny of mammals, flies, bamboos, and yeasts. Another alternative to predefining loci is to use genome-wide site-based species tree estimation tools, such as CASTER (77). However, unlike ROADIES, CASTER requires m-WGA as input, which is computationally intensive to generate and requires an input guide tree. Thus, ROADIES can complement CASTER by providing an accurate guide tree for the m-WGA step. Furthermore, note that CASTER is less suitable for organisms with rampant duplication since it is currently limited to single-copy regions.

In the future, we aim to further improve the capabilities of ROADIES. While our current implementation is scalable for thousands of genome sequences, the runtime might be unreasonably high for tens of thousands of genomes or beyond. We believe GPUs could be leveraged to accelerate the critical stages of our pipeline (115, 116), but this would necessitate the development of new GPU-accelerated libraries. Moreover, incorporating recently proposed divide-and-conquer strategies for species tree construction (117) also presents a promising avenue for improving the runtime of ROADIES and would enable placing new taxons on existing species trees.

Methods

Input Datasets and Reference Phylogenies. Four large-scale genomic assembly datasets from recent studies were used for evaluation: 1) 240 species from the infraclass Placentalia (alternatively referred to as "placental mammals"), 2) 100 species belonging to the subfamily of Drosophilinae and Steganinae, 3) 363 species from the class Aves, and 4) 332 budding yeast species from the subphylum Saccharomycotina. We also used other smaller datasets, such as 1) 48 species from the class Aves for the baseline comparison, 2) 11 species of bamboo from the family Poaceae and subfamily Bambusoideae for evaluating ROADIES with polyploid genomes, and 3) 10 species of whales from the order Cetartiodactyla (subset of 240 species of mammalian dataset) for evaluating gene length variation. Dataset details, including taxonomy and NCBI IDs, are provided in Datasets S1–S8.

The 240 species of placental mammals are collected from the Zoonomia consortium (8). Zoonomia consortium earlier published 241 species with two assemblies from the same species–*Canis lupus familiaris* (Domestic dog). We removed the duplicate dog assembly (with NCBI ID GCF_000002285.3) and kept the one with NCBI ID GCA_004027395.1. The 363 avian species are collected from the Birds 10k Genome Project's dataset (42). For 100 drosophilid genomes, we collected the dataset from NCBI BioProject ID PRJNA675888 used by Kim et al. (40). Genomic data of 332 yeast species were collected from the Figshare repository published by Shen et al. (9). Genomic assemblies of 11 species of bamboo were taken from Ma et al. (41).

We used reference trees from authoritative studies as a proxy for ground truth to compare the accuracy of the species tree estimated by ROADIES. The reference tree of the mammalian dataset from Zoonomia consortium (8), was generated using Progressive Cactus (24). The reference tree of 100 drosophilid genomes, from Kim et al. (40), was generated using 250 single-copy BUSCO genes (using Amino acid sequences) and running MAFFT, RAxML-NG, and ASTRAL-MP (in summary mode) sequentially. For the avian dataset, these studies in the last decade by Kuhl et al. (86), Prum et al. (85), Jarvis et al. (18), Feng et al. (42), and Stiller et al. (5) are considered authoritative. While these studies include all major orders of the avian phylogeny, there is observable discordance between them (*SI Appendix*, Fig. S4). We chose Stiller et al. (5) as our reference [which used 63,430 loci (44,846 intronic, 14,972 exonic, and 4,985 UCE loci)] since it is the largest and the most recent, involving all 363 species we used in our evaluation. The yeast phylogeny followed Shen et al. (9), and the bamboo tree is taken from Ma et al. (41).

Execution Environment and Runtime Estimates. ROADIES was executed in a custom conda environment set up via a custom script roadies_env.sh. All experiments (Figs. 2 and 3) ran on 16-core memory-optimized AWS R6a (r6a.4×large) instances. Input datasets, downloaded from NCBI (IDs in Datasets S1–S5), were stored in AWS EBS volumes as .fa.gz files and provided as the GENOMES parameter in the configuration file. To parallelize ROADIES for higher GENE_COUNT values, we ran the pipeline across multiple AWS instances. For GENE_COUNT=64000, ROADIES was executed on 16 r6a.4xlarge EC2 instances, each processing 4000 genes. After completion, gene trees and mapping files from all instances were merged to create the final input for ASTRAL-Pro3, generating the final species tree. ROADIES also records the total runtime in time_stamps.csv, representing wall-clock execution time. In converge mode, runtimes for all iterations are logged and reported for all experiments in the manuscript.

Cophylogenetic Plots and Accuracy Estimates. ROADIES generates rooted trees by default. To produce rooted trees, we determined the most identical node in the final species tree produced by ROADIES to match the root of the reference tree (which is provided through the REFERENCE parameter in the config.yaml file) through a custom reroot.py script. We manually derived the order-level phylogeny from the final roadies.nwk output. After generating the species and order level tree, we constructed the cophylogenetic plot to compare the tip displacements and topological differences between ROADIES and reference tree using iTOL software (118). We manually rotated the tips for better one-on-one mapping between both trees with minimal tip displacements. ROADIES quantifies tree differences using the normalized Robinson-Foulds distance, estimated with the ETE3 function (119) tree2.compare(tree1).

Stability and Scalability Analysis. To perform the stability test, we ran ROADIES with 100 drosophilid genomes and the same set of configuration parameters (with fixed GENE_COUNT = 4,000). After getting the final species trees from four independent instances, we calculated the normRF for all four species trees with the reference tree. Following the same approach, we performed the scalability test, with the difference of providing different numbers of genomes and varying GENE_COUNT as input across four AWS instances.

Local PP and Quartet Score Analysis. The final species tree generated by ROADIES was compared to the reference tree, with dissimilar branches analyzed using localPP scores and quartet scores as metrics of comparison, which are estimated by ASTRAL-Pro3. LocalPP represents the likelihood of a branch being true based on the input gene trees. ASTRAL-Pro3 provides localPP values for each of the three possible topologies of a quartet formed by a specific internal branch. When the main topology's localPP value is significantly higher than the others, it indicates less contention, a clearer resolution, and strong support from the input gene trees, while similar values across topologies suggest ambiguity. The quartet score quantifies how many gene trees support a specific topology formed by an internal branch. If all three topologies of a branch have nearly equal quartet scores (around 0.33), it suggests no clear winner, whereas a dominant score reflects strong support.

Baseline Methods. To assess ROADIES' performance, we compared the runtime and accuracy of the final tree it generated with those produced by the state-of-the-art pipelines described below. All the baseline experiments were run on a 64-core AWS EC2 instance (r6a.16×large). We estimated the total runtime in seconds by measuring its wall clock time.

Analysis of Read2Tree. We compared ROADIES with another semiautomated phylogenetic approach, Read2Tree (version 0.1.5) (28) which generates species tree from raw reads and orthologous gene markers from reference species as input. We used three species as reference (*Taeniopygia guttata, Gallus gallus, Meleagris gallopavo*), downloading orthologous gene markers through the OMA database (by setting Minimum fraction of covered species as 0.8 and Maximum nr of markers as −1). The rest of the species' input data are downloaded as raw reads from the Sequence Read Archive (SRA) database in .fastq format. After collecting the input data, we ran Read2Tree for all species. We repeated the same experiment entirely for 48 birds species dataset by choosing seven reference species (*Haliaeetus leucocephalus*, *Anas platyrhynchos platyrhynchos*, *Meleagris gallopavo*, *Gallus gallus*, *Taeniopygia guttata*, *Melopsittacus undulatus*, *Tyto alba*) and downloaded their orthologous markers from OMA database and rest of the 41 species as raw reads.

Analysis of BUSCO-based pipeline. For a deeper comparison of ROADIES with other widely used tools and pipelines, we ran BUSCO (version 5.7.1) (25) (https://gitlab.com/ezlab/busco) to generate gene markers and created two pipelines: one using a concatenation-based approach, and another using coalescent-based methods to estimate final species tree of 48-bird dataset. For the concatenation-based pipeline, we adapted an open-source BUSCO pipeline originally designed for protein sequences to support DNA sequences (https://gitlab.com/ezlab/busco_usecases). The six-stage pipeline begins with running BUSCO in genome mode with Augustus as the gene predictor. We selected Aves OrthoDB database (*aves_odb10*) database (31), considering the nature of our input dataset. Next, we prepared a custom script to extract single-copy complete BUSCO genes present in all 48 species, yielding around 500 genes out of a total of 8,338 BUSCOs from the Aves OrthoDB database. These extracted genes were individually aligned using MAFFT (version 7.526) (120) and then filtered using TrimAl (version 1.5.0) (121). Last, we ran a custom script to concatenate all filtered MSAs and to create a supermatrix, which is then passed to RAxML (122) (version 8) to generate the final species tree. Unlike ROADIES, which allows some missing data (up to 4 or 10% of species, whichever is maximum), this pipeline requires genes to be present in all species.

We also created a coalescent version of the BUSCO pipeline to match closely with the ROADIES pipeline, since both are coalescent-based approaches and support missing data (single-copy complete genes, which may not be present in all 48 input genomes). This pipeline consists of five stages. First, we ran BUSCO (version 5.7.1) (25) with the same parameters as the previous pipeline. Next, a custom script extracted all single-copy complete genes that are at least present in one of the 48 species, retrieving 8,328 out of 8,338 genes from the Aves OrthoDB database. We then performed multiple sequence alignment and filtering of all extracted genes using MAFFT (version 7.256) (120) and TrimAl (version 1.5.0) (121), similar to the previous concatenation-based pipeline. In the final stage, we ran ParGenes (version 1.2.0) (123) to generate the gene trees and estimate the final species tree (using --use-astral command). This pipeline closely mirrors ROADIES's coalescent-based approach, except the fact that ROADIES considers multicopy genes and does not require any input gene annotations.

Data, Materials, and Software Availability. The source code of ROADIES is freely available under the MIT License on GitHub (124), and the documentation for ROADIES is available at ref. 125. The details of the input datasets used in the manuscript are listed in Datasets S1–S8. All inferred gene trees and species trees are deposited to Dryad (126).

ACKNOWLEDGMENTS. We thank Tian (Kevin) Liu for his valuable contributions to the ROADIES codebase. We thank Hiram Clawson, Guojie Zhang, Yulong Xie, and Benedict Paten for their helpful feedback. Research reported in this publication was supported by an Amazon Research Award (Fall 2022 Call for Proposals), NIH grant 1R35GM142725 (to S.M.), and funding from the Hellman Fellowship (to Y.T.).

Author affiliations: [a]Department of Computer Science and Engineering, University of California, San Diego, CA 92093; and [b]Department of Electrical and Computer Engineering, University of California, San Diego, CA 92093

1. S. Cheng *et al.*, 10KP: A phylodiverse genome sequencing plan. *GigaScience* **7**, giy013 (2018).
2. Genome 10K Community of Scientists, Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. *J. Hered.* **100**, 659–674 (2009).
3. A. Rhie *et al.*, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
4. N. M. Foley *et al.*, A genomic timescale for placental mammal evolution. *Science* **380**, eabl8189 (2023).
5. J. Stiller *et al.*, Complexity of avian evolution revealed by family-level genomes. *Nature* **629**, 851–860 (2024), 10.1038/s41586-024-07323-1.
6. F. Delsuc, H. Brinkmann, H. Philippe, Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**, 361–375 (2005).
7. S. D. Smith, M. W. Pennell, C. W. Dunn, S. V. Edwards, Phylogenetics is the new genetics (for most of biodiversity). *Trends Ecol. Evol.* **35**, 415–425 (2020).
8. M. J. Christmas *et al.*, Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
9. X.-X. Shen *et al.*, Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545.e20 (2018).
10. A. Marcovitz, R. Jia, G. Bejerano, "Reverse genomics" predicts function of human conserved noncoding elements. *Mol. Biol. Evol.* **33**, 1358–1369 (2016).
11. Y. Turakhia, H. I. Chen, A. Marcovitz, G. Bejerano, A fully-automated method discovers loss of mouse-lethal and human-monogenic disease genes in 58 mammals. *Nucleic Acids Res.* **48**, e91 (2020).
12. A. Hobolth, J. Dutheil, J. Hawks, M. Schierup, T. Mailund, Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* **21**, 349–56 (2011).
13. M. L. Smith, M. W. Hahn, New approaches for inferring phylogenies in the presence of paralogs. *Trends Genet.* **37**, 174–187 (2021).
14. S. Mirarab, L. Nakhleh, T. Warnow, Multispecies coalescent: Theory and applications in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **52**, 247–268 (2021).
15. "Species tree inference: A guide to methods and applications" in *Species Tree Inference*, L. S. Kubatkol, L. Knowles, Eds. (Princeton University Press, 2023), pp. 15–144.
16. J. K. Schull, Y. Turakhia, J. A. Hemker, W. J. Dally, G. Bejerano, Champagne: Automated whole-genome phylogenomic character matrix method using large genomic indels for homoplasy-free inference. *Genome Biol. Evol.* **14**, evac013 (2022).
17. L. Chen *et al.*, Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* **364**, eaav6202 (2019).
18. E. D. Jarvis *et al.*, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
19. S. Lien *et al.*, The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
20. D. Copetti *et al.*, Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12003–12008 (2017).
21. S. Song, L. Liu, S. V. Edwards, S. Wu, Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14942–14947 (2012).
22. P. Kapli, Z. Yang, M. J. Telford, Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444 (2020).
23. M. S. Springer, J. Gatesy, On the importance of homology in the age of phylogenomics. *Syst. Biodivers.* **16**, 210–228 (2018).
24. J. Armstrong *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
25. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
26. B. C. Faircloth, PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinforma. Oxf. Engl.* **32**, 786–788 (2016).
27. L. S. Katz *et al.*, Mashtree: A rapid comparison of whole genome sequence files. *J. Open Source Softw.* **4**, 1762 (2019).
28. D. Dylus, A. Altenhoff, S. Majidian, F. J. Sedlazeck, C. Dessimoz, Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat. Biotechnol.* **42**, 139–147 (2023), 10.1038/s41587-023-01753-4.
29. R. S. Schwartz, K. M. Harkins, A. C. Stone, R. A. Cartwright, A composite genome approach to identify phylogenetically informative data from next-generation sequencing. *BMC Bioinformatics* **16**, 193 (2015).
30. R. Literman, R. Schwartz, Genome-scale profiling reveals noncoding loci carry higher proportions of concordant data. *Mol. Biol. Evol.* **38**, 2306–2318 (2021).
31. D. Kuznetsov *et al.*, OrthoDB v11: Annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
32. J. A. Rick, C. D. Brock, A. L. Lewanski, J. Golcher-Benavides, C. E. Wagner, Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. *Syst. Biol.* **73**, 76–101 (2024).
33. A. Prasad, E. D. Lorenzen, M. V. Westbury, Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Mol. Ecol. Resour.* **22**, 45–55 (2022).
34. J. J. Doyle, Gene trees and species trees: Molecular systematics as one-character taxonomy. *Syst. Bot.* **17**, 144–163 (1992).
35. M. S. Springer, J. Gatesy, The gene tree delusion. *Mol. Phylogenet. Evol.* **94**, 1–33 (2016).
36. H. Philippe *et al.*, Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.*, 283 (2017), 10.5852/ejt.2017.283.
37. S. E. Brenner, Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
38. I. T. Fiddes *et al.*, Comparative annotation toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018).
39. S. König, L. W. Romoth, L. Gerischer, M. Stanke, Simultaneous gene finding in multiple genomes. *Bioinformatics* **32**, 3388–3395 (2016).
40. B. Y. Kim *et al.*, Highly contiguous assemblies of 101 drosophilid genomes. *Elife* **10**, e66405 (2021).
41. P.-F. Ma *et al.*, Genome assemblies of 11 bamboo species highlight diversification induced by dynamic subgenome dominance. *Nat. Genet.* **56**, 710–720 (2024).
42. S. Feng *et al.*, Dense sampling of bird diversity increases power of comparative genomics. *Nature* **587**, 252–257 (2020).

43. L. F. K. Kuderna *et al.*, Identification of constrained sequence elements across 239 primate genomes. *Nature* **625**, 735–742 (2024).
44. L. Shang *et al.*, A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
45. Y. Wu *et al.*, Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).
46. P. Natsidis, P. Kapli, P. H. Schiffer, M. J. Telford, Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience* **24**, 102110 (2021).
47. R. Fernández, T. Gabaldon, C. Dessimoz, "Orthology: Definitions, prediction, and impact on species phylogeny inference" in *Phylogenetics in the Genomic Era*, C. Scornavacca, F. Delsuc, N. Galtier, Eds. (2020), pp. 2.4:1–2.4:14, No commercial publisher|Authors open access book.
48. D. M. Emms, S. Kelly, STAG: Species tree inference from all genes. bioXriv [Preprint] (2018). https://doi.org/10.1101/267914v1 (Accessed 2 March 2025).
49. C. Zhang, S. Mirarab, ASTRAL-Pro 2: Ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**, 4949–4950 (2022).
50. R. Parsons, M. S. Bansal, DupLoss-2: Improved phylogenomic species tree inference under gene duplication and loss. bioXriv [Preprint] (2024). https://doi.org/10.1101/2024.09.05.611565v1 (Accessed 8 March 2025).
51. M. H. Bessa, M. S. Gottschalk, L. J. Robe, Whole genome phylogenomics helps to resolve the phylogenetic position of the Zygothrica genus group (Diptera, Drosophilidae) and the causes of previous incongruences. *Mol. Phylogenet. Evol.* **199**, 108158 (2024).
52. B. C. Genevcius, Dissecting the Pandora's box: Preliminary phylogenomic insights into the internal and external relationships of stink bugs (Hemiptera: Pentatomidae). *Insect Syst. Divers.* **8**, 10 (2024).
53. F. Cedrola *et al.*, Phylogenomics corroborates morphology: New discussions on the systematics of Trichostomatia (Ciliophora, Litostomatea). *Eur. J. Protistol.* **95**, 126093 (2024).
54. S. Reddy *et al.*, Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Syst. Biol.* **66**, 857–879 (2017).
55. E. L. Braun, R. T. Kimball, Data types and the phylogeny of neoaves. *Birds* **2**, 1–22 (2021).
56. O. Jeffroy, H. Brinkmann, F. Delsuc, H. Philippe, Phylogenomics: The beginning of incongruence? *Trends Genet.* **22**, 225–231 (2006).
57. R. S. Harris, "Improved pairwise alignment of genomic DNA," doctoral dissertation, The Pennsylvania State University, University Park, PA (2007).
58. S. Mirarab *et al.*, PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* **22**, 377–386 (2015).
59. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
60. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2–Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
61. Y. Tabatabaee, C. Zhang, S. Arasti, S. Mirarab, Species tree branch length estimation despite incomplete lineage sorting, duplication, and loss. bioXriv [Preprint] (2025). https://doi.org/10.1101/2025.02.20.639320v1 (Accessed 7 March 2025).
62. E. Sayyari, S. Mirarab, Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
63. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
64. D. Thybert *et al.*, Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. *Genome Res.* **28**, 448–459 (2018), 10.1101/gr.234096.117.
65. C. I. Wu, W. H. Li, Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1741–1745 (1985).
66. D. E. Wildman *et al.*, Genomics, biogeography, and the diversification of placental mammals. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14395–14400 (2007).
67. W. J. Murphy, T. H. Pringle, T. A. Crider, M. S. Springer, W. Miller, Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* **17**, 413–421 (2007).
68. M. dos Reis *et al.*, Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. Biol. Sci.* **279**, 3491–3500 (2012).
69. C. C. Morgan *et al.*, Heterogeneous models place the root of the placental mammal phylogeny. *Mol. Biol. Evol.* **30**, 2145–2156 (2013).
70. J. E. Tarver *et al.*, The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.* **8**, 330–344 (2016).
71. J. A. Esselstyn, C. H. Oliveros, M. T. Swanson, B. C. Faircloth, Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. *Genome Biol. Evol.* **9**, 2308–2321 (2017).
72. T. J. D. Halliday, P. Upchurch, A. Goswami, Resolving the relationships of Paleocene placental mammals. *Biol. Rev.* **92**, 521–550 (2017).
73. M. A. O'Leary *et al.*, The placental mammal ancestor and the post–K-Pg radiation of placentals. *Science* **339**, 662–667 (2013).
74. J. Romiguier, V. Ranwez, F. Delsuc, N. Galtier, E. J. P. Douzery, Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.* **30**, 2134–2144 (2013).
75. R. W. Meredith *et al.*, Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521–524 (2011).
76. V. C. Mason *et al.*, Genomic analysis reveals hidden biodiversity within colugos, the sister group to primates. *Sci. Adv.* **2**, e1600633 (2016).
77. C. Zhang, R. Nielsen, S. Mirarab, CASTER: Direct species tree inference from whole-genome alignments. *Science* **387**, eadk9688 (2025).
78. P. Pamilo, M. Nei, Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583 (1988).
79. M.-Y. Chen, D. Liang, P. Zhang, Phylogenomic resolution of the phylogeny of Laurasiatherian mammals: Exploring phylogenetic signals within coding and noncoding sequences. *Genome Biol. Evol.* **9**, 1998–2012 (2017).
80. H. Nishihara, M. Hasegawa, N. Okada, Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9929–9934 (2006).
81. N. S. Upham, J. A. Esselstyn, W. Jetz, Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.* **17**, e3000494 (2019).
82. D. Genereux *et al.*, A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).
83. Y. Du, S. Wu, S. V. Edwards, L. Liu, The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. *BMC Evol. Biol.* **19**, 203 (2019).
84. L. Liu *et al.*, Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7282–E7290 (2017).
85. R. O. Prum *et al.*, A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569–573 (2015).
86. H. Kuhl *et al.*, An unbiased molecular approach using 3′-UTRs resolves the avian family-level tree of life. *Mol. Biol. Evol.* **38**, 108–127 (2021).
87. P. M. O'Grady, R. DeSalle, Phylogeny of the Genus Drosophila. *Genetics* **209**, 1–25 (2018).
88. A. Yassin, Phylogenetic classification of the Drosophilidae Rondani (Diptera): The role of morphology in the postgenomic era. *Syst. Entomol.* **38**, 349–364 (2013).
89. C. A. M. Russo, B. Mello, A. Frazão, C. M. Voloch, Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae). *Zool. J. Linn. Soc.* **169**, 765–775 (2013).
90. A. Suh, The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. *Zool. Scr.* **45**, 50–62 (2016).
91. E. L. Braun, J. Cracraft, P. Houde, "Resolving the Avian tree of life from top to bottom: The promise and potential boundaries of the phylogenomic era" in *Avian Genomics in Ecology and Evolution: From the Lab into the Wild*, R. H. S. Kraus, Ed. (Springer International Publishing, 2019), pp. 151–210.
92. R. T. Kimball *et al.*, A phylogenomic supertree of birds. *Diversity* **11**, 109 (2019).
93. A. Cloutier *et al.*, Whole-genome analyses resolve the phylogeny of flightless birds (palaeognathae) in the presence of an empirical anomaly zone. *Syst. Biol.* **68**, 937–955 (2019).
94. T. B. Sackton *et al.*, Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**, 74–78 (2019).
95. T. Yuri *et al.*, Parsimony and model-based analyses of indels in avian nuclear genes reveal congruent and incongruent phylogenetic signals. *Biology* **2**, 419 (2013).
96. S. Mirarab *et al.*, A region of suppressed recombination misleads neoavian phylogenomics. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2319506121 (2024).
97. X.-X. Shen *et al.*, Reconstructing the backbone of the saccharomycotina yeast phylogeny using genome-scale data. *G3 (Bethesda)* **6**, 3927–3939 (2016).
98. C. T. Hittinger *et al.*, Genomics and the making of yeast biodiversity. *Curr. Opin. Genet. Dev.* **35**, 100–109 (2015).
99. R. Riley *et al.*, Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9882–9887 (2016).
100. M. Groenewald *et al.*, A genome-informed higher rank classification of the biotechnologically important fungal subphylum Saccharomycotina. *Stud. Mycol.* **105**, 1–22 (2023).
101. B. Oxelman *et al.*, Phylogenetics of allopolyploids. *Annu. Rev. Ecol. Evol. Syst.* **48**, 543–557 (2017).
102. J. K. Triplett, L. G. Clark, A. E. Fisher, J. Wen, Independent allopolyploidization events preceded speciation in the temperate and tropical woody bamboos. *New Phytol.* **204**, 66–73 (2014).
103. S. A. Kelchner, Higher level phylogenetic relationships within the bamboos (Poaceae: Bambusoideae) based on five plastid markers. *Mol. Phylogenet. Evol.* **67**, 404–413 (2013).
104. S. Koren *et al.*, Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
105. Z.-H. Guo *et al.*, Genome sequences provide insights into the reticulate origin and unique traits of woody bamboos. *Mol. Plant* **12**, 1353–1365 (2019).
106. B. D. Ondov *et al.*, Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
107. K. Howe, A. Bateman, R. Durbin, QuickTree: Building huge neighbour-joining trees of protein sequences. *Bioinformatics* **18**, 1546–1547 (2002).
108. A. M. Altenhoff *et al.*, OMA orthology in 2024: Improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Res.* **52**, D513–D521 (2024).
109. F. Mölder *et al.*, Sustainable data analysis with Snakemake. *F1000Res.* **10** (2021), 10.12688/f1000research.29032.2.
110. Y. Mao, G. Zhang, A complete, telomere-to-telomere human genome sequence presents new opportunities for evolutionary genomics. *Nat. Methods* **19**, 635–638 (2022).
111. J. E. McCormack *et al.*, Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* **22**, 746–754 (2012).
112. L. Doronina, O. Reising, H. Clawson, D. A. Ray, J. Schmitz, True homoplasy of retrotransposon insertions in primates. *Syst. Biol.* **68**, 482–493 (2019).
113. G. Churakov *et al.*, Rodent evolution: Back to the root. *Mol. Biol. Evol.* **27**, 1315–1326 (2010).
114. B. M. Kirilenko *et al.*, Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107 (2023), 10.1126/science.abn3107.
115. D. L. Ayres *et al.*, BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).
116. S. D. Goenka, Y. Turakhia, B. Paten, M. Horowitz, "SegAlign: A scalable GPU-based whole genome aligner" in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, (2020), pp. 1–13.
117. M. Balaban *et al.*, Generation of accurate, expandable phylogenomic trees with uDance. *Nat. Biotechnol.* **42**, 768–777 (2023), 10.1038/s41587-023-01868-8.
118. I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024), 10.1093/nar/gkae268.
119. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
120. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
121. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
122. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
123. B. Morel, A. M. Kozlov, A. Stamatakis, ParGenes: A tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* **35**, 1771–1773 (2019).
124. A. Gupta, S. Mirarab, Y. Turakhia, ROADIES: Reference-free Orthology-free Annotation-free Discordance-aware Estimation of Species trees. GitHub. https://github.com/TurakhiaLab/ROADIES. Accessed 10 April 2025.
125. A. Gupta, S. Mirarab, Y. Turakhia, ROADIES Documentation. Turakhia Lab, UC San Diego. https://turakhia.ucsd.edu/ROADIES/. Accessed 10 April 2025.
126. A. Gupta, S. Mirarab, Y. Turakhia, Accurate, scalable, and fully automated inference of species trees from raw genome assemblies using ROADIES [Dataset]. Dryad. https://doi.org/10.5061/dryad.tht76hf73. Accessed 10 January 2025.