Molecular Biology and Evolution, 2026, 43, 1–11
https://doi.org/10.1093/molbev/msag073
Published: 23 March 2026
Methods

MBE OXFORD

WASTER: Practical *de novo* Phylogenomics from Low-coverage Short Reads

Chao Zhang [1,2,3,4]*, **Rasmus Nielsen** [3,4,5]

[1]State Key Laboratory of Gene Function and Modulation Research, School of Life Sciences, Peking University, Beijing 100871, China
[2]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[3]Globe Institute, University of Copenhagen, Øster Voldgade 5-7, Copenhagen 1350, Denmark
[4]Department of Integrative Biology, University of California Berkeley, 110 Sproul Hall, Berkeley, CA 94704, USA
[5]Department of Statistics, University of California Berkeley, 110 Sproul Hall, Berkeley, CA 94704, USA
***Corresponding author**: E-mail: chaozhang@pku.edu.cn.
Associate editor: Tal Pupko

Abstract

The advent of affordable whole-genome sequencing has spurred numerous large-scale projects aimed at inferring the tree of life, yet achieving a complete species-level phylogeny remains a distant goal due to significant costs and computational demands. Traditional species tree inference methods, though effective, are hampered by the need for high-coverage sequencing, high-quality genomic alignments, and extensive computational resources. To address these challenges, this study introduces WASTER, a novel *de novo* tool for inferring shallow phylogenies directly from short-read sequences. WASTER employs a k-mer based approach for identifying variable sites, circumventing the need for genome assembly and alignment. Using simulations, we demonstrate that WASTER achieves accuracy comparable to that of traditional alignment-based methods, even for low sequencing depth, and has substantially higher accuracy than other alignment-free methods. We validate WASTER's efficacy on real data, where it accurately reconstructs phylogenies of eukaryotic species with as low depth as 1.5X. WASTER provides a fast and efficient solution for phylogeny estimation in cases where genome assembly and/or alignment may bias analyses or is challenging, for example due to low sequencing depth. It also provides a method for generating guide trees for tree-based alignment algorithms. WASTER's ability to accurately estimate shallow phylogenies from low-coverage sequencing data without relying on assembly and alignment will lead to substantially reduced sequencing and computational costs in phylogenomic projects.

Keywords alignment free, assembly free, low coverage reads, phylogenomics, shallow phylogenies

Introduction

Whole-genome sequencing has recently replaced single marker sequencing as the dominant method for estimating phylogenies, and has been the focus of many recent large-scale efforts for a broad set of species (Lewin et al. 2018; Haussler et al. 2009; Zhang 2015; Fan et al. 2020; Evans et al. 2013; GIGA Community of Scientists 2014; Cheng et al. 2018; Grigoriev et al. 2014; Foley et al. 2023; Stiller et al. 2024). Nevertheless, these efforts are still short of realizing complete species-level phylogenies for most clades. Major obstacles include the costs of high-coverage sequencing, extensive human effort in running species tree inference pipelines, and substantial computational resource demanded (Kapli et al. 2020). In addition, alignment challenges may possibly affect the results in downstream phylogenetic analyses (Liu et al. 2017).

Traditional species tree inference pipelines typically involve several key steps: assembling reads into genomes, aligning orthologous genomic sequences, optionally reconstructing gene trees from aligned genomes, and inferring the species tree from gene trees or directly from the aligned genomes (Fig. 1a). The recently developed CASTER improves the efficiency of species tree inference in the presence of incomplete lineage sorting (ILS) by enabling direct inference from aligned genomes (Zhang et al. 2025). CASTER identifies the optimal tree topology by maximizing the weighted sum of quartet site pattern counts across all genomic sites and species quartets. It assigns negative weights to specific quartet site patterns to counterbalance the effects of "long-branch attraction" and zero weights to invariant sites which contribute no phylogenetic signal. While CASTER helps alleviate the computational bottleneck from alignment to species tree, it still depends on genome assembly and genome

Received: June 10, 2025. **Revised:** February 14, 2026. **Accepted:** February 18, 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

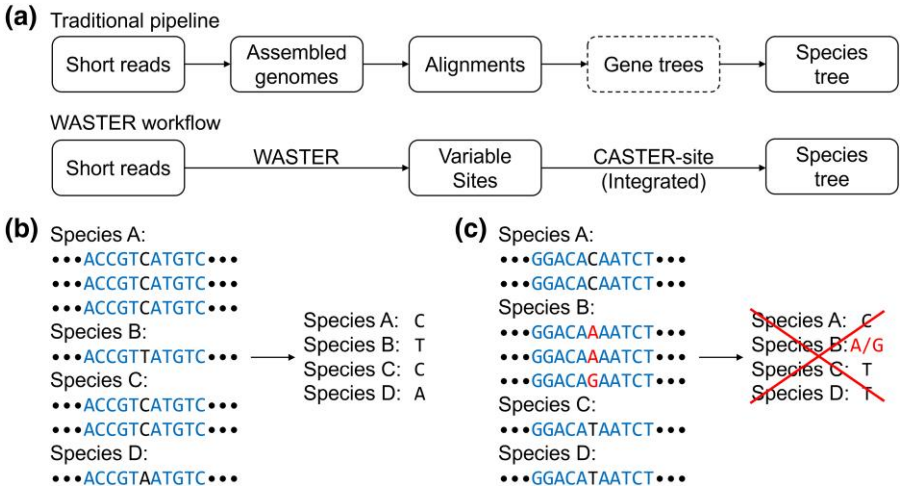

Figure 1 WASTER workflow. a) Overview of the WASTER workflow compared to the traditional species tree inference pipeline. b) WASTER indentifies variable sites by detecting k-mers that differ only at the central base. c) If two variants of a k-mer are found within the same species, WASTER removes the entire aligned site.

alignment, which become new bottlenecks. Additionally, the necessity for high coverage sequencing persists, with traditional pipelines usually requiring coverage above 30X for reliable genome assemblies, imposing financial challenges (Genereux et al. 2020; Feng et al. 2020). Additionally, Progressive Cactus, the leading tool for aligning a large number of genomes, relies on a user-specified guide-tree (Armstrong et al. 2020). The use of a guide-tree for alignment may potentially lead to circular inferences when subsequently estimating the tree using the very same alignment. These factors underscore the need for accurate alignment-free and assembly-free species tree inference methods.

Several assembly-free methods have been proposed (Ondov et al. 2016; Jain et al. 2018; Sarmashghi et al. 2019; Balaban et al. 2020; Dylus et al. 2024; Schwartz et al. 2015), but few are suitable for inferring trees from reads covering entire eukaryotic nuclear genomes. Some methods rely on mapping reads to reference sequences, thus requiring alignments of annotated orthologous sequences as input (Dylus et al. 2024). This limits their applicability, especially in the absence of closely related reference genomes. Other methods construct a "pangenome graph" encompassing all taxa, but this strategy does not scale well to genome-scale data (Schwartz et al. 2015). Alternatively, several k-mer-based methods compute pairwise evolutionary distances and subsequently invoke distance-based algorithms to infer species trees (Ondov et al. 2016; Jain et al. 2018; Sarmashghi et al. 2019; Balaban et al. 2020). However, these methods struggle to infer accurate phylogenies for challenging trees with short internal branches.

In this article, we introduce WASTER (Without Alignment/Assembly Species Tree EstimatoR), a novel *de novo* tool that infers shallow phylogenies directly from short reads. WASTER operates by initially calling variable sites using k-mers from raw sequencing reads and then employs CASTER-site to infer the species tree from the matrix of variable sites (Fig. 1a). We evaluate the method extensively using simulations and applications to real data and show that it substantially outperforms other alignment-free methods and is as accurate as state-of-the-art alignment-based methods, even when assuming that the genome assembly and alignment are error free.

Results

WASTER: k-mer-based variable site identification

The variable site identification mechanism within WASTER involves a k-mer based approach. Initially, WASTER decomposes input reads or genomes into k-mers of a specified length k, where k is an odd number (19 by default). The process then groups these k-mers by matching all positions except the central one. For a given group of k-mers, if the k-mers from each species are identical, including the central position, then the central position of this k-mer group is designated as one variable site (Fig. 1b). In contrast, if any two k-mers from the same species differ, the entire group is disregarded, reducing the possibility of paralogy, alignment errors, or sequencing errors. This approach, however, results in the exclusion of heterozygous sites (Fig. 1c). Following the variable site identification, the concatenated variable sites are utilized as input for CASTER-site, which undertakes the inference of the species tree. Due to WASTER's inability to distinguish between the forward and reverse strands, it employs the T92 substitution model (Tamura 1992), a simplified model that does not account for strand bias.

Evaluation in simulations

SR201 simulated alignments

We first evaluated the performance of WASTER versus alignment-based methods using the SR201 dataset which comprises simulated alignments from 201 species with the heights of species trees similar to that observed in intrafamilial or intraordinal trees in birds and mammals (Zhang et al. 2025). This dataset

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

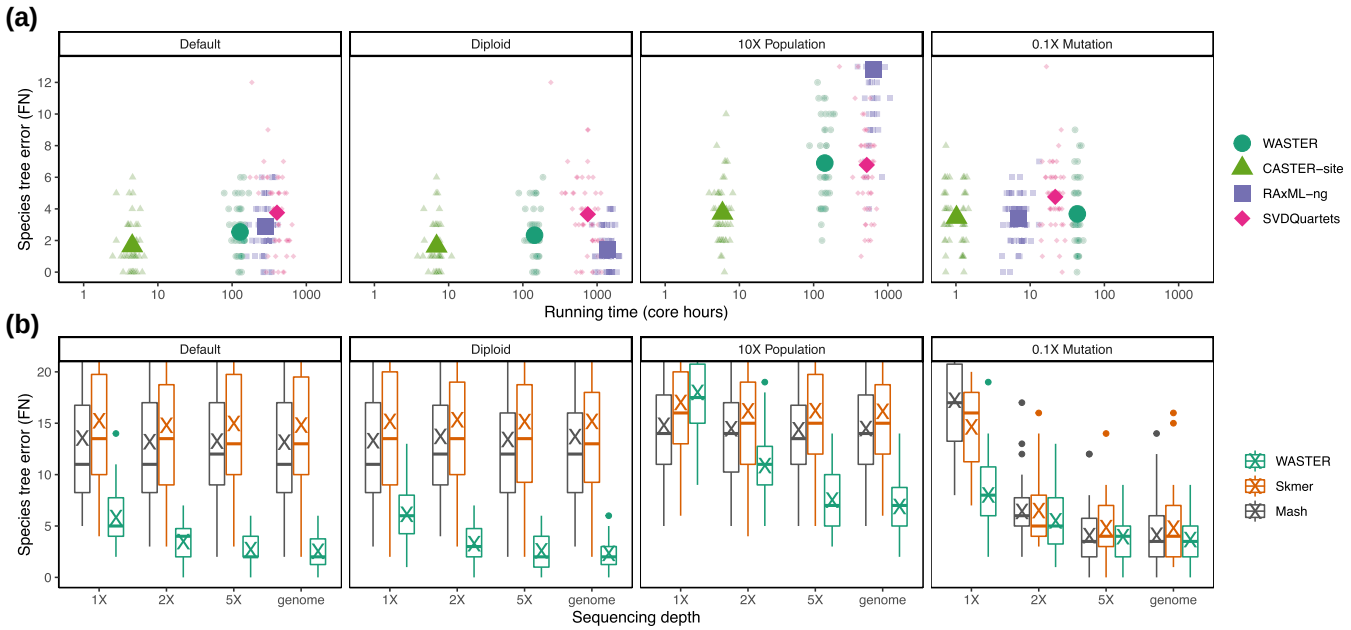

Figure 2 Benchmarking WASTER via SR201 simulations. All simulations are performed on the 201-species phylogeny described in (Zhang et al. 2025). a) The number of branches differing between the inferred and true species trees (bipartition false negatives) versus running time (log scale) for WASTER and three alignment-based species tree inference methods using aligned genomes under various settings. Small symbols represent individual replicates, while large, opaque symbols denote mean values across 50 replicates. b) The number of branches differing between inferred and true species trees (FN) versus sequencing depth for WASTER, Mash, and Skmer on simulated short reads under different settings. Inferred species trees using true genomes are included as control groups. Under 0.1X mutation rate setting, Skmer fails to output phylogenies on 36/24/25/21 replicates for 1X/2X/5X/genome cases, respectively.

incorporates many of the intricacies of molecular evolution. To avoid unrealistic strict molecular clocks, this dataset allows variation in generation time and population size across different branches of the species tree. Simulations were conducted under the Hudson model (Hudson 1983) for recombination and the GTR substitution model (Tavare 1986), with parameters adjusted for varying mutation rates across the genome. Notably, the equilibrium nucleotide frequencies differ from the no strand-bias model assumed by WASTER. We juxtaposed WASTER with several leading alignment-based methods:

1. RAxML-ng (Kozlov et al. 2019), a representative of maximum-likelihood methods which are widely used for reconstructing concatenated sequence tree yet statistically inconsistent (Roch and Steel 2015),
2. SVDQuartets (Chifman and Kubatko 2014), an invariant-based method maintaining statistical consistency under ILS,
3. CASTER-site (Zhang et al. 2025).

Although providing true simulated alignments as inputs, rather than assembled genomes from simulated reads, disproportionately favored alignment-based methods, WASTER still demonstrated accuracy on par with them. Across all replicates under all experimental conditions, WASTER's species tree branch estimation error as measured by bipartition false negatives (FN) averaged at 3.9 branches, significantly lower ($P < 10^{-5}$) than RAxML-ng (5.1 branches) and SVDQuartets (4.7 branches), but higher ($P < 10^{-15}$) than CASTER-site (2.6 branches). Moreover, WASTER's total runtime was comparable to the time required for species tree estimation from alignments alone in traditional workflows, being 5.1 times faster than RAxML-ng and 3.7 times faster than SVDQuartets (Fig. 2a). WASTER uses the integrated CASTER-site to infer species trees directly from sequence data. To evaluate the benefit of using CASTER-site over other methods in this context, we also tested an alternative approach in which variable sites extracted by WASTER were fed into SVDQuartets for species tree inference. Results show that default WASTER with CASTER-site yields higher accuracy than the combination of WASTER with SVDQuartets, highlighting the advantage of CASTER-site (Fig. S1).

We also explored WASTER's performance under various settings, contrasting it with alignment-based methods under different ploidies, ILS levels, and mutation rates. Replacing haploid genomes with diploid genomes did not significantly impact ($P \approx 0.11$) WASTER's accuracy, mirroring results seen with CASTER-site and SVDQuartets. RAxML-ng, employing a diploid model, improved accuracy at the cost of increased runtime. An increase in effective population size by 10X to model elevated ILS levels notably diminished the accuracy of all methods, including WASTER ($P < 10^{-15}$). Here, RAxML-ng's performance was notably poorer than WASTER and the other methods. To simulate less genetic divergence among species, we reduced the mutation rate to 10%, resulting in a statistically significant decrease in accuracy for all methods ($P < 0.05$), attributed to a reduction in informative sites. Under this setting, the runtime of alignment-based methods improved substantially compared to the default setting. However, due to bottlenecks in the variable site identification step, WASTER's runtime improvement was less pronounced (Fig. 2a).

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

SR201 simulated short reads

We extended our analysis to assess WASTER's performance utilizing Illumina-like short reads, generated from the SR201 dataset alignments using the ART simulator (Huang et al. 2012). This evaluation focused on the impact of varying coverage levels on WASTER's accuracy. Our findings indicated a direct correlation between increased coverage and enhanced accuracy of WASTER across all settings. A notable observation was that at 5X coverage, the species tree error for WASTER, based on short reads, averaged 4.2 branches, deteriorating by less than 10% compared to the error obtained using true simulated genomes.

Furthermore, we contrasted WASTER's accuracy with that of Mash (Ondov et al. 2016) and Skmer (Sarmashghi et al. 2019), two frequently used k-mer-based assembly-free and reference-free pairwise evolutionary distance estimators. The k-mer-based distance methods, commonly employed for phylogenetic reconstruction by providing distance matrices to FastME (Lefort et al. 2015), demonstrated consistent species tree errors of approximately 15 branches across all but 0.1 mutation rate settings, irrespective of coverage level. This is likely attributable to their lack of robustness to high evolutionary distances, where fewer k-mers are shared across species, underscoring the superiority of WASTER.

We also evaluated local bootstrap support on WASTER inferred trees in the same simulated dataset. Aggregating across all experimental conditions, replicates, and species tree branches, 94.4% of correctly inferred branches exhibited bootstrap support of at least 95%, while 94.6% of incorrectly inferred branches fell below this threshold, indicating that a 95% cutoff provides well-calibrated measures of statistical uncertainty. At a more stringent cutoff of 99%, 92.2% of correctly inferred branches achieved support of at least 99%, whereas 98.6% of incorrect branches remained below 99% (Fig. S2).

Quartet simulation

We next reused the simulation procedure described in (Zhang et al. 2025) to generate quartet trees for benchmarking WASTER under challenging conditions involving both high levels of incomplete lineage sorting (ILS) and mutation-rate heterogeneity, which together can induce both "long-branch attraction" and "long-branch repulsion." We simulated 100 replicate alignments, each comprising 20 million base pairs (Mbp), for each of three possible species-tree topologies. Simulations were conducted under the Hudson coalescent model (Hudson 1983) using `msprime` (Baumdicker et al. 2022), with varying tree height ($\lambda = \{1, 0.5, 0.2\}$) and internal branch length ($x = \{0.001, 0.002, 0.005\}$). The branch lengths for the baseline condition ($\lambda = 1, x = 0.001$) were chosen to reflect those observed in a ratite dataset (Sackton et al. 2019) (Fig. S3a).

Benchmarking results on simulated alignments show that CASTER-site consistently outperforms all other methods across all tested conditions (Fig. S4a). RAxML-ng exhibits signs of long-branch repulsion in the Farris-zone topology ($BC \mid AD$), particularly when the internal branch is short (Fig. S4b). Both alignment-free methods display a substantial increase in species tree estimation error under deep phylogenies ($\lambda = 1$). In contrast, under shallow phylogenies ($\lambda \leq 0.5$), WASTER outperforms MASH when internal branches are short ($x \leq 0.002$), although WASTER is more sensitive to shorter alignments (≤ 2 Mbp). In addition to alignments, we also simulated short-read datasets at varying coverage levels ($\{1, 2, 5\}\times$) using ART (Huang et al. 2012) under shallow phylogenies ($\lambda \leq 0.5$). Overall, WASTER achieves higher accuracy with increasing coverage, while Mash performs better at very low coverage (Fig. S4c). Notably, Mash suffers from long-branch attraction in deeper phylogenies when the internal branch is short (Fig. S4d).

GDL simulation

To evaluate the performance of WASTER and Mash on multi-copy gene families, we extended the quartet simulations to include gene duplication and loss (GDL). Simulations were conducted under varying GDL rates, with tree height and internal branch length fixed at $\lambda = 0.2$ and $x = 0.002$, respectively. To enable the emergence of gene superfamilies, duplications and losses were simulated beginning 0.07 substitution units above the root (Fig. S3b). We varied the gene duplication rate across $\{0, 1, 2, 4, 8\}\times$, maintaining a fixed 1:1 ratio between gene duplication and loss rates. For each replicate, we simulated 10,000 gene family trees using SimPhy (Mallo et al. 2016), followed by recombination-free alignments of 100 to 10,000 base pairs (bp) generated with INDELible (Fletcher and Yang 2009). Additionally, we simulated short-read sequencing data at varying coverage levels ($\{1, 1.5, 2, 5\}\times$) using ART (Huang et al. 2012).

Simulation results show that Mash performs better when genomes are short and sequencing coverage is low. In contrast, WASTER outperforms Mash for longer genomes and higher coverage levels. Notably, as genome length increases, the species tree error rate of WASTER approaches 0% regardless of the GDL rate or sequence depth, demonstrating statistical consistency under GDL. In comparison, Mash exhibits a marked increase in error rate as the GDL rate increases. Furthermore, under high GDL rates ($8\times$), Mash displays a counter-intuitive trend where its error rate increases with greater sequencing depth—implying that in practice, higher genomic coverage may not necessarily yield more accurate phylogenies when using Mash. At high GDL levels ($8\times$) and high coverage ($5\times$ or true genome), increasing the number of gene families by 100-fold results in only slight improvement in species tree inference accuracy, suggesting either a lack of statistical consistency or an impractically slow convergence rate (Fig. S5).

Applications in real data

Applications to raw reads

We also compared phylogenies estimated using WASTER, Mash, and Skmer from low-coverage short reads, to phylogenies constructed using traditional methods employing high-coverage reads. Our analyses span multiple taxonomic groups:

1. **Rats:** We evaluated 18 rat species, particularly focusing on the *Niviventer* genus (Ge et al. 2021). The established phylogeny, derived from 14X coverage reads (approximately 39 Gbps), was obtained using MrBayes (Ronquist et al. 2012) (concatenation) and ASTRAL (Zhang et al. 2018), both yielding identical results. We downsampled the reads to 1.5X coverage (3.75 Gbps) for reconstructing phylogenies with WASTER, Mash, and Skmer. The phylogeny

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

reconstructed by WASTER matched the established one, while the phylogenies by Mash and Skmer differed from the established phylogeny by two and five branches, respectively (Fig. 3a and Fig. S6a).

2. **Butterflies:** This study involved 35 butterfly species from the Nymphalidae family (Wang et al. 2022b), with a published phylogeny derived from RAxML on genome-wide variable sites. We standardized the downsampled reads to uniformly 2 Gbps across species, which corresponds to approximately 5X coverage, with the exact coverage level varying according to the genome size of each species. WASTER's reconstruction deviated from the published phylogeny by three branches. Notably, the phylogeny observed within the *Kallima* genus aligns with the phylogeny inferred by CASTER-site, suggesting potential introgression (Zhang et al. 2025). The discrepancies in the deeper splits could potentially be attributed to rapid radiation causing short internal branch lengths associated with high statistical uncertainty (Wang et al. 2022b). Mash and Skmer, however, differed by 8 and 9 branches, respectively (Fig. 3b and Fig. S6b).
3. **Birches:** We analyzed high-coverage RAD sequences (3 Gbps on average) for *Betula utilis* FC and *Betula buggsii* (Wang et al. 2022a) and low-coverage RAD sequences (0.3 Gbps on average) for 20 other birch species (Wang et al. 2021). A reference phylogeny was constructed by mapping RAD sequences to the *Betula pendula* genome assembly and inferring a species tree using ASTRAL from orthologous locus trees; this topology was consistent with the RAxML phylogeny based on the supermatrix. We used this reference tree to evaluate our reconstruction methods. WASTER, Mash, and Skmer all placed *B. lenta* as the sister of *B. chichibuensis*, rather than the sister of *B. michauxii*, likely due to introgression (Fig. S7a,b), which is confirmed using Patterson's D-statistic (Patterson et al. 2012) (Fig. S7c). Beside *B. lenta*, the WASTER inferred phylogeny differed from the published phylogeny by one branch, while Mash and Skmer differed by five and seven branches, respectively (Fig. 3c and Fig. S6c). Noticeably, Mash misgrouped the two species with high coverage as sister taxa, suggesting a potential bias in the method under conditions of non-uniform coverage.
4. **Lizards:** We conducted a similar analysis on 11 Anolis lizard species (Corbett-Detig et al. 2020). The original phylogeny was based on 9X coverage reads. ExaML (Kozlov et al. 2015) (concatenation) and ASTRAL generated consistent phylogenies. Upon downsampling to 1.5X coverage (2.25 Gbps), the reconstructed phylogenies by WASTER and Mash matched the published one, while Skmer's reconstruction deviated by one branch (Fig. S6d).
5. **Mandarin fishes:** Our analysis included three mandarin fish species and one outgroup (He et al. 2020), with *Siniperca chuatsi* sequenced using high-coverage long reads and the others with approximately 40X coverage short reads. The published phylogeny was deduced using RAxML (Stamatakis 2014). Using downsampled short reads with 2X coverage (1.5 Gbps), as well as the assembled *Siniperca chuatsi* genome, all three methods successfully reconstructed the published phylogeny (Fig. S6e).
6. **Mushrooms:** We examined short reads from four species in the *Tricholomopsis* genus and one outgroup (Wang et al. 2023). The established phylogeny, inferred using RAxML on 450 single-copy orthologous genes, was compared with our phylogenies using downsampled reads (1 Gbps). WASTER and Mash successfully reconstructed the published phylogeny, but Skmer did not (Fig. S6f).

These results demonstrate that phylogenies reconstructed using WASTER on low-coverage reads closely align with those derived from traditional pipelines from high-coverage reads, significantly outperforming Mash and, even more so, Skmer in accuracy.

Applications to assembled genomes

In addition to inferring phylogenies of closely related species, WASTER can also generate guide trees for more distantly related species with sufficient accuracy, particularly beneficial when using Progressive Cactus. We demonstrate this with guide trees created by WASTER, Mash, and Skmer for assembled genomes of species in the Primates and Passeriformes orders:

1. **Primates:** Using a dataset of 28 assembled primate genomes (Shao et al. 2023), WASTER accurately recovered the published phylogeny reconstructed by ExaML on aligned genomes. However, the reconstructions by Mash and Skmer deviated by one and two branches, respectively (Fig. 4a and Fig. S8a).
2. **Perching birds:** The dataset of 173 perching bird genomes (Stiller et al. 2024), well-known for their challenging phylogeny, was analyzed. The published phylogeny, reconstructed by ASTRAL+RAxML on aligned genomes, was compared with our reconstructions. WASTER and Mash successfully recovered major clades of Passeriformes (Passerida, Muscicapida, Sylviida, Corvides, and Tyranni), differing from the published phylogeny by 17 and 24 branches, respectively (Fig. 4b). Skmer, however, failed to recover 64 out of 170 branches in the published phylogeny, including misplacing the cinnamon ibon (Hypocryptadius cinnamomeus) among Sylviida, instead of Passerida (Fig. S8b).

Conclusion and discussion

The increasing availability of whole-genome sequencing data has not resulted in a corresponding increase in phylogenetic inference using aligned whole genomes. A major impediment has been the extensive computational resources and significant human effort required, rendering such approaches impractical for many researchers. WASTER addresses this challenge by offering a computationally efficient and automated solution, facilitating broader adoption of whole-genome phylogenetic analyses. Additionally, WASTER's output can reduce the cost of many projects by providing high-quality phylogenies using assembly-free low-coverage sequencing, thereby greatly reducing the cost of phylogenomics. This type of application will enable whole genome approaches to be adopted by a much larger share of the taxonomic community that does not have access to the type of resources necessary to produce high-quality assemblies

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

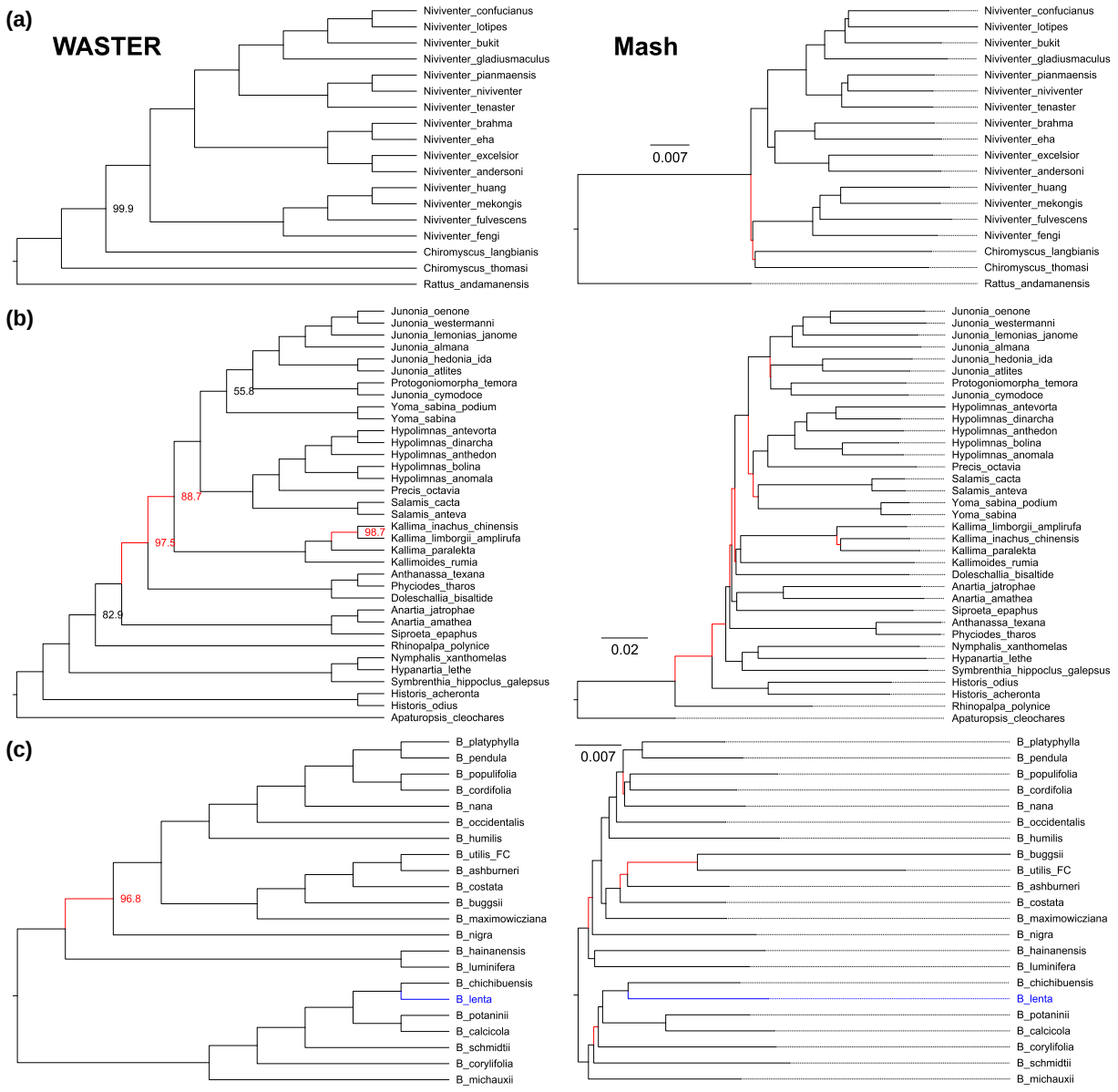

Figure 3 WASTER and Mash on low-coverage short reads. Phylogenies reconstructed using a) 1.5X WGS reads of rat species, b) 2 Gbp butterfly WGS reads, and c) 0.3–3 Gbp birch RAD-seq reads. WASTER branch supports (%) are shown next to the branches, and 100% supports are omitted. Branches that are incongruent with published phylogenies based on high-coverage reads using traditional pipelines are colored red. B. lenta is placed differently from the published phylogeny due to introgression. See also Fig. S6.

for all species of interest. Even when assembly is possible, WASTER can play a crucial role in verifying the accuracy of the phylogenetic inferences that rely on many steps of assembly and alignment subject to potential human errors. Furthermore, WASTER can be used to produce guide trees for progressive alignment.

When designing WASTER, the following decisions were made to achieve faster and more accurate phylogenetic reconstruction, as well as to provide a better user experience for average cases:

1. Variant Site Calling: WASTER operates under the presumption that whole genomes contain overwhelmingly abundant phylogenetic signals, and that capturing an unbiased small subset of these signals is sufficient for accurate phylogenetic reconstruction. However, this presumption may not hold for very small genomes (e.g. bacteria or viruses) or in cases of rapid radiation (e.g. the Neoavian crown group), where signal abundancy may be insufficient.
2. Paralogy handling: Paralogy can introduce substantial noise and significantly dilute phylogenetic signals. Therefore, WASTER excludes an entire variant site whenever it detects two k-mer variants differing only at the central base within a single species. This approach greatly reduces the number of paralogous sites, albeit at the cost of losing some genuine phylogenetic signals. Certain signals that could have been

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

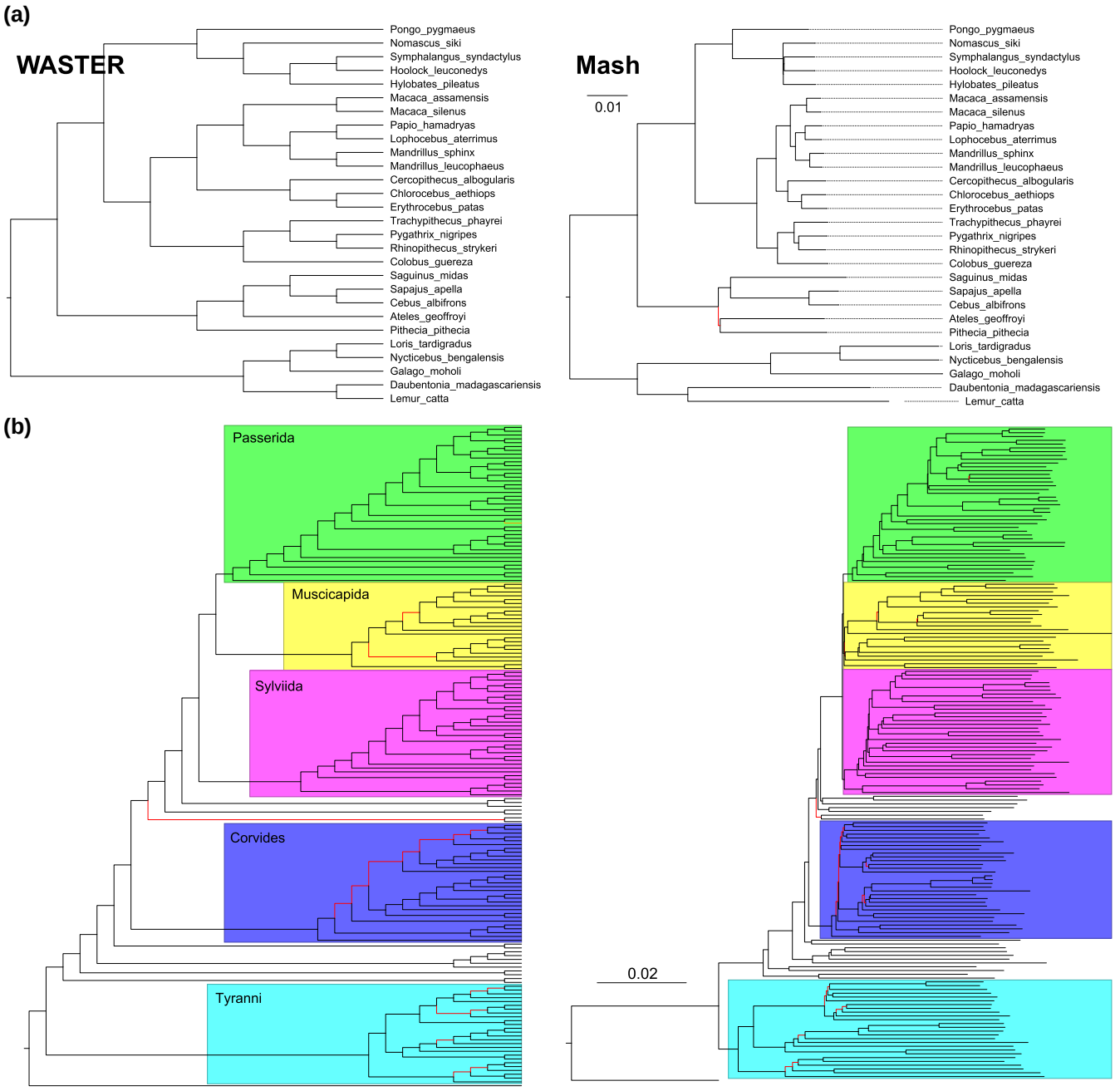

Figure 4 WASTER and Mash on assembled genomes. a) Phylogenies reconstructed using 28 assembled primate genomes. WASTER provides 100% supports for all branches. b) Phylogenies reconstructed from assembled genomes of 173 perching birds, with species names and support values omitted. Branches incongruent with the published phylogenies are highlighted red. See also Fig. S8.

restored are intentionally omitted by design. For instance, while sequencing errors could be mitigated by counting k-mers, WASTER deliberately avoids k-mer counting to save memory.

3. The choice of k-mer size (k): WASTER sets the default k value to 19, ensuring compatibility with machines equipped with 48 GB of memory. Increasing k to 21 or higher would require at least 512 GB of memory, making it impractical for average users. Conversely, setting k to 17 or lower is only effective for genomes significantly smaller than 2 billion base pairs (e.g. fungal genomes). Therefore, we strongly recommend using the default k value for all datasets.

WASTER also presents several avenues for future development and refinement. For example, there exists a promising potential for adapting WASTER to identify possible gene-flow events, including ancient ones, using phylogenetic asymmetries and for inferring phylogenetic networks. Another key area for advancement is to extend WASTER to estimate branch lengths in presence of ILS and, potentially, gene-flow. Finally, integrating WASTER with other

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

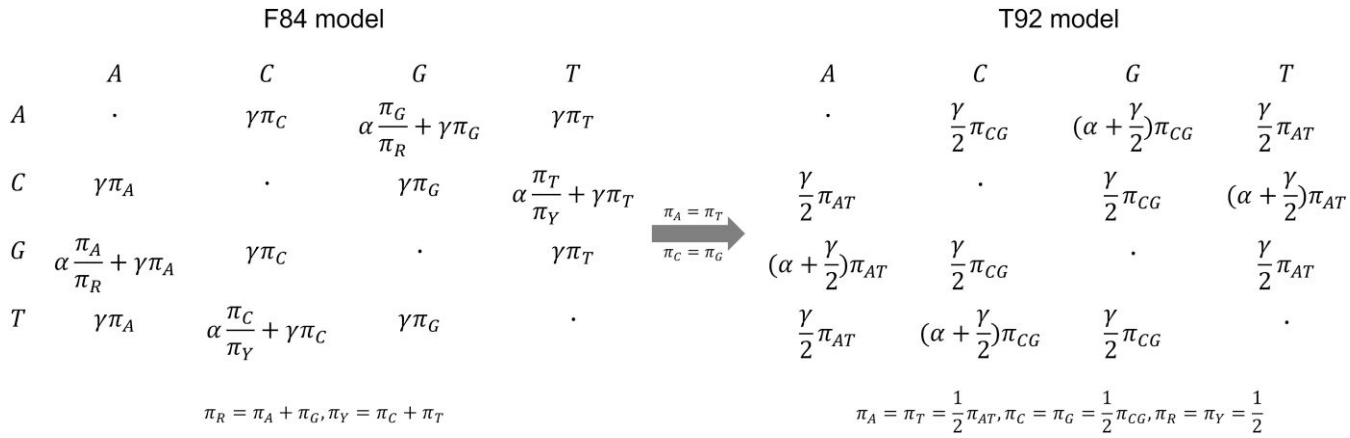

Figure 5 No-strand bias model. By assuming equal equilibrium frequencies of complementary nucleotides, the F84 model reduces to the T92 model.

variable-site-based tools would create a more comprehensive and user-friendly platform, broadening the scope of assembly-free phylogenomic and evolutionary studies.

Despite its broad applicability, WASTER has several limitations:

1. As a k-mer-based method, WASTER is not well suited for reconstructing deep phylogenies, such as intraordinal, family-level insect trees. In such cases, it may produce incorrect or even misleading topologies.
2. Being assembly-free, WASTER assumes genomes lack strand bias and exhibit uniform GC content. In reality, GC content often varies across regions, and strand biases can occur, leading to potential inference errors.
3. Unlike distance-based approaches, WASTER does not provide branch lengths by default. Estimating branch lengths under ILS is already challenging, and ascertainment bias toward conserved regions further complicates this task.

Future work could address these issues by integrating a robust branch length estimation algorithm into WASTER and enhancing its performance on deep phylogenies.

Methods

WASTER

WASTER initially identifies aligned variable sites using Algorithm 1, described below, before leveraging CASTER-site for species tree inference. Acknowledging its limitation in distinguishing between forward and reverse DNA strands, WASTER operates under the assumption of no strand-bias, equating the equilibrium frequencies of A to T and C to G. Consequently, differently from the conventional CASTER-site that assumes the F84 model, WASTER operates under the T92 model. This model can be easily reduced from F84 model by assuming equal equilibrium frequencies of complementary nucleotides (Fig. 5). WASTER computes local bootstrap support for each branch using 1,000 site-based bootstrap replicates. The local support for a branch is defined as the percentage of replicates in which the species tree topology is favored over both of its nearest-neighbor interchange (NNI) alternatives around that branch.

Algorithm 1 WASTER–s algorithm to indentify a set of aligned variable sites S using sets of k-mers $K_1, \ldots, K_n$, where each set represents all k-mers in the input reads/genomes of a species.

Procedure ADDKMER(M, i, s) ▷ M will be modified in this procedure
  $L, c, R \leftarrow$ the first $\frac{k-1}{2}$-mer, the central 1-mer, and the last $\frac{k-1}{2}$-mer of s
  $j \leftarrow$ the $(k-1)$-mer combining L and R
  if $M[i, j] \neq$ "-" and $M[i, j] \neq c$ **then**
    $M[i, j] \leftarrow$ "-" ▷Marked as inconsistent
  else
    $M[i, j] \leftarrow c$
  end if
end procedure
procedure WASTER
  $S \leftarrow$ empty alignment, $M \leftarrow$ empty look-up table ▷a map or a dictionary
  for i in $\{1, \ldots, n\}$ **do**
    for j in all $(k-1)$-mers **do**
      $M[i, j] \leftarrow$ "-" ▷ M is initialized with all gaps
    end for
    for s in K_i **do**
      $L, R \leftarrow$ the first $\frac{k-1}{2}$-mer, and the last $\frac{k-1}{2}$-mer of s
      $\bar{R} \leftarrow$ reverse complement of R
      if $L < \bar{R}$ lexographically **then**
        ADDKMER(M, i, s)
      end if
      if $L > \bar{R}$ lexographically **then**
        ADDKMER(M, i, reverse complement of s)
      end if
    end for
  end for
  for j in all $(k-1)$-mers **do**
    if $M[\cdot, j]$ doesn't contain "N" and $M[\cdot, j]$ is not invariant **then**
      Add $M[\cdot, j]$ as one column of S
    end if
  end for
end procedure

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

Simulating low-coverage reads

Reads with 1X, 2X, and 5X coverage in the simulation studies were simulated using ART Illumina v2.5.8 with the following command:

```
for i in 1 2 5; do
    art_illumina –ss HS25 –sam –i $GENOME \
      –p –l 150 –f $i –m 800 –s 20
      –o ${i}X_READS
      cat ${i}X_READS{1, 2} . fq > \
    ${i}X_READS_COMBINED . fq
  done
```

Each input file contains one FASTA sequence representing a haploid genome or two FASTA sequences representing a diploid genome. Each output file contains reads from both ends.

Species tree inference

We inferred species trees using short reads or whole genomes from simulated and real data with WASTER, Mash, and Skmer. Both Mash and Skmer are distance-based methods and only generate distance matrices. Following the instructions from Skmer's official GitHub repository, we used FastME to infer phylogenies from the distance matrices produced by Skmer. For consistency, we also applied FastME to the distance matrices generated by Mash. WASTER is not distance-based and directly produces phylogenies. Note that since WASTER is incapable of generating distance matrices, the impact of FastME on the comparison of phylogenetic reconstruction accuracy between WASTER and distance-based methods is unavoidable.

Aside from those alignment free methods, we also reused inferred species trees by CASTER-site, RAxML-ng, and SVDQuartets on simulated true alignments from a previous study (Zhang et al. 2025) for comparison (Fig. 2a).

WASTER

The following command was used to infer species trees:

```
waster –C –t 16 –i $INPUT–o $OUTPUT
```

The input file here is a list of species names followed by the corresponding input files:

```
SpeciesName1      Species1.fq
SpeciesName2      Species2.fq
…
```

Mash

The following script was used to infer species trees:

```
mash triangle –s 1000000 $INPUT > mash.tsv
python3 mashFormatFastme . py mash . tsv \
    mash . phy
  fastme –i mash . phy    o $OUTPUT
```

The python script "mashFormatFastme.py" was used to format the output distance matrix by Mash as the input format of FastME:

```
import sys
with open(sys . argv[2] , "w") as f:
    lines = []
    for line in open(sys.argv[1]):
        lines.append(line.split())
    n = int(lines[0][0])
    f . write(lines[0][0] + "\n")
    for i in range(1, n+1):
        f . write("\t" . join([lines[i][0] . split(" . ")[0]]
      + [lines[i][j] for j in range(1, i)] + ["0"]
      + [lines[j][i] for j in range(i + 1, n + 1)])
      + "\n")
```

Skmer

The following script was used to infer species trees:

```
#generating ref_dist_mat . txt below
skmer reference $INPUT –p 16 –t
bash skmerFormatFastme . bash
  ref_dist_mat.txt skmer . phy
fastme –i skmer . phy –o $OUTPUT
```

the python script "skmerFormatFastme.bash" was used to format the output distance matrix by Skmer as the input format of FastME:

```
tail –n + 2 $1 | wc -l > $2
tail –n +2 $1 >> $2
sed –i "s/\t/ __ /g" $2
```


Acknowledgments

We thank Yulong Xie from Zhejiang university for advice on avian phylogenetics.

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

Author contributions

C.Z. developed the idea and implemented the program; C.Z. conceived and conducted the experiments; C.Z. and R.N. analyzed the results and wrote the manuscript.

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

This work was supported by National Natural Science Foundation of China, School of Life Sciences at Peking University, Peking-Tsinghua Center for Life Sciences, VILLUM FONDEN grant 40582, and NIH grant R35 GM153400-01.

Conflicts of interest

The authors declare that there is no conflict of interest.

Data availability

WASTER code is available on Github (github.com/chaoszhang/ASTER). Data is available on FigShare (figshare.com/s/93bcc97f2df8838eda00).

References

Armstrong J *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*. 2020:587:246–251. ISSN 1476-4687. https://doi.org/10.1038/s41586-020-2871-y.

Balaban M, Sarmashghi S, Mirarab S. APPLES: scalable distance-based phylogenetic placement with or without alignments. *Syst Biol*. 2020:69:566–578. ISSN 1063-5157. https://doi.org/10.1093/sysbio/syz063.

Baumdicker F *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2022:220:iyab229. ISSN 1943-2631. https://doi.org/10.1093/genetics/iyab229.

Cheng S *et al.* 10KP: a phylodiverse genome sequencing plan. *Gigascience*. 2018:7:1–9. ISSN 2047217X. https://doi.org/10.1093/gigascience/giy013.

Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. *Bioinformatics*. 2014:30:3317–3324. ISSN 1367-4803. https://academic.oup.com/bioinformatics/article/30/23/3317/206559. https://doi.org/10.1093/bioinformatics/btu530.

Corbett-Detig RB, Russell SL, Nielsen R, Losos J. Phenotypic convergence is not mirrored at the protein level in a lizard adaptive radiation. *Mol Biol Evol*. 2020:37:1604–1614. ISSN 0737-4038. https://academic.oup.com/mbe/article/37/6/1604/5728642. https://doi.org/10.1093/molbev/msaa028.

Dylus D, Altenhoff A, Majidian S, Sedlazeck FJ, Dessimoz C. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat Biotechnol*. 2024:42:139–147. ISSN 1546-1696. https://www.nature.com/articles/s41587-023-01753-4. https://doi.org/10.1038/s41587-023-01753-4.

Evans JD *et al.* The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered*. 2013:104:595–600. ISSN 0022-1503. https://doi.org/10.1093/jhered/est050.

Fan G *et al.* Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *Gigascience*. 2020:9:1–8. ISSN 2047217X. https://doi.org/10.1093/gigascience/giaa080.

Feng S *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature*. 2020:587:252–257. ISSN 1476-4687. https://www.nature.com/articles/s41586-020-2873-9. https://doi.org/10.1038/s41586-020-2873-9.

Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 2009:26:1879–1888. ISSN 0737-4038. https://doi.org/10.1093/molbev/msp098.

Foley NM *et al.* A genomic timescale for placental mammal evolution. *Science*. 2023:380:eabl8189. ISSN 0036-8075. https://doi.org/10.1126/science.abl8189.

Ge D *et al.* Demographic history and genomic response to environmental changes in a rapid radiation of wild rats. *Mol Biol Evol*. 2021:38:1905–1923. ISSN 15371719. https://doi.org/10.1093/molbev/msaa334.

Genereux DP *et al.* A comparative genomics multitool for scientific discovery and conservation. *Nature*. 2020:587:240–245. ISSN 1476-4687. https://doi.org/10.1038/s41586-020-2876-6.

GIGA Community of Scientists. The global invertebrate genomics alliance (GIGA): developing community resources to study diverse invertebrate genomes. *J Hered*. 2014:105:1–18. ISSN 0022-1503. https://doi.org/10.1093/jhered/est084.

Grigoriev IV *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res*. 2014:42:D699–D704. ISSN 0305-1048. https://doi.org/10.1093/nar/gkt1183.

Haussler D *et al.* Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species. *J Hered*. 2009:100:659–674. ISSN 0022-1503. https://doi.org/10.1093/jhered/esp086.

He S *et al.* Mandarin fish (Sinipercidae) genomes provide insights into innate predatory feeding. *Commun Biol*. 2020:3:1–13. ISSN 2399-3642. https://www.nature.com/articles/s42003-020-1094-y. https://doi.org/10.1038/s42003-020-1094-y.

Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012:28:593–594. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btr708.

Hudson RR. Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*. 1983:23:183–201. ISSN 0040-5809. https://doi.org/10.1016/0040-5809(83)90013-8.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018:9:1–8. ISSN 2041-1723. https://www.nature.com/articles/s41467-018-07641-9. https://doi.org/10.1038/s41467-018-07641-9.

Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. *Nat Rev Genet*. 2020:21:428–444. ISSN 1471-0064. https://www.nature.com/articles/s41576-020-0233-0. https://doi.org/10.1038/s41576-020-0233-0.

Kozlov AM, Aberer AJ, Stamatakis A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*.

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026

2015:31:2577–2579. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btv184.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. 2019:35:4453–4455. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btz305.

Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol*. 2015:32:2798–2800. ISSN 0737-4038. https://doi.org/10.1093/molbev/msv150.

Lewin HA *et al.* Earth BioGenome project: sequencing life for the future of life. *Proc Natl Acad Sci U S A*. 2018:115:4325–4333. ISSN 10916490. https://doi.org/10.1073/pnas.1720115115.

Liu L *et al.* Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc Natl Acad Sci U S A*. 2017:114:E7282–E7290. ISSN 10916490. https://doi.org/10.1073/pnas.1616744114.

Mallo D, De Oliveira Martins L, Posada D. SimPhy : phylogenomic simulation of gene, locus, and species trees. *Syst Biol*. 2016:65:334–344. ISSN 1063-5157. https://academic.oup.com/sysbio/article/65/2/334/2427219. https://doi.org/10.1093/sysbio/syv082.

Ondov BD *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016:17:1–14. ISSN 1474760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0997-x. https://doi.org/10.1186/S13059-016-0997-X.

Patterson N *et al.* Ancient admixture in human history. *Genetics*. 2012:192:1065–1093. ISSN 00166731. https://doi.org/10.1534/genetics.112.145037.

Roch S, Steel M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol*. 2015:100:56–62. ISSN 1096-0325. https://pubmed.ncbi.nlm.nih.gov/25545843/. https://doi.org/10.1016/j.tpb.2014.12.005.

Ronquist F *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012:61:539–542. ISSN 1063-5157. https://academic.oup.com/sysbio/article/61/3/539/1674894. https://doi.org/10.1093/sysbio/sys029.

Sackton TB *et al.* Convergent regulatory evolution and loss of flight in paleognathous birds. *Science*. 2019:364:74–78. ISSN 10959203. https://www.science.org/doi/10.1126/science.aat7244. https://doi.org/10.1126/science.aat7244.

Sarmashghi S, Bohmann K, Gilbert PM, Bafna V, Mirarab S. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biol*. 2019:20:34. ISSN 1474760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1632-4. https://doi.org/10.1186/s13059-019-1632-4.

Schwartz RS, Harkins KM, Stone AC, Cartwright RA. A composite genome approach to identify phylogenetically informative data from next-generation sequencing. *BMC Bioinformatics*. 2015:16:1–10. ISSN 14712105. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0632-y. https://doi.org/10.1186/S12859-015-0632-Y.

Shao Y *et al.* Phylogenomic analyses provide insights into primate evolution. *Science*. 2023:380:913–924. ISSN 10959203. https://doi.org/10.1126/science.abn6919.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014:30:1312–1313. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btu033.

Stiller J *et al.* Complexity of avian evolution revealed by family-level genomes. *Nature*. 2024:629:851–860. ISSN 1476-4687. https://www.nature.com/articles/s41586-024-07323-1. https://doi.org/10.1038/s41586-024-07323-1.

Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G +C-content biases. *Mol Biol Evol*. 1992:9:678–687. ISSN 0737-4038. https://doi.org/10.1093/oxfordjournals.molbev.a040752.

Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. In: *Some Mathematical Questions in Biology—DNA Sequence Analysis*. Am Math Soc. Vol. 17; 1986. p. 57–86. ISSN 0075-8523.

Wang G-S *et al.* Phylogenetic and taxonomic updates of Agaricales, with an emphasis on Tricholomopsis. *Mycology*. 2023:15:180–209. ISSN 2150-1203. https://www.tandfonline.com/doi/abs/10.1080/21501203.2023.2263031. https://doi.org/10.1080/21501203.2023.2263031.

Wang L *et al.* Molecular and morphological analyses clarify species delimitation in section Costatae and reveal Betula buggsii sp. nov. (sect. Costatae, Betulaceae) in China. *Ann Bot*. 2022a:129:415–428. ISSN 0305-7364. https://doi.org/10.1093/aob/mcac001.

Wang N, Kelly LJ, McAllister HA, Zohren J, Buggs RJA. Resolving phylogeny and polyploid parentage using genus-wide genome-wide sequence data from birch trees. *Mol Phylogenet Evol*. 2021:160:107126. ISSN 1055-7903. https://www.sciencedirect.com/science/article/pii/S1055790321000592. https://doi.org/10.1016/j.ympev.2021.107126.

Wang S *et al.* The evolution and diversification of oakleaf butterflies. *Cell*. 2022b:185:3138–3152. ISSN 0092-8674. https://doi.org/10.1016/j.cell.2022.06.042.

Zhang C, Nielsen R, Mirarab S. CASTER: direct species tree inference from whole-genome alignments. *Science*. 2025:387:eadk9688. ISSN 0036-8075. https://doi.org/10.1126/science.adk9688.

Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 2018:19:153. ISSN 14712105. https://doi.org/10.1186/s12859-018-2129-y.

Zhang G. Bird sequencing project takes off. *Nature*. 2015:522:34–34. ISSN 1476-4687. https://www.nature.com/articles/522034d. https://doi.org/10.1038/522034d.

Downloaded from https://academic.oup.com/mbe/article/43/4/msag073/8537509 by University of California Library - Berkeley Library user on 14 April 2026