Perspective
(MBE)

Upsampling of Tiny Subsamples Tames Big Data Evolutionary Analysis in Phylogenomics

Sudhir Kumar[1,2,*], Koichiro Tamura[3,4], and Sudip Sharma[1,2]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA

[2]Department of Biology, Temple University, Philadelphia, PA 19122, USA

[3]Department of Biological Sciences, Tokyo Metropolitan University, Tokyo, Japan

[4]Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Tokyo, Japan

**Corresponding author (s.kumar@temple.edu)*

Keywords: phylogenomics, confidence limits, model selection, parameter estimation, hypothesis testing, incomplete lineage sorting

Abstract

Long runtime, high memory demands, and the need for high-performance computing increasingly limit evolutionary analysis of large phylogenomic datasets. We review a scalable framework in which phylogenomic inference is performed on small site subsamples, which are then expanded by upsampling to match the length of the full alignment. In this framework, which involves the analysis of phylogenomic subsamples and upsampling (PSU), each analysis uses only a small number of sites, greatly reducing computation, while upsampling restores the total number of sites and substitutions in the full dataset. Evidence from simulated and empirical datasets shows that PSU can approximate full-data results for bootstrap support, substitution-model selection, likelihood-based hypothesis testing, and estimation of evolutionary parameters and their variances, often with substantial reductions in time and memory, enabling phylogenomics on personal desktops. PSU also provides distributions of clade support across independent subsamples, revealing concordant and conflicting phylogenetic signals that may be obscured in conventional concatenated analyses. Adaptive procedures for choosing the subsample size, the number of subsamples, and the number of upsampling replicates make PSU practical and reproducible. These attributes position PSU as a flexible, accessible, and environmentally favorable strategy for scalable phylogenomic inference.

INTRODUCTION

Advances in sequencing technology and the rapid growth of sequence databases have made it increasingly easier to assemble large collections of sequences spanning thousands of loci across diverse organisms, individuals, and strains (Fig. 1). This expansion has transformed molecular phylogenetics into phylogenomics, which involves analyzing genome-scale alignments to infer evolutionary relationships, estimate divergence times, and explore patterns of molecular evolution (Philippe et al. 2005; Kumar et al. 2012).

Although longer alignments increase statistical power and improve phylogenetic resolution, the computational demands of their analysis grow with the number of sites and substitutions. These demands can become particularly onerous when applying computationally intensive methods, such as Maximum Likelihood (ML) (Fig. 2). For example, selecting the best-fit nucleotide substitution model for an alignment of 394 kilobase pairs from 200 bird species (hereafter, the 200×394 dataset) took over 10 days of computation (258 hours) and used 14 GB of memory (Prum et al. 2015; Sharma and Kumar 2022). Analyses of these and larger phylogenomic datasets cannot be conducted on standard desktop computers with only a few GB of RAM and a few CPU cores. Similar computational challenges arise when ML and other sophisticated methods are used to infer phylogenies, evaluate bootstrap support, and estimate divergence times (Sharma and Kumar 2021; Kumar 2022; Sharma and Kumar 2022).

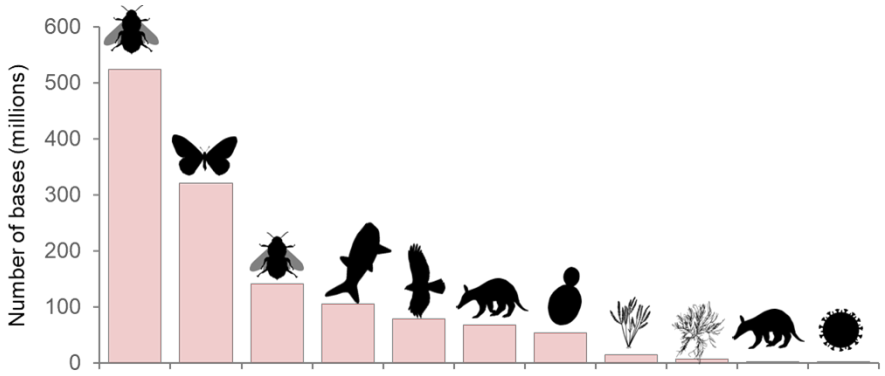

Figure 1. Phylogenomic datasets across the Tree of Life.
Each bar corresponds to a phylogenomic dataset, with the bar height proportional to the number of bases (sites × sequences). Dataset references are in *Supplementary Table S1*.

These computational demands hinder discovery and reduce scientific rigor. Long runtimes discourage robustness checks across models, assumptions, and data treatments, while the need for high-performance computing excludes many laboratories with limited access to advanced hardware. Reproducibility also suffers when reanalyses require days of computation or more memory than standard desktop systems provide. In addition, repeated large-scale analyses carry an environmental cost because runtime directly affects energy use and carbon footprint (Kumar 2022).

Computational innovations are therefore needed to make evolutionary analysis scalable for big-data phylogenomics. One promising strategy is to analyze multiple small subsamples of the alignment and then combine their results. However, such strategies suffer from reduced power because the number of substitutions and sites in a subsample is much lower than in the full

dataset, and it is not straightforward to adjust estimates of bootstrap support and variances to match those from the full-data analysis (Sharma and Kumar 2021; Sharma and Kumar 2022). For these reasons, phylogenomic subsamples are used only to evaluate the stability of evolutionary inferences from concatenated datasets and to explore heterogeneity in phylogenetic signal across partitions and loci (Seo 2008; Faircloth et al. 2012; Song et al. 2012; Edwards 2016; Mongiardino Koch 2021; Lozano-Fernandez 2022).

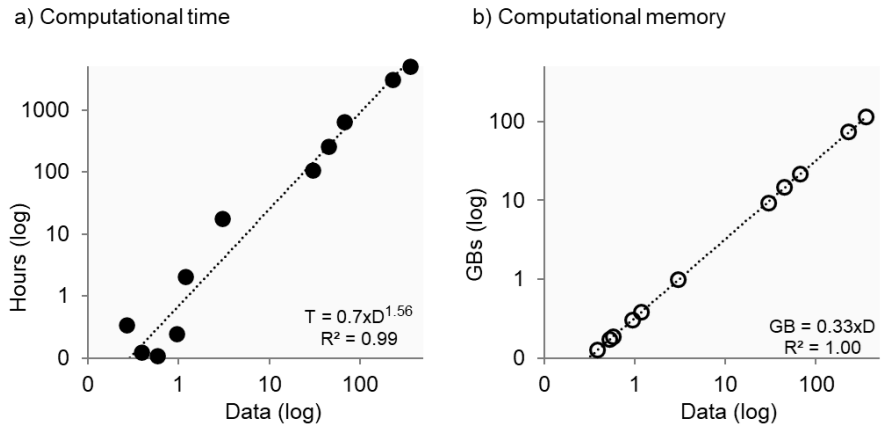

Figure 2. Increasing Computational demands. (**a**) Runtime and (**b**) memory requirements for ML analyses of empirical and simulated datasets. Data size (D) on the x-axis is the product of the number of sequences and the number of unique site configurations, as identical sites are collapsed during ML analysis. Source data are from Table 1 of Sharma and Kumar (2022).

Ideally, subsample analyses will retain the statistical power of the full dataset while reducing computational cost. In this review, we synthesize an emerging framework to address this challenge by upsampling phylogenomic subsamples prior to evolutionary analysis, aiming to approximate the statistical behavior of the full dataset (Kleiner et al. 2014). This idea builds on the statistical foundation of the bag-of-little-bootstraps framework (Kleiner et al. 2014), but extends it to the distinctive needs of phylogenomics, including clade support estimation, model selection, likelihood-based hypothesis testing, and evolutionary parameter estimation. In the sections that follow, we review the logic, performance, applications, and limitations of PSU as a unifying framework for scalable phylogenomic inference.

However, PSU should be viewed as a scalable approximation of full-data inference. This places it alongside many widely used heuristic strategies in phylogenetics, including approximate tree searches and numerical optimization procedures used in maximum-likelihood analysis (Strimmer and Von Haeseler 1996; Price et al. 2010; Stamatakis 2014; Minh et al. 2020; Azouri et al. 2021; Kumar et al. 2023; Kumar et al. 2024). Its value, therefore, rests on empirical performance, convergence behavior, and practical utility across realistic datasets. As reviewed below, PSU targets a different dimension of the computational problem by reducing the number of distinct sites used while preserving full-alignment scale through upsampling.

Estimating confidence in inferred relationships using site subsamples

Felsenstein's bootstrap approach. Felsenstein (1985) adapted Efron's (1979) bootstrap method to assess confidence in clades derived from molecular phylogenetic analysis. In Felsenstein's bootstrap, *R* replicate alignments are generated by resampling sites with replacement from the full dataset of *N* sites. Each replicate dataset also has *N* sites. A phylogeny is inferred from each

replicate dataset, and the proportion of replicates that reconstruct a particular clade is used as its bootstrap support (*FBS*) (Fig. 3a). This analysis evaluates the statistical stability of inferred clades under sampling variation introduced by resampling sites from the full dataset. Felsenstein's bootstrap procedure becomes computationally expensive because hundreds of phylogenetic analyses must be conducted for the bootstrap replicate datasets, each containing ~63.2% of the original sites. This is because the probability that a site is not chosen in a given draw is (1 - 1/*N*), and thus the probability that it is never chosen in all *N* draws is $(1 - 1/N)^N$. Therefore, the probability that the site appears at least once is $1 - (1 - 1/N)^N$, which approaches $1 - e^{-1} \approx 0.632$ as *N* becomes large. Thus, each bootstrap replicate contains approximately 63.2% of the original sites at least once.

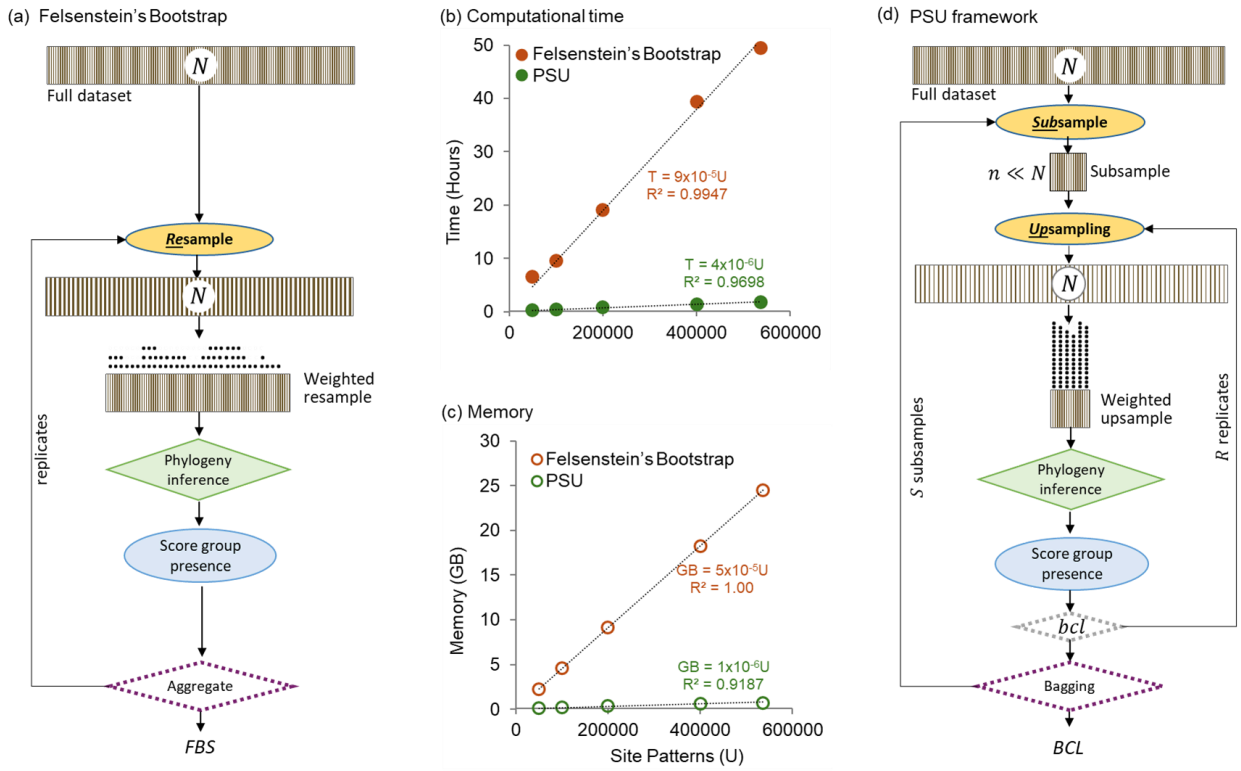

Figure 3. Flowcharts and computational demands of Bootstrap analysis. (**a**) In Felsenstein's bootstrap, replicate datasets are generated by resampling sites from the full dataset (*N* sites), maintaining both dataset size and the site pattern diversity. Felsenstein's bootstrap support (*FBS*) for a clade is the proportion of replicate phylogenies containing the clade of interest. (**b**) runtime and (**c**) memory usage for inferring ML phylogeny for one bootstrap replicate for datasets containing increasing numbers of unique site configurations (*U*) in the full alignment. (**d**) In the PSU framework, small subsamples of randomly selected *n* sites (*n* << *N*) are upsampled to full length (*N* sites). Clade support is estimated for each subsample (*bcl*) and then aggregated across *S* subsamples to obtain the median of subsample *bcl* values (*BCL*), which is used to estimate *FBS*. Panels *b* and *c* show the time and memory required to analyze a PSU replicate dataset. Results are from the analysis of simulated datasets from a previous study (Tamura et al. 2012) containing 446 sequences and alignments ranging from 50,000 to 536,534 bases, as reported in Fig. 1b of Sharma and Kumar (2021).

Since computational time scales linearly with the number of sites in the dataset (Fig. 3b), Felsenstein's bootstrap analysis with just 100 replicates is expected to take more than 63 times longer than a single ML phylogeny inferred from the full dataset. These requirements escalate with more replicates and can become onerous for very long phylogenomic alignments (Fig. 3b). Memory requirements per replicate also increase linearly with sequence length, which can far exceed the memory available on standard desktops (Fig. 3c). These demands motivated the need for alternative strategies that can generate accurate estimates of bootstrap support while greatly reducing computational costs.

Bootstrapping subsamples to avoid full-data bootstrapping. Sharma and Kumar (2021) adapted the bag-of-little-bootstraps approach of Kleiner et al. (2014) for phylogenomic analysis (Fig. 3d). In PSU, a small subsample of n sites, where $n << N$, is initially selected from the entire dataset, and bootstrap replicate datasets are generated by resampling N sites with replacement from that subsample. Each replicate alignment matches the full dataset length (N) but contains only n distinct sample sites. This upsampling step is a key innovation that restores the total number of sites and substitutions represented in the analysis while keeping the number of distinct sites the same as in the subsample. As shown in Figure 4, the average number of substitutions in PSU replicate datasets closely matches that in standard bootstrap replicates and in the full dataset. Thus, PSU deliberately trades complete site diversity for computational efficiency while preserving much of the statistical scale needed for confidence estimation.

Figure 4. Number of substitutions in PSU datasets. The number of substitutions in PSU replicate datasets (green) as compared to those in the standard bootstrap replicate datasets (red) and the full dataset (grey). Results are from the analysis of simulated datasets containing 446 species and MSAs with 50,000 to 536,534 sites, as reported by Sharma and Kumar (2021). The number of substitutions is the sum of ML estimates of branch lengths multiplied by the number of sites.

In the PSU analysis, multiple upsampled datasets are generated from each subsample. A phylogeny is inferred for each PSU replicate, and the proportion of PSU trees containing a specific clade yields the subsample bootstrap confidence level (*bcl*). Because each subsample includes only a fraction of the sites in the full dataset, multiple subsamples must be analyzed to obtain a stable estimate of full-data bootstrap support. Thus, the collection of bcl values across these subsamples is aggregated to estimate *FBS* for a given clade.

The way subsample *bcl* values are aggregated is important. Sharma and Kumar (2021) found that the mean *bcl*, although strongly correlated with *FBS*, tended to underestimate high *FBS* values and overestimate low *FBS* values. The median of the subsample *bcl* values, denoted *BCL*, provided a better approximation of conventional bootstrap support. This difference likely

reflects the indicator nature of clade support: each replicate tree either contains or does not contain a clade, unlike the continuous estimators emphasized in the original bag-of-little-bootstraps framework of Kleiner et al. (2014). The Appendix provides a simple illustration of why median aggregation can outperform mean aggregation for such indicator-style confidence measures.

BCL estimation requires only a fraction of the time and memory needed for full-data bootstrap analysis. For the 200×394 dataset, PSU reduced computation time by 79% and memory use by 95% (Table 1), with similar gains reported for other empirical datasets (Sharma and Kumar 2021). In many cases, subsampling of fewer than 5% of sites was sufficient, and approximately 10 subsamples with about 10 upsampling replicates each produced stable results. These reduced memory demands also make PSU analyses easier to parallelize across cores, further decreasing wall-clock time. PSU is therefore especially useful when full-data bootstrap analysis would exceed desktop memory limits or require days of computation.

Table 1. Performance of the PSU approach.

	PSU					Full Data	
	Subsample	Savings		Time	Memory	Time	Memory
Analysis type	size (max)	*Time*	*Memory*	(seconds)	(MBs)	(seconds)	(MBs)
Clade support	2.1%	79%	96%	994,142	144	4,734,008	3,773
Model selection	1.8%	99%	98%	11,160	292	928,701	14,688
Gamma parameter	8.4%	78%	91%	1,555	542	7,106	5,875
Model substitution rates	6.3%	86%	93%	980	410	7,106	5,875
Branch lengths (BLs)	8.4%	78%	91%	1,555	542	7,106	5,875
Variance (BLs)	8.4%	50%	86%	68,258	542	135,559	3,771
Molecular Clock Test	4.2%	85%	44%	2,651	1,358	17,650	2,427

Note. The avian phylogenomic dataset comprised 259 nuclear loci from 200 species (Prum et al. 2015). The concatenated multiple sequence alignment contains 394,686 sites and 336,490 unique site configurations.

Distributions of clade support from PSU. Unlike full-data bootstrapping, which reports a single support value for each clade (*FBS*), PSU reports a distribution of support values across subsamples. Some clades show consistently high support across subsamples, indicating strong genome-wide concordance (Fig. 5a). Others show broad or multimodal distributions, with some subsamples supporting the clade and others contradicting it (Fig. 5b). Such heterogeneity can reflect alignment or orthology errors, model misspecification, hidden biases, incomplete lineage sorting, or other sources of phylogenetic discordance (Lanfear and Hahn 2024; Sharma and Kumar 2024; Sharma and Kumar 2025). PSU detects this heterogeneity directly from random site subsamples rather than from predefined genes or partitions.

The distribution of *bcl* values can be used to estimate Net Bootstrap Support (*NBS*), a measure designed to reduce overconfidence in concatenated phylogenomic analyses (Sharma and Kumar 2025). Whereas *BCL* uses the median *bcl* to approximate conventional bootstrap

support (*BCL ~ FBS*), *NBS* is calculated as the mean of the subsample *bcl* distribution. This distinction is important. The median is useful when the goal is to approximate *FBS*, but it can downplay conflicting subsamples. The mean incorporates the entire distribution, including subsamples that weakly support or contradict a clade, and, therefore, provides a theoretically supported measure that is more sensitive to phylogenetic conflict (Sharma and Kumar 2022).

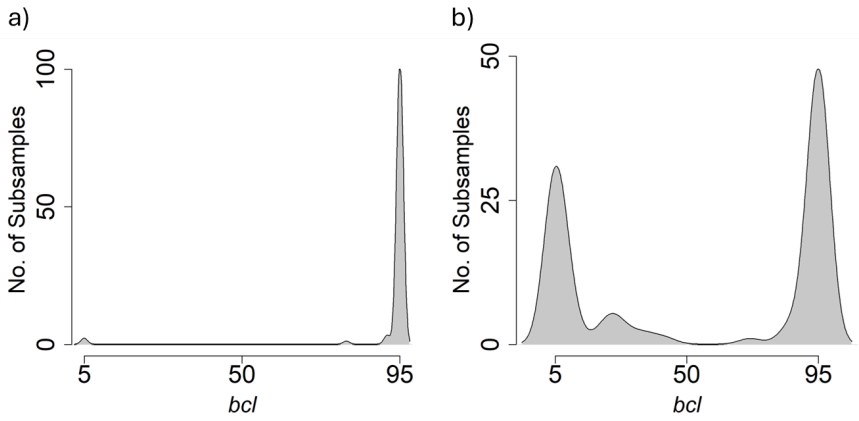

Figure 5. Distribution of subsample bootstrap support values (*bcl*) for two rodent clades. (**a**) Consistently high support across subsamples for a clade. (**b**) A distribution indicating conflicting phylogenetic signals across subsamples for the inferred clade. Results are from Sharma and Kumar (2025) and correspond to clades R1 and R1' in the rodent phylogeny presented in Figure 3 in Sharma and Kumar (2025).

Analyses of some simulated datasets suggest that *NBS* can perform similarly to MSC methods across phylogenomic datasets generated under low, moderate, and high levels of ILS (Fig. 6a) (Sharma and Kumar 2025). Interestingly, *NBS* could surpass MSC-based local posterior probabilities when gene tree estimation error (GTEE) was high (Fig. 6b). This outcome is biologically plausible because MSC methods depend on the quality of the input gene trees, which can become unreliable when individual gene alignments are short or contain few phylogenetically informative substitutions (Simmons and Gatesy 2015; Shen et al. 2021; Sharma and Kumar 2025). Because *NBS* calculation does not use gene (or partition) trees, it appears to be robust to high GTEE.

In addition to ILS and GTEE, one or a few loci with unusual phylogenetic histories, alignment issues, and hidden biases can drive bootstrap support for incorrect or contentious clades in concatenated analyses of hundreds of loci (Brown and Thomson 2016; Shen et al. 2017; Sharma and Kumar 2024; Sharma and Kumar 2025). In the PSU analysis, some subsamples will exclude such disruptive loci and sites, which can lead to reduced support for spurious clades driven by those sites (Sharma and Kumar 2025). In these cases, *NBS* is expected to be lower than FBS, which was indeed observed in the concatenated analysis of many empirical datasets (Sharma and Kumar 2025). Thus, *NBS* can serve as a practical way to reduce overconfidence.

Thus, PSU provides more than computational acceleration. By summarizing how clade support varies across independent site subsamples, it offers a direct diagnostic of concordant and conflicting phylogenetic signals within concatenated alignments. This information is often hidden when support is reported only as a single full-data bootstrap value.

Phylogenomics without data partitions. Conventional approaches for addressing phylogenetic heterogeneity often divide an alignment by gene, codon position, genomic region, or other biologically defined categories. These partitions may then be modeled separately within

concatenated analyses or used to reconstruct gene trees that are summarized by consensus or multispecies coalescent methods (Bull et al. 1993; Chippindale and Wiens 1994; Brandley et al. 2005; Gadagkar et al. 2005; Lanfear et al. 2017). Such strategies are powerful and widely used, but they also require decisions about how to partition the data, and different partitioning schemes can lead to different conclusions. In addition, fitting separate models to many partitions or estimating many gene trees can be computationally demanding, and short partitions may suffer from gene tree estimation error (Simmons and Gatesy 2015; Shen et al. 2021; Sharma and Kumar 2025). PSU offers a complementary strategy. Rather than requiring predefined partitions, it uses random site subsamples to ask whether support for a clade is consistent across the alignment. This does not replace model-based approaches that explicitly represent gene tree variation, but it provides a practical and computationally efficient way to detect phylogenetic conflict directly from concatenated data. We therefore suggest reporting NBS alongside FBS or BCL for clades inferred from concatenated phylogenomic analyses.

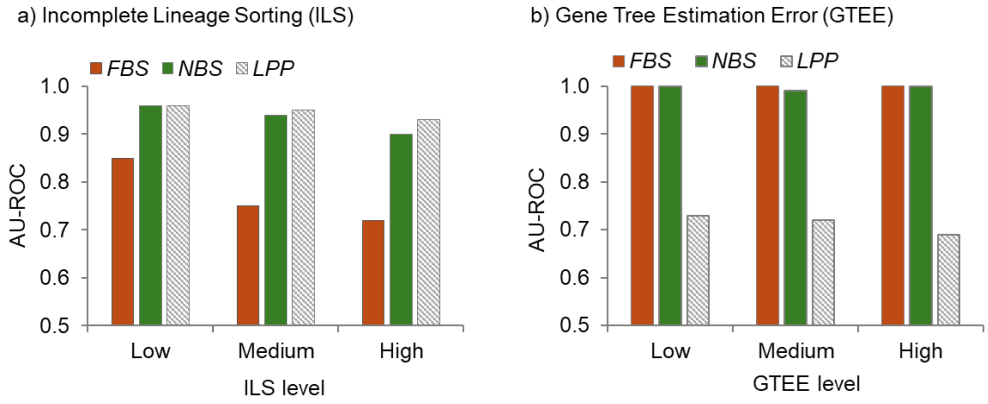

Figure 6. Accuracy of phylogeny support measures under heterogeneous conditions. Performance of support metrics in distinguishing correct and incorrect clades under varying levels of (**a**) incomplete lineage sorting, ILS, and (**b**) gene tree estimation error, GTEE, redrawn using data in figures 7 and 8 of Sharma and Kumar (2025). Simulated datasets were obtained from Mirarab et al. (2014) and simulated with low, medium, and high levels of ILS using a species tree of 37 mammalian species within a multispecies coalescent framework. Ten datasets were analyzed from each ILS category, each comprising 100 genes (1,600 sites per gene). Simulated datasets with GTEE were obtained from Shen et al. (2021), in which branch lengths were reduced to introduce increasing levels of low, medium, and high GTEE, with 1,000 gene sequence alignments for each category. Net Bootstrap Support (*NBS*), derived from the mean of subsample support values, incorporates both supporting and conflicting signals and achieves performance comparable to or exceeding local posterior probability (*LPP*) produced by multispecies coalescent (MSC)-based approaches under challenging conditions.

Automated tuning of PSU parameters. PSU analysis requires choices for three tuning parameters: the number of sites per subsample (n), the number of independent subsamples (S), and the number of upsampling replicates per subsample (R). These choices affect both accuracy and computational efficiency, so adaptive protocols have been developed to select them objectively (Sharma and Kumar 2021; Sharma and Kumar 2025). The general strategy is to begin with an empirically guided subsample size and a small number of subsamples and upsampling replicates. Support values are then estimated and monitored across iterations. Additional subsamples are added until average support values for clades in different support ranges stabilize within predefined tolerances. If convergence is not achieved, the subsample size is increased, and the process is repeated. Once stability criteria are met, BCL and NBS are estimated from the resulting collection of PSU replicate phylogenies. This adaptive approach

reduces the need for ad hoc parameter tuning and makes PSU analyses more reproducible. It is also practically important because optimal choices depend on dataset-specific properties such as alignment length, divergence, missing data, and phylogenetic complexity.

Evolutionary Hypothesis Testing Using PSU

PSU is not limited to bootstrap support. Many tasks in molecular phylogenetics involve likelihood-based model comparison or hypothesis testing, and these analyses can also be accelerated by analyzing subsamples using an upsampling and aggregation strategy. Here, we consider three examples: substitution-model selection, tests of nested evolutionary hypotheses, and comparisons of non-nested tree topologies.

Selecting the optimal substitution model. Choosing nucleotide or amino acid substitution models is a standard step in likelihood-based phylogenetics, typically involving likelihood-ratio tests or information criteria (Posada and Crandall 1998; Posada 2008; Kalyaanamoorthy et al. 2017; Darriba et al. 2020; Sharma and Kumar 2022). For long phylogenomic alignments, this step can become expensive because many candidate models must be optimized and compared. PSU reduces this cost by evaluating candidate models on upsampled subsamples rather than on the full alignment (Sharma and Kumar 2022). For the 200×394 dataset, PSU selected the same final model as the full-data analysis while reducing runtime from 258 CPU hours to 1.5 CPU hours and memory use from 14 GB to 0.15 GB (Table 1). Across additional empirical datasets, PSU showed orders-of-magnitude reductions in memory use and substantially improved scaling with data size (Fig. 7b and c). For model selection, only one upsampled replicate is needed for each subsample size, since the goal is not to estimate bootstrap support. The subsample size is increased progressively until the same substitution model is selected in consecutive analyses (Fig. 7a) (Sharma and Kumar 2022). In empirical applications, convergence was usually achieved using less than 5% of the distinct site configurations in the full dataset.

Testing nested evolutionary hypotheses. PSU can also be used for likelihood-ratio tests of nested evolutionary hypotheses. For example, the molecular clock hypothesis can be tested by comparing the log-likelihood of a rooted phylogeny under clock and no-clock models and computing the usual test statistic, $2\Delta lnL$. In a PSU implementation, the test is first conducted on an upsampled subsample. The subsample size is then progressively increased until consecutive analyses yield the same statistical conclusion at a chosen significance threshold. The final test statistic and *P*-value are taken from the last converged analysis. Applied to the 200×394 dataset, this protocol reached convergence using only 4.2% of the full dataset, and it produced the same biological conclusion as the full-data analysis, rejecting the molecular clock with a highly significant P-value (Table 1).

Although illustrated here with the molecular clock test, the same strategy can be applied to other nested evolutionary hypotheses, including local-clock models and nested constraints on substitution processes or selection regimes (e.g., (Yoder and Yang 2000)). PSU therefore provides a scalable route to likelihood-based hypothesis testing when full-data calculations may be time-consuming or impractical.

Testing non-nested hypotheses. PSU can also be used for likelihood-based hypothesis testing among tree topologies (e.g., T1 and T2) when parametric likelihood-ratio tests are not suitable.

In a typical bootstrap setup, the distribution of Δ*ln*L = *ln*L(T1) - *ln*L(T2) can be computed across bootstrap replicates under a fixed substitution model. The null hypothesis Δ*ln*L = 0 can be tested by constructing a confidence interval for the statistic using PSU, based on the estimated variance of Δ*ln*L. We conducted such an analysis for two contrasting rodent trees: one from a concatenated alignment (T1) and the other from an MSC analysis (T2). They differ in the placement of a single taxon (Roycroft et al. 2020; Shen et al. 2021; Sharma and Kumar 2025).

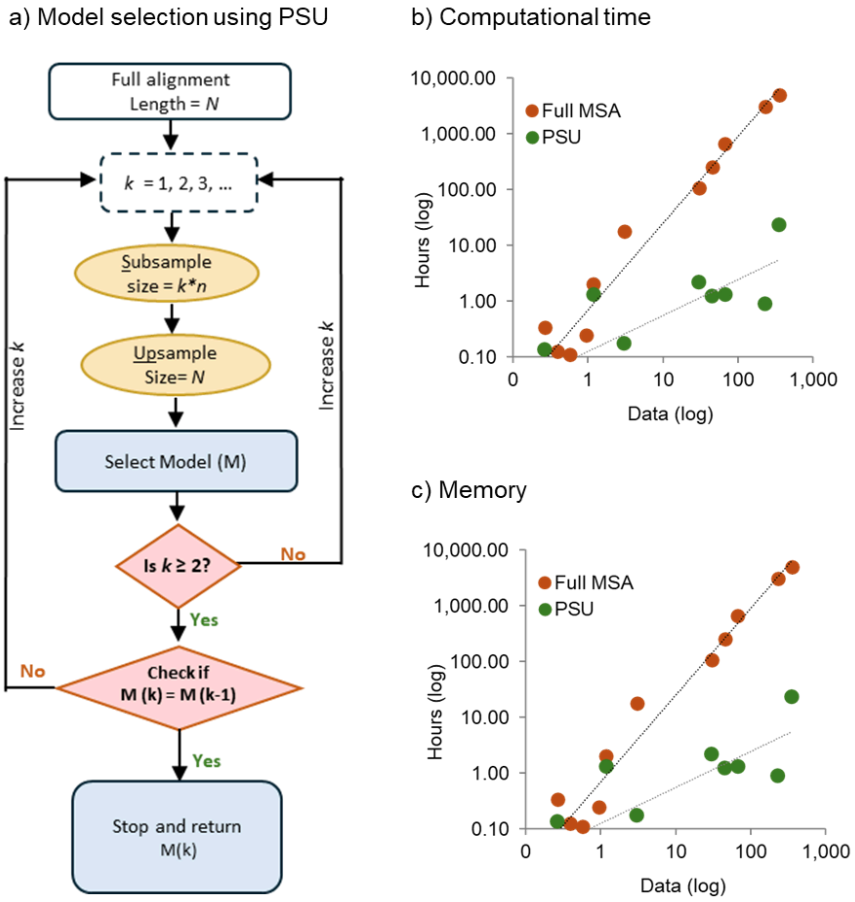

Figure 7. Scaling behavior of PSU versus full-data analysis for model selection. (**a**) A flowchart demonstrating the steps of the PSU framework for selecting the best-fit model of substitution. (**b**) Memory and (**c**) runtime requirements for substitution model selection as a function of data size (*D*, log scale), using data from Sharma and Kumar (2022). PSU exhibits sublinear scaling because the number of distinct site patterns evaluated remains limited, even as the total number of sites and substitutions scales to full dataset size through upsampling. PSU time and memory needs are of order $O(D^{0.64})$ and $O(D^{0.23})$, respectively, compared to the full-data analysis, which are of order $O(D^{1.56})$ and $O(D^{1.0})$, respectively, for the data set analyzed (see Figure 2).

We estimated the variance of ΔlnL across multiple upsampling replicates for each subsample, then averaged these estimates across subsamples, following the logic of Kleiner et al. (2014). Additional subsamples were added until the variance estimate stabilized within a predefined tolerance. In this example, convergence was achieved with seven subsamples. The resulting 95% confidence interval for Δ*ln*L ranged from -78.6 to 231.3, including 0, indicating that the concatenation tree was not significantly better supported than the alternative topology. This result is consistent with the low NBS for the defining clade of T1 and with its low posterior probability in the MSC analysis (Roycroft et al. 2020; Shen et al. 2021; Sharma and Kumar 2025).

Estimating Evolutionary Parameters

PSU also extends to the estimation of evolutionary parameters, including substitution model parameters, branch lengths, divergence times, and their associated variances. In these

applications, the target is not clade support but the stabilization of numerical estimates across progressively larger upsampled subsamples.

Estimating substitution model parameters. Substitution-model parameters estimated during PSU model selection were reported to be nearly identical to those obtained from full-data analysis, including the gamma shape parameter for among-site rate variation (Sharma and Kumar 2022). This suggests that PSU can be used not only to select models but also to estimate their parameters efficiently. For example, PSU can estimate the gamma shape parameter (α) by progressively increasing the subsample size and monitoring changes in α across iterations. The process stops when consecutive estimates fall within a predefined tolerance, 1% by default. Applied to the 200×394 dataset, this protocol converged after the fourth iteration, using at most 8.38% of all sites, and required less memory and runtime than full maximum-likelihood estimation using the complete dataset. Across empirical datasets, PSU-based estimates of α (0.398) closely matched full-data estimates (0.394). The same strategy can be used for substitution-rate parameters. In this case, convergence is assessed by comparing rate estimates across consecutive iterations, for example, by requiring a correlation coefficient greater than 0.99. For the 200×394 dataset, convergence was achieved after the third iteration using only 6.3% of sites, and the resulting rate estimates closely matched those from the full alignment (Fig. 8a**;** slope = 1.00, correlation = 0.999).

Estimating branch lengths and their variances. Branch lengths are central to molecular evolutionary analysis because they measure the amount of evolutionary change along lineages and are used to identify rate variation, test evolutionary hypotheses, and estimate divergence times (Tamura et al. 2012). In large phylogenomic alignments, estimating branch lengths by maximum likelihood can be computationally demanding because branch lengths and model parameters must be optimized across many site patterns. A PSU protocol similar to that used for substitution-model parameters can be applied to branch length estimation on a given phylogeny. Here, the convergence of estimates can be assessed by monitoring the stability of short branch-length estimates as the subsample size increases. Short branches provide a stringent target because they are generally harder to estimate reliably than long branches. In our analyses, a stopping rule based on a correlation greater than 0.99 for the shortest 25% of branch lengths between successive iterations was effective.

Applied to the 200×394 dataset, this protocol produced branch length estimates that closely matched those from the full alignment (Fig. 8b; slope = 0.95, correlation = 0.99), while reducing runtime and memory use by 4.6-fold and 10.8-fold, respectively. PSU can also estimate branch length variances, which are needed for hypothesis testing (Dopazo 1994) and for confidence intervals in divergence-time methods such as RelTime (Tamura et al. 2012; Tamura et al. 2018; Tao et al. 2020). After a stable subsample size is identified for branch length estimation, multiple subsamples and upsampled replicates are analyzed. Branch-specific variances are calculated within each subsample and then averaged across subsamples, following the bag-of-little-bootstraps strategy of Kleiner et al. (2014). Convergence is assessed using the coefficients of variation for the shortest 25% of branches, with estimates considered stable when the correlation between successive iterations exceeds 0.99. For the 200×394 dataset, this approach closely matched full-bootstrap variance estimates while requiring substantially less time and memory (Fig. 8c; Table 1).

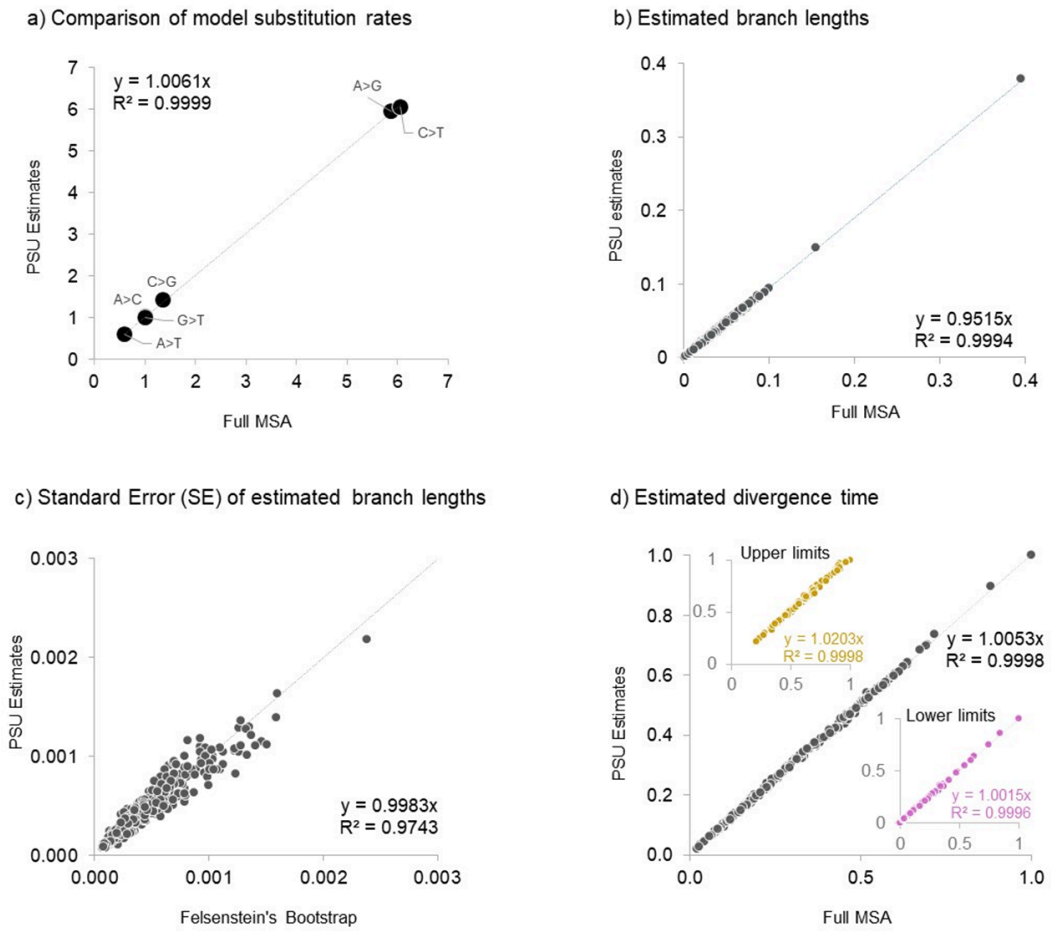

Figure 8. PSU estimates of model parameters, branch lengths, and divergence times for the 200×394 dataset. (**a**) Comparison of model substitution rate parameter estimates obtained from PSU and full dataset analyses. (**b**) Comparison of branch lengths for all internal and tip branches estimated using PSU and full dataset analysis. Branch lengths are estimated for the bird phylogeny and the GTR+G4 model of substitutions. (**c**) Comparison of branch length standard error (square root of estimated variance) from the standard bootstrap (x-axis) and PSU approach (y-axis). (**d**) Comparison of the relative time estimated using the RelTime approach for the given phylogeny with estimated branch lengths. The gray circle represents the relative node times, while the purple and yellow circles represent the upper and lower limits of 95% confidence intervals.

The divergence times obtained from the RelTime framework (Tamura et al. 2012; Tamura et al. 2018; Tao et al. 2020) using branch lengths estimated from the PSU approach were highly concordant with those estimated from the full dataset (Fig. 8d; slope = 1.005, R^2 = 0.999). In addition to the relative node time, their 95% confidence intervals from the PSU approach closely matched those obtained from the full dataset analysis (Fig. 8d, CI_{lower}: slope = 1.001 and R^2 = 0.999, CI_{upper}: slope = 1.020 and R^2 = 0.999).

Bootstrap Consensus Phylogeny and Branch Lengths

The consensus of bootstrap replicate trees is often used as the inferred phylogeny (Felsenstein 1985). Because phylogenomic datasets frequently yield high support for most clades,

PSU-derived replicate trees can also be used to construct majority-rule consensus phylogenies. In such cases, clade hypotheses, support values, and estimates of phylogenetic heterogeneity are obtained within the same computational analysis.

Branch lengths and their variances can also be estimated directly from the PSU replicate trees used to build the consensus. Replicate trees are scanned to identify branches corresponding to clades in the PSU consensus tree. For each such branch, lengths are extracted across replicate trees within each subsample, and their means and variances are then averaged across subsamples. To avoid unstable estimates, variance calculations can be restricted to cases in which the clade appears in a minimum number of replicate trees within a subsample.

Our tests using empirical datasets showed that branch length (slope = 0.95, correlation = 0.99) and variance (slope = 0.97, correlation = 0.98) estimates obtained in this way closely matched those from full data analyses. These results have an important practical implication: in many large phylogenomic applications, PSU may reduce or eliminate the need for a separate full-data ML analysis. A single PSU analysis can yield a consensus phylogeny, clade support, branch length estimates, model parameters, and associated measures of uncertainty.

Applying PSU beyond ML

Although this article has focused primarily on maximum-likelihood analyses, the PSU logic may extend to other phylogenetic methods when computational cost is strongly influenced by the number of distinct site patterns or when repeated analyses of long alignments are required. For example, maximum parsimony analyses can benefit because repeated site patterns need not be evaluated independently. Distance-based methods, including Neighbor-Joining (Saitou and Nei 1987), may also benefit when distance estimation or bootstrap testing must be repeated many times for very long alignments. Bayesian approaches may be more complex because posterior sampling behavior depends on additional factors, but upsampled subsamples could still provide useful approximations in some settings.

Thus, PSU is best viewed as a general computational template: subsample sites, upsample to full-alignment scale, compute the target quantity, and aggregate results across subsamples. Its usefulness will depend on the inference method, the computational bottleneck, and the quantity being estimated. This flexibility should allow PSU to benefit from continuing advances in scalable phylogenetic algorithms.

Scaling with Increasing Numbers of Sequences

While PSU addresses the computational burden in phylogenomics arising from increasing alignment length, the number of sequences (taxa) is also growing rapidly in modern datasets (Pattengale et al. 2010; Minh et al. 2013; Sharma and Kumar 2021). This growth introduces additional computational challenges for which fast heuristic methods have been developed to speed up the bootstrap analysis of many sequences (Stamatakis et al. 2008; Minh et al. 2013). These approaches can be effectively combined with PSU to achieve computational efficiency in both dimensions of an alignment (many sites and many taxa). For example, Sharma and Kumar (2021) estimated subsample-wise clade support using the Ultrafast Bootstrap (UFB) method (Minh et al. 2013; Hoang et al. 2018). Briefly, each subsample is analyzed using UFBoot, which repeatedly upsamples the subsample alignment R times, typically $R \geq 1000$, to generate replicate phylogenies from which clades and their corresponding subsample bootstrap support

values, bcl, are estimated. These clade-specific *bcl* values are then used to estimate *BCL* and *NBS*.

For a mammal dataset containing 37 species and 1,391,742 sites, Sharma and Kumar (2021) reported that the combined PSU+UFB approach required only 0.83 hours and 0.16 GB of memory on a standard multi-core computer, using 10 subsamples and default UFB settings. This represented a substantial improvement over using UFB alone, which required 7.50 hours and 6.7 GB of memory, and PSU alone, which required 18.9 hours. Similar to the FBS results, all clades except one were recovered with BCL values ≥ 95%, whereas the remaining clade received 90% BCL. These results illustrate that PSU can be synergistically combined with fast bootstrap heuristics to achieve substantial gains in both runtime and memory efficiency.

Other advances in phylogenetic inference can also be integrated into the PSU framework, including improved tree-search strategies, approximate likelihood calculations, and alternative measures of branch support such as transfer bootstrap expectation (Lemoine et al. 2018). Because PSU operates through a general subsample-upsample-compute-aggregate template, it can be paired with different inference engines as long as the computational burden is reduced by limiting the number of distinct sites evaluated. This makes PSU complementary to ongoing algorithmic improvements for large numbers of taxa.

Limitations of PSU

PSU is a theoretically motivated approximation rather than an exact reformulation of full-data phylogenetic inference. Its connection to the bag-of-little-bootstraps framework provides statistical motivation, but it does not establish equivalence between PSU and full-data maximum-likelihood analysis. At present, the strongest support for PSU comes from empirical performance across simulated and real phylogenomic datasets. PSU should therefore be viewed as a complementary framework that prioritizes scalability while approximating quantities that would otherwise require more expensive full-data analysis.

The effectiveness of PSU depends on whether small subsamples capture representative patterns of evolutionary change. When datasets contain extensive missing data, weak phylogenetic signals, or strong heterogeneity, larger subsamples may be required to achieve stable estimates. In such cases, the computational advantage of PSU will decrease, and in the limiting case where the required subsample size approaches the full alignment, PSU offers little benefit.

Because PSU relies on random site subsamples, results can vary across independent runs, especially when the number of subsamples is small or when subsample sizes are insufficient. This variability can affect estimates of clade support, model parameters, branch lengths, and likelihood-based statistics. Adaptive protocols mitigate this problem by increasing the number or size of subsamples until estimates stabilize, but optimal settings remain dataset-dependent.

Finally, PSU does not replace methods that explicitly model particular biological processes. For example, NBS can reveal conflict in concatenated alignments, but it is not a substitute for multispecies coalescent models when the goal is to model gene tree variation directly. Further systematic evaluation across diverse phylogenomic scenarios will be important for refining convergence criteria, identifying failure modes, and defining best practices for PSU analyses.

CONCLUSIONS

PSU offers a practical strategy for scaling phylogenomic inference by analyzing many small-site subsamples, which are then upsampled to full alignment length. Its efficiency arises from a useful separation between computational cost and statistical scale: likelihood-based computation is strongly influenced by the number of distinct sites evaluated, whereas many measures of inferential power depend on the total number of sites and substitutions represented. By reducing the former while restoring the latter through upsampling, PSU can approximate many full-data analyses at substantially lower computational cost.

The framework also expands what standard bootstrap analysis can provide. Instead of reporting only a single support value for each clade, PSU yields a distribution of support values across subsamples. The median of this distribution, BCL, approximates conventional bootstrap support, whereas the mean, NBS, incorporates conflicting subsample signals and can reduce overconfidence in concatenated analyses. These distributions make PSU not only a computational shortcut but also a diagnostic tool for detecting heterogeneity in phylogenomic signals.

Beyond clade support, PSU provides a common framework for substitution model selection, likelihood-based hypothesis testing, branch-length estimation, variance estimation, and potentially divergence-time analysis. In fact, a single PSU analysis can yield the consensus phylogeny, support values, model parameters, branch lengths, and associated uncertainty measures, reducing or eliminating the need for separate full-data analyses. With adaptive protocols for tuning subsample size, number of subsamples, and number of upsampling replicates, PSU is becoming increasingly practical for routine use.

As phylogenomic alignments continue to grow, scalable methods will be essential for making rigorous evolutionary inference accessible on standard computing platforms. With the upcoming implementation in widely used software packages, such as MEGA (Kumar et al. 2024), PSU is well positioned to be an option for big data phylogenomics.

Data and Code availability

The phylogenomic datasets discussed in this article are publicly available from the corresponding references. R codes for performing PSU analyses are available on GitHub at https://github.com/ssharma2712.

Acknowledgments

We thank John Allard and Glen Stecher for many helpful comments. This work was supported by a research grant from the National Institutes of Health (R35GM139540-06) and the National Science Foundation (DBI-2505985).

References

Azouri D, Abadi S, Mansour Y, Mayrose I, Pupko T. 2021. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat. Commun.* 12:1–9.

Brandley MC, Schmitz A, Reeder TW. 2005. Partitioned Bayesian analyses, partition choice,

and the phylogenetic relationships of scincid lizards. *Syst. Biol.* 54:373–390.

Brown JM, Thomson RC. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* 66:517–530.

Bull JJ, Huelsenbeck JP, Cunningham CW, Swofford DL, Waddell PJ. 1993. Partitioning and Combining Data in Phylogenetic Analysis. *Syst. Biol.* 42:384–397.

Chippindale PT, Wiens JJ. 1994. Weighting, Partitioning, and Combining Characters in Phylogenetic Analysis. *Syst. Biol.* 43:278–287.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. 2020. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* 37:291–294.

Dopazo J. 1994. Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. *Journal of Molecular Evolution 1994 38:3* 38:300–304.

Edwards SV. 2016. Phylogenomic subsampling: a brief review. *Zoologica Scripta* 45:63–74.

Efron B. 1979. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* 7:1–26.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* 61:717–726.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. B Mol. Dev. Evol.* 304:64–74.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* 35:518–522.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.

Kleiner A, Talwalkar A, Sarkar P, Jordan MI. 2014. A Scalable Bootstrap for Massive Data. *J. R. Stat. Soc. Series B Stat. Methodol.* 76:795–816.

Kumar S. 2022. Embracing Green Computing in Molecular Phylogenetics. *Mol. Biol. Evol.* 39:msac043.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol. Biol. Evol.* 29:457–472.

Kumar S, Stecher G, Suleski M, Sanderford M, Sharma S, Tamura K. 2024. MEGA12: Molecular evolutionary genetic analysis version 12 for adaptive and green computing. *Mol. Biol. Evol.* 41:msae263.

Kumar S, Tao Q, Lamarca AP, Tamura K. 2023. Computational reproducibility of molecular

phylogenies. *Mol. Biol. Evol.* 40:msad165.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Mol. Biol. Evol.* 34:772–773.

Lanfear R, Hahn MW. 2024. The meaning and measure of concordance factors in phylogenomics. *Mol. Biol. Evol.* 41:msae214.

Lemoine F, Domelevo Entfellner JB, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* 556:452–456.

Lozano-Fernandez J. 2022. A practical guide to design and assess a phylogenomic study. *Genome Biol. Evol.* 14:evac129.

Minh BQ, Nguyen MAT, Von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30:1188–1195.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37:1530–1534.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.

Mongiardino Koch N. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. *Mol. Biol. Evol.* 38:4025–4038.

Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. 2010. How many bootstrap replicates are necessary? *J. Comput. Biol.* 17:337–354.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* 36:541–562.

Posada D. 2008. jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol.* 25:1253–1256.

Posada D, Crandall KA. 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569–573.

Roycroft EJ, Moussalli A, Rowe KC. 2020. Phylogenomics Uncovers Confidence and Conflict in the Rapid Radiation of Australo-Papuan Rodents. *Syst. Biol.* 69:431–444.

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

Seo T-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Mol. Biol. Evol.* 25:960–971.

Sharma S, Kumar S. 2021. Fast and accurate bootstrap confidence limits on genome-scale phylogenies using little bootstraps. *Nat. Comput. Sci.* 1:573–577.

Sharma S, Kumar S. 2022. Taming the Selection of Optimal Substitution Models in Phylogenomics by Site Subsampling and Upsampling. *Mol. Biol. Evol.* 39:msac236.

Sharma S, Kumar S. 2024. Discovering fragile clades and causal sequences in phylogenomics by evolutionary sparse learning. *Mol. Biol. Evol.* 41:msae131.

Sharma S, Kumar S. 2025. Robust and efficient confidence limits for phylogenomic inference of organismal relationships. *Mol. Biol. Evol.* 42:msaf296.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol* 1:126.

Shen X-X, Steenwyk JL, Rokas A. 2021. Dissecting Incongruence between Concatenation- and Quartet-Based Approaches in Phylogenomic Data. *Syst. Biol.* 70:997–1014.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Mol. Phylogenet. Evol.* 91:98–122.

Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U. S. A.* 109:14942–14947.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stamatakis A, Hoover P, Rougemont J. 2008. A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst. Biol.* 57:758–771.

Strimmer K, Von Haeseler A. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964–969.

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U. S. A.* 109:19333–19338.

Tamura K, Tao Q, Kumar S. 2018. Theoretical foundation of the reltime method for estimating divergence times from variable evolutionary rates. *Mol. Biol. Evol.* 35:1770–1782.

Tao Q, Tamura K, Mello B, Kumar S. 2020. Reliable Confidence Intervals for RelTime Estimates of Evolutionary Divergence Times. *Mol. Biol. Evol.* 37:280–290.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* 17:1081–1090.

Appendix

The mean and median are two commonly used summary statistics for aggregating subsample-wise estimates in little bootstrap analyses (Kleiner et al. 2014). The relative performance of these bagging strategies can be evaluated within the framework of the indicator bootstrap, analogous to phylogenetic bootstrap analysis, in which the presence or absence of a clade is recorded for each upsample generated from a subsample. In the indicator bootstrap, each bootstrap upsample is assigned a binary score: 1 if it supports the null hypothesis and 0 otherwise. The nonparametric P-value is then estimated as the proportion of replicates that do not support the null hypothesis. Similarly, in phylogenetic bootstrapping, the occurrence of a clade in a bootstrap replicate tree is scored as 1, whereas its absence is scored as 0, providing a connection between indicator-based statistical inference and clade support estimation. However, bootstrap support for a clade is defined as the proportion of replicate trees that contain that clade, whereas the p-value in the indicator bootstrap represents the complementary quantity, the proportion of bootstrap replicates that do not support the null hypothesis.

The performance of the two bagging strategies was evaluated using data simulated from a normal distribution with mean −5 and standard deviation 1 ($X \sim Normal(\mu = -5, \sigma = 1)$) as the population. From this population, we generated 100 random samples of size N = 25,000 to test a one-tailed hypothesis. The hypothesis test was formulated as follows: the null hypothesis $H_0: \mu = -5.01$ was tested against the one-sided alternative $H_1: \mu < -5.01$. The threshold was chosen near the population mean to yield a broad range of p-values. For each sample, p-values were computed using both a parametric one-sample t-test and a nonparametric bootstrap test. To implement the little bootstrap framework, subsamples of size $n = N^g$ were used, with g varying from 0.6 to 0.9. For each subsample size category, S = 1000 subsamples were analyzed, with R = 100 upsampled replicates per subsample, to investigate the asymptotic behavior of the estimates.

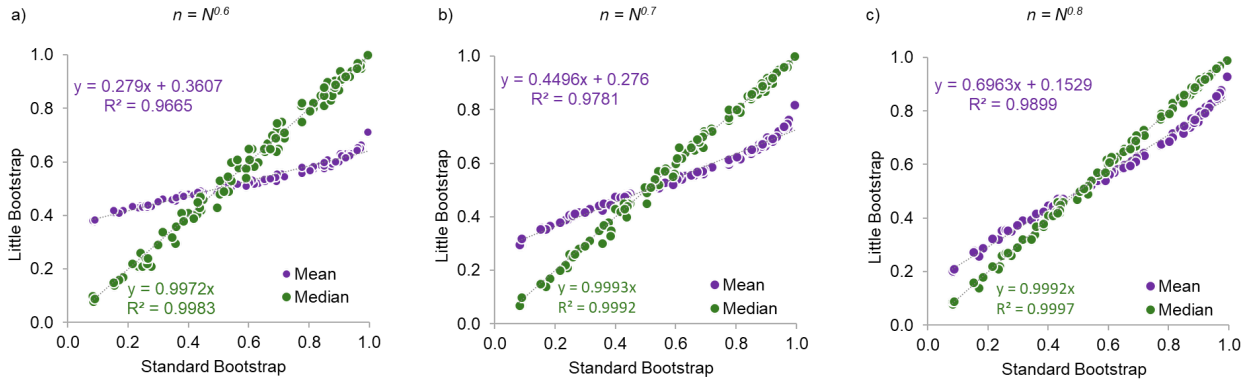

Figure A1 | Comparison of p-values from the standard bootstrap and little bootstrap with mean (blue) and median bagging (orange). For the little bootstrap, subsamples of **(a)** $N^{0.6}$, **(b)** $N^{0.7}$, and **(c)** $N^{0.8}$ have been analyzed, where N is the number of total data points in the sample.

The goal of the little bootstrap analysis is to approximate the results obtained from the standard bootstrap analysis. Therefore, p-value estimates obtained using mean and median bagging were compared with those from the standard bootstrap approach. The analysis shows that median bagging yields more accurate p-value estimates than mean bagging (Fig. A1 a-c). This is because the mean aggregation is sensitive to a small number of subsamples that strongly

support or reject the hypothesis, which can bias the estimated mean toward 0.5, the expected value under randomness. As a result, mean bagging tends to overestimate p-values when the true p-value is <0.5 and to underestimate them when the true p-value is >0.5. A similar pattern has been observed in little bootstrap analyses in phylogenomics (see Fig. 1f in Sharma and Kumar (2021)). These results suggest that median bagging is a more reliable aggregation strategy when the parameter of interest is defined in terms of an indicator variable.

Supplementary Table S1

Data	Base type	Species	Sites	Data size (Species ✖ Sites)	Data size (Million)	DOI
Butterflies	DNA	61	5,267,461	321,315,121	321.32	https://doi.org/10.1093/sysbio/syz030
Insects A	DNA	174	3,011,544	524,008,656	524.01	https://doi.org/10.1016/j.cub.2017.01.027
Insects B	DNA	48	2,938,039	141,025,872	141.03	https://doi.org/10.1016/j.ympev.2017.12.005
Mammal A	DNA	39	1,391,742	54,277,938	54.28	https://doi.org/10.1073/pnas.1211733109
Birds	DNA	200	394,684	78,936,800	78.94	https://doi.org/10.1038/nature15697
Plants	DNA	16	4,246,454	67,943,264	67.94	https://doi.org/10.1016/j.ympev.2018.08.011
Mammal B	DNA	274	7,370	2,019,380	2.02	https://doi.org/10.1098/rspb.2012.0683
Lassa Virus	DNA	179	3,186	570,294	0.57	https://doi.org/10.1016/j.cell.2015.07.020
Vertebrate	AA	58	1,806,035	104,750,030	104.75	https://doi.org/10.1093/sysbio/syv059
Yeast	AA	23	634,530	14,594,190	14.59	https://doi.org/10.1038/nature12130
Green plants	AA	360	19,449	7,001,640	7.00	https://doi.org/10.1186/1471-2148-14-23

Supplementary Table 1*:* Data refer to genomic sequence alignments from the indicated species. Bases denote the type of nucleotide or amino acid data. Data size is calculated as the product of the number of species and the sequence length, representing the total number of aligned characters and the computational time and memory required for ML analysis. DOI provides the digital object identifiers of the studies in which these datasets were originally published and analyzed.