

StarBeast3: Adaptive Parallelized Bayesian Inference under the Multispecies Coalescent

JORDAN DOUGLAS^{1,*}, CINTHY L. JIMÉNEZ-SILVA², AND REMCO BOUCKAERT¹

¹School of Computer Science, University of Auckland, 9 Symonds Street Level 1 Student Commons, Auckland 1010, New Zealand and ²School of Biological Sciences, University of Auckland, 9 Symonds Street Level 1 Student Commons, Auckland 1010, New Zealand;

*Correspondence to be sent to: School of Computer Science, University of Auckland, 9 Symonds Street Level 1 Student Commons, Auckland 1010, New Zealand; E-mail: jordan.douglas@auckland.ac.nz.

Received 17 October 2021; reviews returned 1 February 2022; accepted 8 February 2022

Associate Editor: Rayna Bell

Abstract.—As genomic sequence data become increasingly available, inferring the phylogeny of the species as that of concatenated genomic data can be enticing. However, this approach makes for a biased estimator of branch lengths and substitution rates and an inconsistent estimator of tree topology. Bayesian multispecies coalescent (MSC) methods address these issues. This is achieved by constraining a set of gene trees within a species tree and jointly inferring both under a Bayesian framework. However, this approach comes at the cost of increased computational demand. Here, we introduce StarBeast3—a software package for efficient Bayesian inference under the MSC model via Markov chain Monte Carlo. We gain efficiency by introducing cutting-edge proposal kernels and adaptive operators, and StarBeast3 is particularly efficient when a relaxed clock model is applied. Furthermore, gene-tree inference is parallelized, allowing the software to scale with the size of the problem. We validated our software and benchmarked its performance using three real and two synthetic data sets. Our results indicate that StarBeast3 is up to one-and-a-half orders of magnitude faster than StarBeast2, and therefore more than two orders faster than *BEAST, depending on the data set and on the parameter, and can achieve convergence on large data sets with hundreds of genes. StarBeast3 is open-source and is easy to set up with a friendly graphical user interface. [Adaptive; Bayesian inference; BEAST 2; effective population sizes; high performance; multispecies coalescent; parallelization; phylogenetics.]

Existing methods for testing macroevolutionary and macroecological questions have not kept pace with the explosion of next-generation sequence data now available (Blom et al. 2016b; Bragg et al. 2017; Stenson et al. 2017). Despite burgeoning databases of within- and between-species genomic diversity (Blom et al. 2016b; Bragg et al. 2017; Stenson et al. 2017), it is still common practice to ignore the gene-tree discordance that underlies any species phylogeny inferred from multilocus sequences and instead infer species ancestry based on concatenated sequence data taken to represent all underlying gene histories (Degnan and Rosenberg 2009; Heled and Drummond 2010; Jones 2017; Ogilvie et al. 2017; Rannala and Yang 2017). While this approach can perform well for inferring topologies when branches are long and incomplete lineage sorting (ILS) is absent, these conditions are rarely met.

Species trees inferred from concatenated sequences are often topologically incorrect (Degnan and Rosenberg 2009; Heled and Drummond 2010; Ogilvie et al. 2017), provide biased estimates for branch lengths and substitution rates (Kubatko et al. 2011; Ogilvie et al. 2016; Mendes and Hahn 2016), and underestimate uncertainty in tree topology, resulting in an unjustified degree of confidence in the wrong tree (Heled and Drummond 2010; Ogilvie et al. 2017). Such biases are exacerbated by subsampling of incongruent genes (Edwards et al. 2016; Mendes and Hahn 2016) and hold even for deep splits in the tree (Oliver 2013). These are crucial concerns in themselves and, more generally, can lead to biased estimates and erroneous inferences about fundamental evolutionary and ecological processes that require accurate phylogenetic trees, such as rates of speciation and extinction (Cadena et al. 2011; Rowe et al. 2011; Pepper

et al. 2013), rates of substitution in DNA sequences (Bouckaert et al. 2013) and morphological characters (Pepper et al. 2013), species ancestry and ancestral age estimation (Mitchell et al. 2014), geographical history and origins (Lemey et al. 2009; Bouckaert 2016), and species delimitation (Yang and Rannala 2010; Grummer et al. 2013; Leaché et al. 2014; Yang and Rannala 2014).

The multispecies coalescent (MSC; Maddison 1997; Edwards 2009; Liu et al. 2009) is an approach designed to minimize these potential biases by modeling macroevolution as a distribution of gene trees constrained by a species tree (Degnan and Rosenberg 2009; Heled and Drummond 2010; Jones 2017; Ogilvie et al. 2017; Rannala and Yang 2017). In doing so, the MSC provides a more biologically realistic framework for phylogenetic inference that captures the process of ILS underlying most multilocus phylogenies. Furthermore, by explicitly modeling both species and gene trees, the MSC can address questions that cannot be addressed under a concatenation approach—such as automatic species delimitation (Fujita et al. 2012), with important implications for biodiversity assessment and conservation (Bickford et al. 2007).

A number of software packages have implemented the MSC in various ways (see review by (Liu et al., 2015)). Our work at the Centre for Computational Evolution at the University of Auckland has led the development of *BEAST (STARBeast; Heled and Drummond 2010) and StarBeast2 (Ogilvie et al. 2017)—full Bayesian MSC frameworks for species-tree estimation from multilocus sequence data—and UglyTrees for visualizing these models (Douglas 2020). By explicitly modeling the MSC and avoiding the biases associated with concatenation methods (Heled and Drummond 2010; Ogilvie et al.

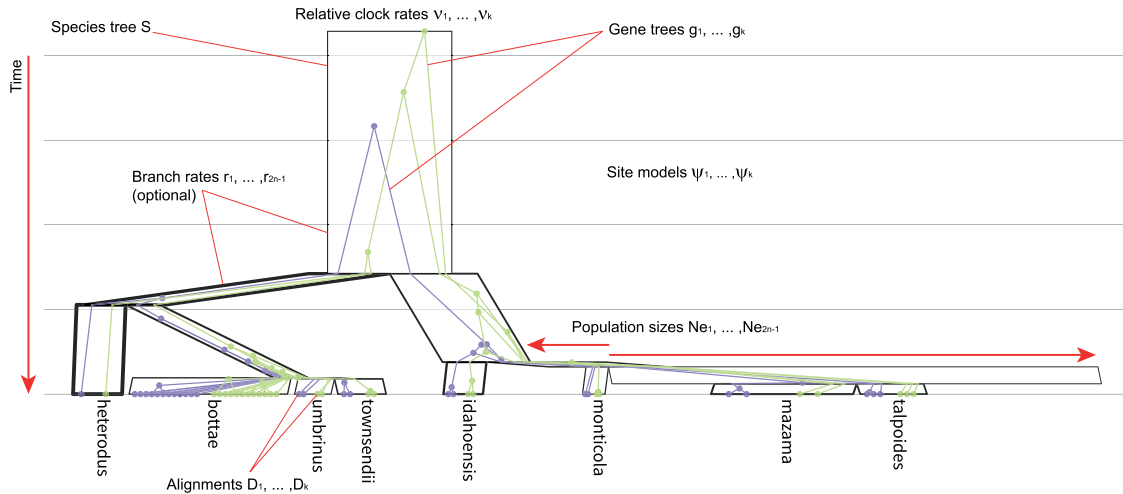


FIGURE 1. Depiction of the multispecies coalescent model, with $k=2$ gene trees constrained within a single species tree S with $n=8$ species. In this depiction, node heights (age) run along the y -axis and species-tree node widths are proportional to effective population sizes (arbitrary units). The relative molecular substitution rate of each species-tree branch is proportional to line thickness. Tree was built from a Gopher data set (Belfiore et al. 2008) and visualized using UglyTrees (Douglas 2020).

2016, 2017), an analysis using either of these software packages can significantly improve the conclusions drawn from data.

However, despite some advances in computational efficiency of the full Bayesian MSC (Jones 2017; Ogilvie et al. 2017; Rannala and Yang 2017), these complex models remain computationally intractable for large next-generation sequence data sets of 100's of sequenced loci across hundreds of individuals (i.e., 10^4 – 10^6 samples \times loci). As a result, existing applications of the approach have tended to consider smaller data sets (Kang et al. 2014; Blom et al. 2016a) or to ignore much of the available data (Blom et al. 2016b; Bragg et al. 2017; Stenson et al. 2017), which reduces accuracy and increases uncertainty in species-tree estimates (Song et al. 2012; Ogilvie et al. 2017). One approach to this problem has been the development of much simpler summary coalescent methods that utilize distributions of estimated gene-tree topologies as input to rapidly process large data sets (Liu et al. 2015). These include the rooted triplet method MP-EST (Liu et al. 2010) and the quartet method ASTRAL (Mirarab et al. 2014). However, summary coalescent methods are sensitive to gene-tree errors (Mirarab and Warnow 2015; Xi et al. 2015) and produce trees in coalescent units, and thus time and population size estimates used by downstream analyses are confounded.

Here, we aim to perform Bayesian inference on large data sets using the Markov chain Monte Carlo (MCMC) algorithm as our workhorse. As illustrated in Figure 1, the number of parameters involved is quite large, as is the accompanying state space. We develop a set of new MCMC proposals to explore state space in a much more efficient way than previous implementations and demonstrate we can handle data sets several times faster than *BEAST and StarBeast2. The resulting software package StarBeast3 is available as an open-source BEAST 2 package (Bouckaert et al. 2019).

METHODS

The MSC

Our objective is to develop efficient methods in a Bayesian framework for analyzing models where there is a phylogeny, S , such as a species or language tree, that forms a constraint on a set of k trees $\mathbf{G} = \{g_1, \dots, g_k\}$, such as gene trees. Each taxon within \mathbf{G} is assigned to a single taxon within S , from some fixed individual-to-species mapping function (Fig. 1). Species tree $S = (T_S, t_S)$ consists of a topology T_S and divergence times t_S , as does the set of gene trees $\mathbf{G} = (T_G, t_G)$.

All trees are assumed to be binary rooted time trees, where branch lengths describe the passing of time from the root of the tree down to the tips. Taxon node heights are assumed to be fixed and are typically extant (with height 0). Each gene tree g_i consists of $2n_{g_i} - 1$ nodes and $2n_{g_i} - 2$ branches for taxon count n_{g_i} , while S consists of $2n_S - 1$ nodes and $2n_S - 1$ branches, including a root branch, for species count n_S . Gene-tree taxa are associated with data $D = \{D_1, \dots, D_k\}$, for example, nucleotide sequences or cognate data. Let θ be a set of model parameters, for instance, those related to the speciation or nucleotide substitution processes. Consider the posterior density function $p(S, \mathbf{G}, \theta | D)$:

$$p(S, \mathbf{G}, \theta | D) = \underbrace{\frac{1}{Z}}_{\text{normalization constant}} \underbrace{p(S | \theta)}_{\text{species tree prior}} \times \underbrace{\prod_{i=1}^k p(g_i | S, \theta)}_{\text{gene tree prior}} \times \underbrace{\prod_{i=1}^k p(D_i | g_i, \theta)}_{\text{gene tree likelihoods}} \times \underbrace{p(\theta)}_{\text{parameter hyper prior}} \quad (1)$$

The MSC model is therefore hierarchical. S can follow a range of tree prior distributions $p(S|\theta)$, such as the Yule (Yule 1925) or birth–death models (Nee et al. 1994). Whereas, each gene tree g_i is assumed to follow the MSC process (Degnan and Rosenberg 2009; Heled and Drummond 2010; Jones 2017; Ogilvie et al. 2017; Rannala and Yang 2017), under which species-tree branches are associated with independently and identically distributed (effective) population sizes N_e which govern the coalescent process of G , where $|N_e| = 2n_S - 1$. Gene trees are thus assumed to be contained within S (Fig. 1).

Site evolution is assumed to follow a continuous-time Markov process (Felsenstein 1981) under some substitution model and clock model:

$$p(D_i|g_i, \theta) = p(D_i|g_i, \psi_i, \mathbf{r}, \mu). \quad (2)$$

ψ_i can adopt a range of molecular substitution models, such as the HKY nucleotide evolution model (Hasegawa et al. 1985) or the WAG amino acid evolution model (Whelan and Goldman 2001). Tree g_i has relative molecular substitution rate $v_i \in \psi_i$. Branches in S are associated with substitution rates \mathbf{r} , which govern the rate of site evolution of G along the respective branch, where $|\mathbf{r}| = 2n_S - 1$ (Fig. 1). Branch rates \mathbf{r} are assumed to be independently and identically distributed under a log-normal distribution with standard deviation σ (i.e., the MSC relaxed clock model; Drummond et al. 2006; Ogilvie et al. 2017). Lastly, the clock rate μ can be estimated when accompanied by time-calibration data, such as ancient fossil records (Sauquet et al. 2011; Heled and Drummond 2012; Ballesteros and Sharma 2019), or left fixed when no such data are available. Overall, the total substitution rate of any given branch in g_i is the product of v_i , μ , and a subset of the elements in \mathbf{r} (weighted by their coverage of the gene-tree branch; Ogilvie et al. 2017).

In this article, we develop tools that allow the MSC to be applied to large data sets using complex models of evolution. Although we focus on MSC models, we anticipate that in the future other models of the form expressed in Eq. (1) will be developed, for example, models that allow some lateral gene transfer and therefore allow some gene-tree branches to cross species boundaries in the species tree. We design a number of MCMC operators which generate proposals that explore the state space more efficiently—using a Gibbs sampler for population sizes, a combination of Bactrian (Yang and Rodríguez 2013; Thawornwattana et al. 2018) and adaptable variance multivariate normal (Baele et al. 2017) proposal kernels, a parallel operator for sampling gene trees and substitution model parameters, and an MCMC operator which selects other operators based on their exploration efficiency (Douglas et al. 2021b). Moreover, in the special case of the multispecies relaxed clock model (Ogilvie et al. 2017), we introduce methods for operating on the species tree, the gene trees, and the clock model simultaneously (Zhang and Drummond 2020; Douglas et al. 2021b).

Effective Population Size Gibbs Operator

The StarBeast2 (Ogilvie et al. 2017) and DISSECT (Jones et al. 2015) packages have the capability of integrating effective population sizes N_e when using an inverse gamma distributed prior on N_e , based on a technique introduced by (Liu et al., 2008) and detailed out by (Jones, 2017). This approach greatly reduces the state space. However, consequently the posterior Eq. 1 can no longer be broken down in a product over components over individual gene trees:

$$\prod_{i=1}^k p(g_i|S, \theta) = \int \prod_{N_e}^k p(g_i|S, N_e) dN_e. \quad (3)$$

Thus, the technique is not suitable for gene-tree operator parallelization, and therefore, we estimate N_e instead.

Suppose that $N_{e_b} \in N_e$, for species-tree branch b , follows an inverse gamma prior distribution $\text{Inv-}\Gamma(\alpha_N, \mu_N)$, where the shape α_N is fixed at 2 and therefore the scale μ_N is the expected value (because $\mathbb{E}(N_{e_b}) = \frac{\beta}{\alpha-1}$). Following the results by (Jones, 2017), the posterior of N_{e_b} follows an inverse gamma $\text{Inv-}\Gamma(\alpha_{N'}, \mu_{N'})$, such that $\alpha_{N'} = \alpha_N + a$ and $\mu_{N'} = \mu_N + c$ where a is the total number of coalescent events of all gene trees in branch b and $c = \sum_j \frac{1}{p_j} \sum_i c_{jbi} \binom{n_{jb}-i}{2}$. Here, p_j is the ploidy of gene g_j , c_{jbi} the size of the i th coalescent interval for gene g_j in branch b , and n_{jb} the number of lineages of gene tree g_j at the tip-side of branch b (so that $n_{jb} - i$ is the number of lineages at the start of the i th coalescent interval for g_j).

Instead of integrating N_e , our GibbsPopulation operator samples from the posterior. All $2n_S - 1$ elements in N_e are proposed simultaneously. As demonstrated later, this turns out to be more efficient than standard N_e random walk operators, with the added advantage of sampling effective population sizes—which may be a parameter of interest—as well as the ability to parallelize gene-tree proposals. This technique is readily applicable for periodically sampling and logging N_e to implementations that do integrate this term out.

Bactrian Operators for Trees

The step size of a proposal kernel should be such that the proposed state x' is sufficiently far from the current state x to explore vast areas of parameter space, but not so far that the proposal is rejected too often (Gelman et al. 1997). The Bactrian distribution (Yang and Rodríguez 2013; Thawornwattana et al. 2018) has minimal probability mass around the center, and a higher concentration flanking the center, akin to the humps of a Bactrian camel (Fig. 2; left). This distribution is a preferred alternative to standard uniform- or normal-distributed random walk kernels, as it places minimal probability on step sizes that are too large or too small, and has successfully improved phylogenetic inference in

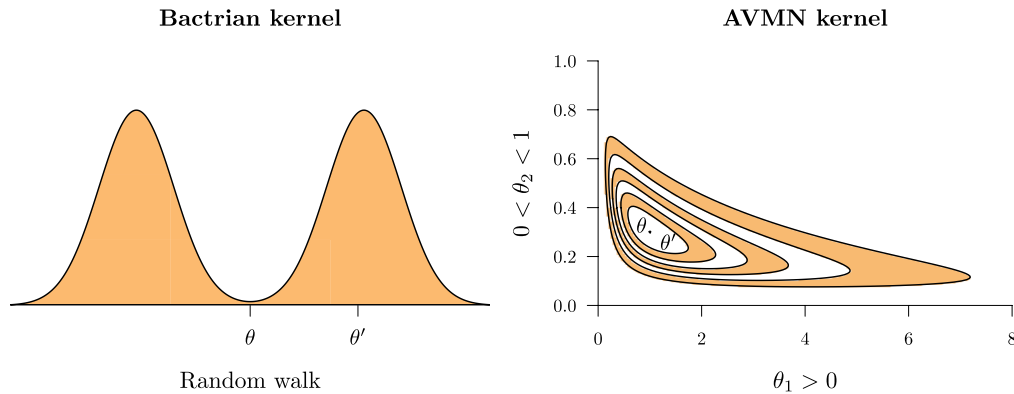


FIGURE 2. Depiction of random walks $\theta \rightarrow \theta'$ under varying proposal kernels. Left: The random walk occurs from the origin between the two modes, where the vertical axis shows the probability density function of the kernel $p(\theta'|\theta)$ (Yang and Rodríguez 2013). Right: A 2D random walk on inversely correlated parameters $\theta = (\theta_1, \theta_2)$ with different domains (Baele et al. 2017). Contours describe the joint probability density function $p(\theta_1, \theta_2)$ under a transformed multivariate normal distribution learned during MCMC.

previous studies (Yang and Rodríguez 2013; Zhang and Drummond 2020; Douglas et al. 2021b).

In this article, we apply Bactrian proposals to trees. The standard set of tree node height proposals in BEAST 2 consists of a `Scale` operator which embarks all nodes in the tree on a random walk (in log-space), a `RootScale` operator which does so for only the root of a tree, an `UpDown` operator which changes species/gene node heights and various continuous parameters simultaneously (Drummond et al. 2002), a `SubtreeSlide` operator which slides a node up or down branches (Hohna et al. 2008), and constant distance operators when a relaxed clock model is applied (Zhang and Drummond 2020). Each operator would normally draw a random variable from a uniform distribution, but here we instead use a Bactrian distribution and apply appropriate transformations. We also introduce the `Interval` operator, which transforms parameters with lower- and upper-bounds (such as tree node heights) by applying a Bactrian random walk in their real-space transformations.

Adaptive Variance Multivariate Normal Operator

An adaptive variance multivariate normal (AVMN) operator (Baele et al. 2017) provides proposals for a set of real-space parameters by learning the posterior throughout the run of the MCMC algorithm and approximating it as a multivariate normal distribution to capture correlations between parameters (Fig. 2; right). The space spanned by such a set of continuous parameters may need to be transformed (in order to satisfy the assumption that all parameters lie in real-space), by applying a log-transformation to parameters with positive domains (such as substitution rates), or a log-constrained sum transformation to multivariate parameters with unit sums (such as nucleotide frequencies), for instance. AVMN has been demonstrated to be more efficient in estimating phylogenetic parameters than standard random walk or scale operators

(Baele et al. 2017; Bouckaert 2020; Douglas et al. 2021b).

Consider a single gene tree g_i and its substitution model ψ_i , consisting of substitution rates and nucleotide frequencies for instance. Performing a single proposal for any single parameter would require a full recalculation of the tree likelihood $p(D_i|g_i, \psi_i, \mathbf{r}, \mu)$ (see peeling algorithm by (Felsenstein, 1981)). Therefore, proposing all site model parameters ψ_i simultaneously can reduce the number of likelihood calculations required and thus lower the computational runtime.

Parallel Gene-Tree Operator

During MCMC, operators are typically sampled proportionally to fixed weights (or proposal probabilities), to ensure the chain is ergodic. Here, we present an alternative method, where a single gene tree g_i and its substitution model ψ_i is selected, and N_p operators are sequentially sampled and applied to g_i and ψ_i , before returning to the full parameter space. This is equivalent to running a small MCMC chain of N_p steps—applying only gene tree and substitution model operators on g_i and ψ_i —and then accepting the resulting g_i' and ψ_i' afterwards with probability 1, as if it were a single Gibbs sampling operation (Geman and Geman 1984).

Observe that because only g_i and its associated parameters change, part of Eq. (1) can be rewritten as:

$$\begin{aligned} & \underbrace{\prod_{i=1}^k p(D_i|g_i, \psi_i, \mathbf{r})}_{\text{gene tree likelihoods}} \times \underbrace{\prod_{i=1}^k p(g_i|S, \mathbf{N}_e)}_{\text{gene tree priors}} \\ & \quad \underbrace{\hspace{10em}}_{\text{gene tree posteriors}} \\ & = \prod_{i=1}^k p(D_i|g_i, \psi_i, \mathbf{r}) \times p(g_i|S, \mathbf{N}_e). \end{aligned} \quad (4)$$

Thus, the posterior distribution can be decomposed into the product of contributions of individual gene trees and their substitution models. Assuming that substitution model parameters ψ_i are distinct for each gene tree g_i , an N_p -step MCMC chain could be run for each of g_i and g_j for ($i \neq j$) in parallel, and the resulting g_i' and g_j' each accepted with probability 1, as if two Gibbs operators were sequentially applied. Because the posterior density for g_i is proportional to $p(D_i|g_i, \theta)p(g_i|S, \theta)$ and that of g_j proportional to $p(D_j|g_j, \theta)p(g_j|S, \theta)$ then provided that any shared parameters (such as \mathbf{r} , S , and \mathbf{N}_e) are not being operated on, these two N_p -step MCMC chains can run in parallel.

Where there are N_t threads available, the k gene trees are split into N_t groups (assuming $k \geq N_t$). The N_t sets of N_p -step MCMC chains are run in parallel and the resulting gene trees \mathbf{g} are accepted into the main MCMC chain. Here, we introduce a parallel operator `ParallelGeneTreeOperator`(\mathbf{G}, ψ). This operator partitions gene trees into N_t threads and operates on their topologies, node heights, and substitution models. Tree node height proposals employ the Bactrian kernel where applicable (Fig. 2), and substitution model proposals invoke the AVMN kernel (Fig. 2). The chain length N_p of each thread is learned during MCMC (Fig. 3).

Since each small MCMC chain for a thread can be considered a single Gibbs proposal, for N_t threads in principle N_t steps should be added to the main chain. If the operator is selected just before logging a state, in principle some threads may need to be disregarded before logging in order to maintain exactly equal intervals in the trace log. Due to the low frequency at which the operator is selected, and the logging intervals being orders of magnitude larger than the number of threads, this does not appear to be a problem in practice.

Species Tree Relaxed Clock Model Operators

The constant distance operator family exploits the negative correlations between divergence times and branch substitution rates by proposing both terms simultaneously (Zhang and Drummond 2020). This technique has yielded a parameter convergence rate of one to two orders of magnitude faster, particularly for large data sets that come with peaked posterior distributions (Douglas et al. 2021b). Under the MSC relaxed clock model used by StarBeast2, the branch rate of gene-tree branch b is the length-weighted branch rate \mathbf{r} of all species-tree branches that contain b (Ogilvie et al. 2017). Moreover, effective population sizes \mathbf{N}_e are positively correlated with divergence times, so this correlation could also be readily exploited.

Extending the work by (Zhang and Drummond, 2020), we introduce the `ConstantDistanceMSC` operator. This operator proposes a node height t_X for species-tree internal node X , the three branch rates (elements of \mathbf{r}) and population sizes (elements of \mathbf{N}_e) incident to X , and heights for all gene-tree non-leaf nodes that are

contained within these three incident branches (Fig. 4). t_X is embarked on a Bactrian random walk (Yang and Rodríguez 2013) to give t'_X , then \mathbf{r} and the node heights in \mathbf{G} are proposed such that all genetic distances are conserved following the change in t_X , and \mathbf{N}_e is proposed such that the positive correlation between itself and the branch lengths incident to X is respected (see Algorithm S1).

Previously, we introduced the narrow exchange rate (NER) operator (Douglas et al. 2021b). This operator combined the simple `NarrowExchange` operator (i.e., a proposal which swaps a subtree with its uncle subtree; (Drummond et al., 2002)) with the `ConstantDistance` operator (Zhang and Drummond 2020), by applying a small topological change to the tree and then recomputing branch substitution rates such that evolutionary distances are preserved. We demonstrated that this operator assisted the traversal of tree topology space on longer alignments compared with shorter ones.

Here, we combine this work with the `CoordinatedExchange` operator implemented by (Ogilvie et al., 2017)—based on work by (Jones, 2017) and (Rannala and Yang, 2017)—and introduce the coordinated narrow exchange rate (CNER) operator. This operator exchanges a species-tree node with its uncle node adjusts gene-tree topologies \mathbf{g} to preserve compatibility with S , and proposes three nearby branch rates in \mathbf{r} to preserve genetic distances (Algorithm S2).

Adaptive Operator Weighing

Previously, we developed the `AdaptableOperatorSampler`(x) operator (Douglas et al. 2021b). This operator learns the weights (or proposal probabilities) behind a set of suboperators during MCMC, by rewarding operators which bring about large changes to parameter x in short computational runtime, with respect to some distance function: Euclidean distance when x is real, and RNNI distance (Collienne and Gavryushkin 2021) when x is tree topology. This approach can account for the scenario when an operator's performance is conditional on the data set. When a data set contains very little signal with respect to a certain parameter x and its prior distribution, then resampling that parameter from its prior distribution using the `SampleFromPrior`(x) operator may be more efficient than embarking x on a random walk, for instance (Douglas et al. 2021b). In contrast, data sets with more signals are likely to prefer smarter operators which account for correlations in the posterior distribution, such as the constant distance or NER operators (Zhang and Drummond 2020; Douglas et al. 2021b).

Here, we have applied the `AdaptableOperatorSampler` to seven areas of parameter space: the species and gene-tree node heights (t_s and t_G), the relaxed clock model rates \mathbf{r} and standard deviation σ , the mean effective population size μ_N , the species-tree birth rate λ (assuming a Yule speciation model; Yule 1925), and

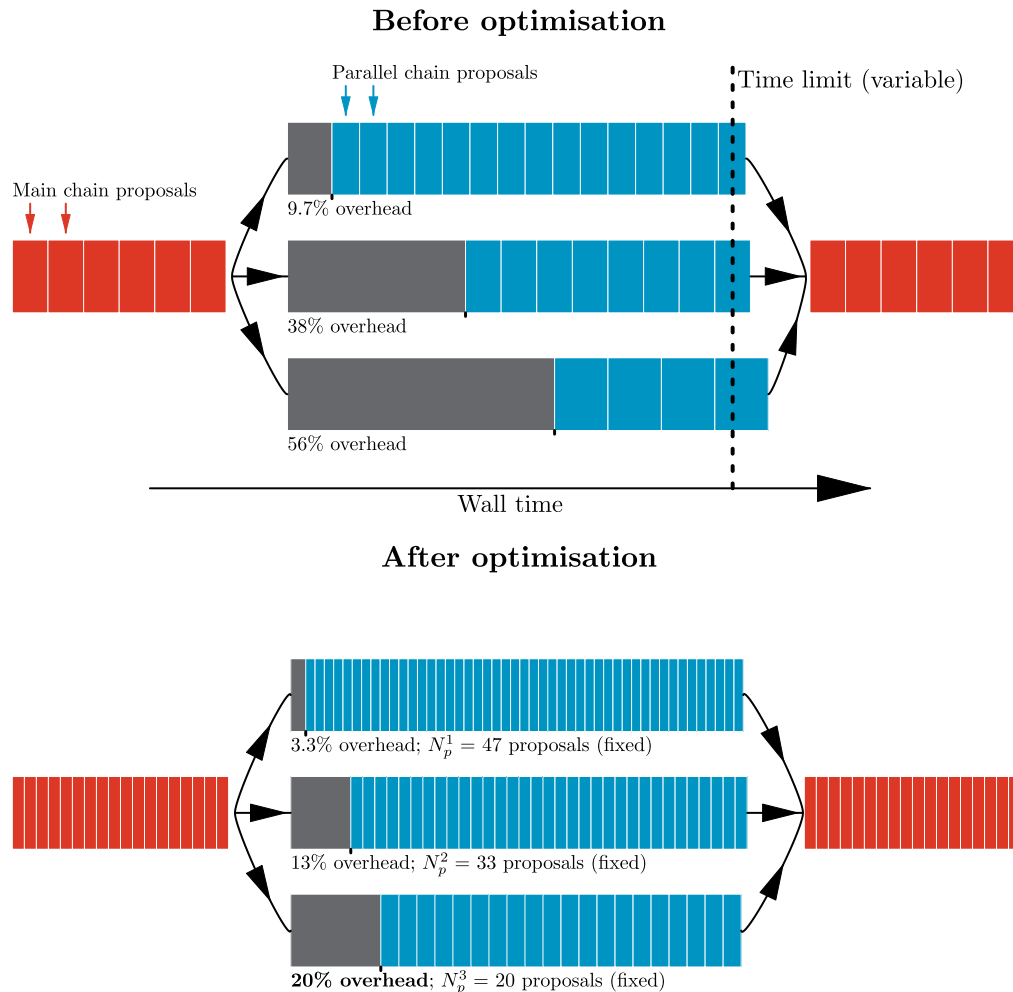


FIGURE 3. Optimization of gene-tree parallel operator chain lengths. Top: The time limit of each parallel MCMC chain is randomized on each call so that the overhead (intercept) and time-per-proposal (slope) can be learned as a linear regression model. Bottom: The linear regression model is applied, and parallel MCMC chain lengths are set such that the slowest thread attains the user-specified target overhead (i.e., the bottom thread has attained 20% overhead in the example above).

the species-tree topology T_S . These operator schemes are explicated in Tables 1 and 2.

RESULTS

In this section, we first validate the correctness of StarBeast3 through a well-calibrated simulation study. Then, we demonstrate that StarBeast3 is efficient at doing Bayesian inference on large data sets compared with StarBeast2. We did not compare to *BEAST directly, since it does not provide relaxed clock models on species trees, but note that (Ogilvie et al., 2017) benchmarked StarBeast2 against *BEAST for strict clocks and found StarBeast2 to be an order faster than *BEAST, so any gain over StarBeast2 will be more so over *BEAST.

Validation

In order to validate the correctness of StarBeast3, we performed two well-calibrated simulation studies.

These were achieved by simulating nucleotide alignments (of two varying sizes) using parameters directly sampled from the prior distribution, and then recovering the posterior estimates of these parameters by doing Bayesian inference on the simulated alignments using StarBeast3. For each study, the 95%-coverage of each parameter was approximately 95% (meaning that the true parameter estimate was within the 95% highest posterior density interval approximately 95% of the time). Therefore, these experiments provide confidence in StarBeast3's correctness and are presented in Figure 5 and Section S4 of Supplementary material available on Dryad at <http://dx.doi.org/10.5061/dryad.f1vhhmgzk>.

Performance Benchmarking

We evaluated the performance of StarBeast3 for its ability to achieve multispecies coalescent parameter convergence in a Bayesian framework, compared with

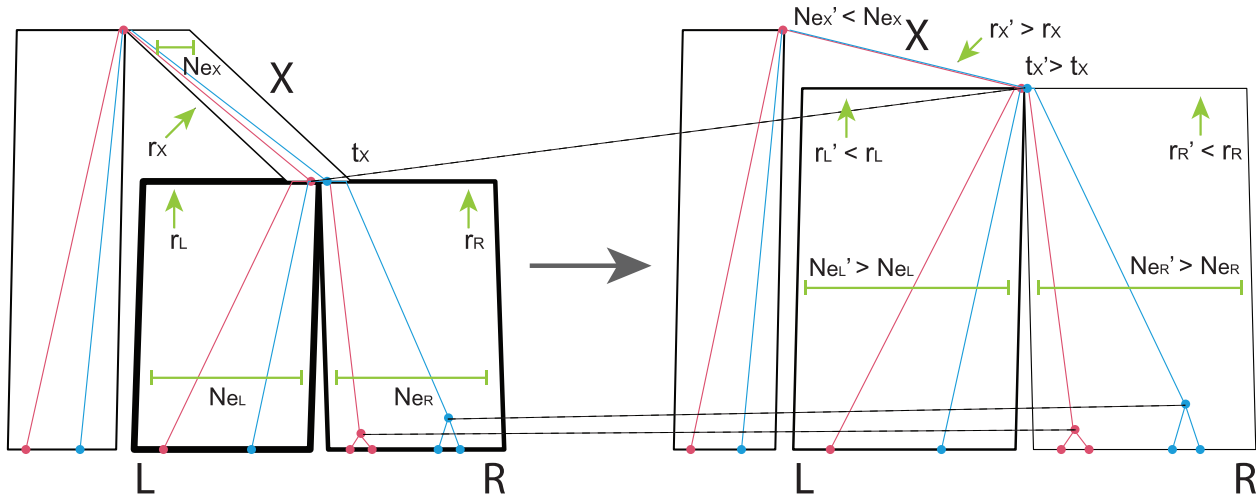


FIGURE 4. An example of an ConstantDistanceMSC proposal, acting on species nodes X and its two children L and R . First, the height of X (t_X) is increased to t'_X . Then, the relative substitution rates of branches L (r_L) and R (r_R) are decreased to r'_L and r'_R , and r_X is increased to r'_X . These compensations in branch length ensure that the genetic distance of each branch (X , L , and R) is maintained. The thicknesses of the species node lines are proportional to these substitution rates. Finally, the effective population size of L and R are increased to N_{eL}' and N_{eR}' , while that of X is decreased to N_{eX}' . These compensations in node height ensure that the ratio between branch length and branch population size are maintained. Species node widths are proportional to their effective population size. During this operation, gene-tree nodes always remain constrained by the species tree. Figure was generated by UglyTrees (Douglas 2020).

that of StarBeast2. Although it is a nontrivial problem to determine if an MCMC chain has converged, the effective sample size (ESS) can serve as a useful metric. Thus, we computed the number of effective samples generated per hour (ESS/h) across multiple replicates of MCMC, using three real and two simulated data sets (Table 3). The ESS of any parameter should be over 200 in order to estimate its posterior distribution (Tracer; Rambaut et al. 2018). To allow both software packages to perform at their best, effective population sizes were integrated by StarBeast2, but were estimated by StarBeast3. This section provides a general comparison of StarBeast3 and StarBeast2; however, the performances of individual operators can be found in Sections S5 and S6 of Supplementary material available on Dryad.

The ESS/h was evaluated in five distinct areas of parameter space. First, we considered generic summaries of convergence: the ESS/h of the posterior density $p(\theta|D)$, the likelihood $p(D|\theta)$, and the prior density $p(\theta)$. Second, species tree S convergence was evaluated in terms of its height h_S , its length l_S , and hyperparameters λ —the Yule model birth rate (Yule 1925)—and μ_N —the mean effective population size. In the case of StarBeast3, where effective population sizes are estimated, we also measured the mean ESS/h associated with species-tree leaf nodes of N_e . Third, gene-tree convergences were evaluated by their heights h_G , their lengths l_G , and the RNNI distances (Collienne and Gavryushkin 2021) to their UPGMA D_{UPGMA} (Sokal 1958) and neighbor-joining D_{NJ} trees (Saitou and Nei 1987). As there are multiple gene trees, we only considered the mean ESS/h of each term. Fourth, substitution model convergence (HKY substitution model; Hasegawa et al. 1985) was measured from the transition–transversion ratio κ , nucleotide frequencies f , and gene-tree substitution rates

v , where the ESS/h of each term was averaged across all k substitution models. Lastly, relaxed clock model convergence was evaluated by considering the mixing of branch rate empirical mean \bar{r} and variance $\text{var}(r)$, as well as the relaxed clock standard deviation parameter σ .

These results showed that, depending on the data set, the “slowest” parameter generally converged considerably faster for StarBeast3 than it did for StarBeast2 (see the min term in Figs. 6 and 7). On the smallest data set considered (Frog), StarBeast2 and 3 performed comparably well overall (and no significant difference in min). However, StarBeast3 performed better on all of the other data sets, with the “slowest” parameter converging between 4 and 37 \times as fast, and the posterior density $p(\theta|D)$ converging between 2 and 36 \times as fast, often at a statistically significant level. For StarBeast3, the absolute time needed to converge varied a lot across the data sets, and even across multiple replicates of the same data set (see final column of Table 3). The fastest data sets—Frog and Simulated(12)—required 1–2 days to converge, while the Spider data set required over a month.

Notably, relaxed clock model parameters converged up to 35 \times as fast under StarBeast3. This was credited to the use of a real-space branch rate parameterization (where branch rates are real numbers as opposed to discrete bins, as implemented in StarBeast2) as well as constant distance operators, which adjust branch rates and divergence times simultaneously (Zhang and Drummond 2020; Douglas et al. 2021b). The disparity between StarBeast3 and StarBeast2 was less extreme for the smaller $k=26$ gene tree Frog data set (Barrow et al. 2014), consistent with previous experiments (Douglas et al. 2021b).

Substitution model parameters ψ generally converged faster for StarBeast2 than they did for StarBeast3. Note,

TABLE 1. StarBeast3 operator scheme, assuming a Yule tree prior on the species tree with birth rate λ .

Operator	Weight	Reference
<i>Species tree</i>		
NodeReheight (T_S, t_S, T_G, t_G)	30	Ogilvie et al. (2017)
CoordinatedUniform (t_S, t_G)	30	Ogilvie et al. (2017), Jones (2017)
CoordinatedExponential (t_S, t_G)	15	Ogilvie et al. (2017), Jones (2017)
SubtreeSlide (T_S, t_S) ^a	15	Hohna et al. (2008)
WilsonBalding (T_S)	15	Drummond et al. (2002)
WideExchange (T_S)	15	Drummond et al. (2002)
AdaptableOperatorSampler (T_S)	15	
NarrowExchange (T_S)		Drummond et al. (2002)
CoordinatedExchange (T_S, T_G)		Ogilvie et al. (2017)
NER (T_S, r)		Douglas et al. (2021b)
CNER (T_S, r, T_G)		<i>Species Tree Relaxed Clock Model Operators</i>
Uniform (t_S) ^a	3	
RootScale (t_S) ^a	3	
Interval (t_S) ^a	3	<i>Bactrian Operators for Trees</i>
AdaptableOperatorSampler (t_S)	100	
Uniform (t_S) ^a		
TreeScale (t_S) ^a		
Interval (t_S) ^a		<i>Bactrian Operators for Trees</i>
ConstantDistanceMSC (t_S, t_G, r, N_e) ^a		<i>Species Tree Relaxed Clock Model Operators</i>
CoordinatedUniform (t_S, t_G)		Ogilvie et al. (2017), Jones (2017)
CoordinatedExponential (t_S, t_G)		Ogilvie et al. (2017), Jones (2017)
UpDown ($[t_S, t_G, N_e, \mu_N], [v, \lambda, \mu]$) ^a		Drummond et al. (2002)
<i>Gene trees/site models</i>		
ParallelMCMC (T_G, t_G, ψ)	3.42	Table 2
<i>Tree hyperparameters</i>		
GibbsPopulation (N_e)	50	<i>Effective Population Size Gibbs Operator</i>
AdaptableOperatorSampler (μ_N)	5	
Scale (μ_N) ^a		
UpDown ($[t_S, t_G, N_e, \mu_N], [v, \lambda, \mu]$) ^a		Bouckaert et al. (2019)
SampleFromPrior (μ_N)		Douglas et al. (2021b)
AdaptableOperatorSampler (λ)	5	
Scale (λ) ^a		
UpDown ($[t_S, t_G, N_e, \mu_N], [v, \lambda, \mu]$) ^a		Bouckaert et al. (2019)
SampleFromPrior (λ)		Douglas et al. (2021b)
<i>Relaxed clock model</i>		
AdaptableOperatorSampler (r)	30	
Scale (r) ^a		
ConstantDistanceMSC (t_S, t_G, r, N_e) ^a		<i>Species Tree Relaxed Clock Model Operators</i>
SampleFromPrior (r)		Douglas et al. (2021b)
AdaptableOperatorSampler (σ)	5	
Scale (σ) ^a		
SampleFromPrior (σ)		Douglas et al. (2021b)

Notes: The ParallelMCMC operator weight was set such that it is sampled 1% of the time. Further operator details can be found in Drummond and Bouckaert (2015).

^aBactrian kernel applied to random walk (Yang and Rodríguez 2013).

however, that this is by design. The total operator weight assigned to ψ parameters was 50% smaller in StarBeast3, in order to ensure balanced convergence across all areas of parameter space. In all data sets considered, substitution models converged significantly faster than any other area of parameter space, despite receiving relatively little operator weight, and therefore computational resources that were being spent on the substitution model were better off spent in “slower” areas of parameter space, such as gene-tree node heights.

The AdaptableOperatorSampler operators (Table 1) confirmed the value in the NER and ConstantDistanceMSC operators for operating on their respective areas of parameter space. The ConstantDistanceMSC operator almost always

outperformed other operators at proposing species node heights t_S (Table 4). The exception to this was the Skink data set, for which the UpDown operator was superior at proposing branch lengths, and the Frog data set, for which ConstantDistanceMSC, CoordinatedExponential, and UpDown were all on a par. In general, very little operator weight was rewarded to the Uniform, Interval, TreeScale CoordinatedUniform, and CoordinatedExponential operators for their abilities to propose species node heights. Similarly, among NarrowExchange variants evaluated by AdaptableOperatorSampler (T_S), the NER operator was marginally favored by all data sets (Table 5). This was due to the operator making larger or more

TABLE 2. StarBeast3 parallel operator scheme for gene trees and their associated site models (assumed to be an HKY model with transition–transversion ratio κ and nucleotide frequencies f)

Operator	Weight	Reference
ParallelMCMC (T_G, t_G, Ψ)		<i>Parallel Gene-Tree Operator</i>
Gene trees		
$\forall_{i \in 1, \dots, k}$ WilsonBalding (T_{g_i})	15	Drummond et al. (2002)
$\forall_{i \in 1, \dots, k}$ WideExchange (T_{g_i})	15	Drummond et al. (2002)
$\forall_{i \in 1, \dots, k}$ NarrowExchange (T_{g_i})	15	Drummond et al. (2002)
$\forall_{i \in 1, \dots, k}$ SubtreeSlide (T_{g_i}, t_{g_i}) ^a	10	Hohna et al. (2008)
$\forall_{i \in 1, \dots, k}$ Uniform (t_{g_i})	30	
$\forall_{i \in 1, \dots, k}$ RootScale (t_{g_i}) ^a	10	
$\forall_{i \in 1, \dots, k}$ Interval (t_{g_i}) ^a	10	
$\forall_{i \in 1, \dots, k}$ AdaptableOperatorSampler (t_{g_i})	100	<i>Bactrian Operators for Trees</i>
TreeScale (t_{g_i}) ^a		
Uniform (t_{g_i}) ^a		
SubtreeSlide (T_{g_i}, t_{g_i}) ^a		Hohna et al. (2008)
EpochOperator (t_{g_i}) ^a		Bouckaert (2021)
Site models		
$\forall_{i \in 1, \dots, k}$ AVMN (ψ_i, t_{g_i})	5	Baele et al. (2017)
$\forall_{i \in 1, \dots, k}$ Scale (κ_i) ^a	0.5	
$\forall_{i \in 1, \dots, k}$ Scale (ν_i) ^a	0.5	
$\forall_{i \in 1, \dots, k}$ DeltaExchange (f_i)	0.5	

Notes: Each operator is applicable to a single gene tree g_i or its site model ψ_i . AVMN (ψ_i, t_{g_i}) generated proposals for the site model and complete set of tree node heights simultaneously. Operator weights are normalized into proposal probabilities within a single MCMC chain called by ParallelMCMC. Further operator details can be found in (Drummond and Bouckaert, 2015).

^aBactrian kernel applied to random walk (Yang and Rodríguez 2013).

frequent topological changes to the species tree, in faster computational runtime, especially compared with CoordinatedNarrowExchange and CNER. Overall, this experiment reinforced the value of learning operator weights on a problem-by-problem basis. A full breakdown of the remaining four adaptive operators can be found in Section S6 of Supplementary material available on Dryad.

Lastly, we evaluated the effect of threading on StarBeast3, by comparing its performance under 1, 2, 4, 8, and 16 threads allotted to the ParallelMCMC gene-tree operator (Fig. 8). There was a positive-but-modest correlation between the number of threads and the overall rate of convergence among the terms considered, with an overall log-linear slope coefficient of 0.19. This can be interpreted as follows: across the range of threads and data sets considered, doubling the number of threads was associated with an increase in mixing by 14%. Multithreading provided the strongest boost for the Skink and Spider data sets and made little difference to the simulated data set (48 taxa). This is an unexpected result, because the Skink and Spider data sets have fewer genes ($k=50$ compared with $k=100$), and may be due to the former data sets having more taxa and thus larger trees.

Benchmarking on Large Data Sets

We benchmarked the performance of StarBeast3 on simulated data sets with increasingly large numbers of gene trees k , ranging from 250 to 1000 genes. Each gene was 200 nucleotides in length. In order to achieve convergence in a timely manner, we performed inference under a strict clock model (i.e., all branch rates fixed $r=1$) and with a small sample size ($n_s=4$ species $n_G=12$ taxa). These experiments showed that StarBeast3

required more time to produce one sample for larger k , and therefore more time to produce one effective sample, as expected (Fig. 9). The $k=250$ gene data set would require ≈ 10 h for the average ESS to exceed 200 in all areas of parameter space, while the $k=1000$ gene data set would require ≈ 40 h. Furthermore, we confirmed that gene-tree parallelization gave a noticeable-but-modest improvement to runtime (Fig. 9). Although the trees were small, this experiment showed that StarBeast3 is indeed capable of running on large data sets with several hundred genes.

DISCUSSION

The Next Generation of Bayesian MCMC Operators

In recent years, Bayesian MCMC proposals have advanced significantly beyond that of the unidimensional random walk. The use of adaptive algorithms and advanced proposal kernels have become increasingly prevalent (Haario et al. 2001; Vihola 2012; Yang and Rodríguez 2013; Benson and Friel 2018). In phylogenetic inference in particular, tree proposals have been guided by conditional clade probabilities and parsimony scores (Höhna and Drummond 2012; Zhang et al. 2020), and mirror kernels learn target distributions which act as “mirror images” (Thawornwattana et al. 2018), for instance.

Here, we introduced a range of recently developed MCMC operators to the MSC, including Bactrian proposal kernels (Yang and Rodríguez 2013), which have been successfully applied to bird phylogeny (Maliot et al. 2019), and tree “flex” operators (BICEPS; Bouckaert 2021), which have been applied to coronavirus disease-2019 genomic data (Douglas et al. 2021a). We also

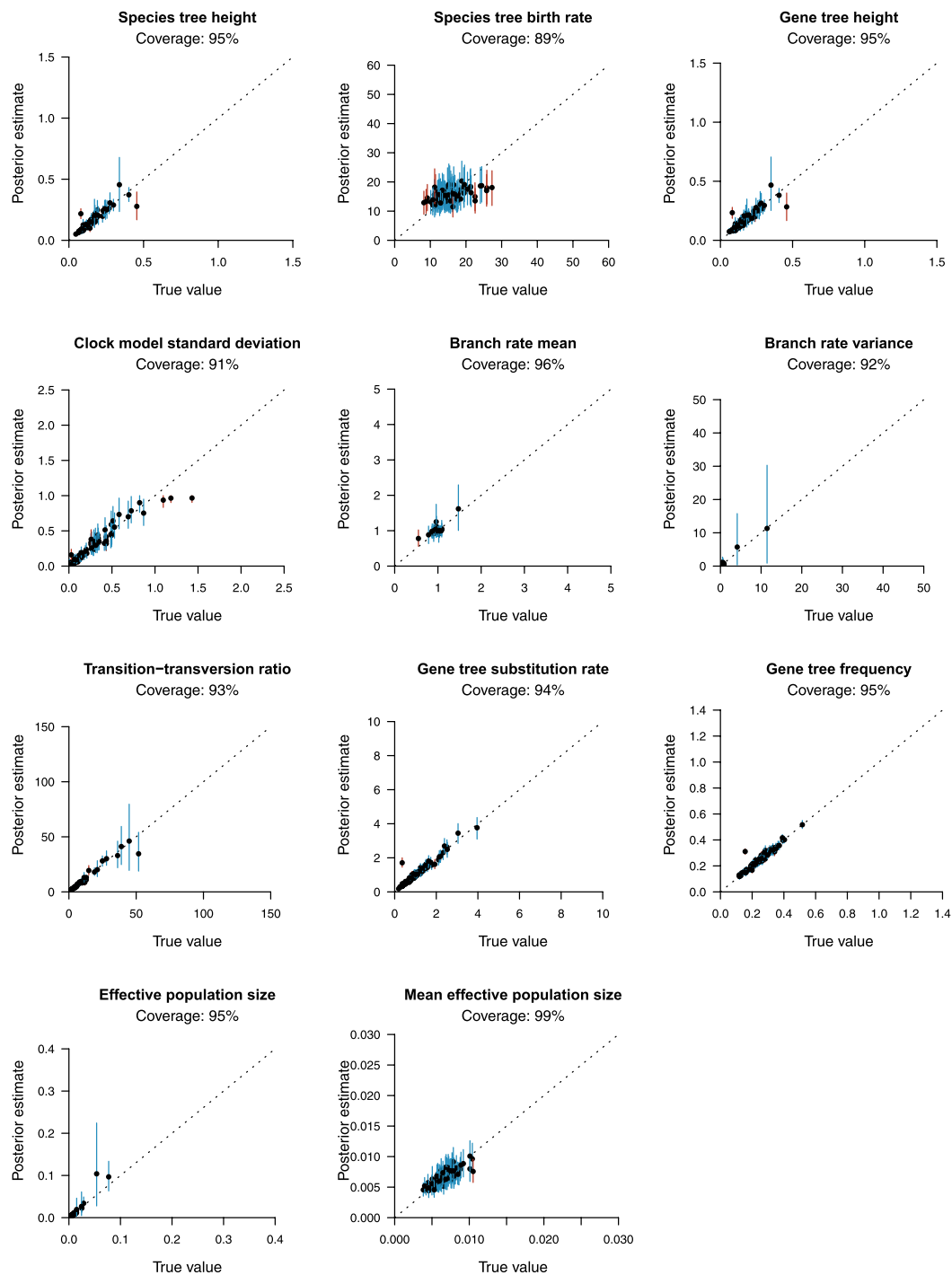


FIGURE 5. Well-calibrated simulation study analyzing $n_S = 16$ species, $n_G = 48$ taxa, and $k = 50$ genes. One-hundred simulations were performed to recover the coverage between “true” simulated values and their estimates under the posterior distribution. 95% highest posterior density (HPD) intervals of parameters are represented by vertical lines. Each line represents a single simulation, and is colored blue when the true value was contained within the 95% interval, or red otherwise. The top of each plot shows the coverage of each parameter (i.e., the number of MCMC simulations for which the “true” parameter value was contained within the 95% HPD).

invoked a series of more meticulous operators which account for known correlations, such as the AVMN kernel (Baele et al. 2017), constant distance operators (Zhang and Drummond 2020), and the NER operator

(Douglas et al. 2021b), as well as adaptive operators that improve over the course of MCMC, such as the adaptable operator sampler (Douglas et al. 2021b), parallel gene-tree operators, and the AVMN kernel

TABLE 3. Benchmark data sets

Data set	No. of species n_S	No. of taxa n_G	No. of gene trees k	Time (h)
Frog (Barrow et al. 2014)	21	88	26	25–41
Skink (Bryson Jr et al. 2017)	10	59	50	30–54
Spider (Hamilton et al. 2016)	36	83	50	660–1100
Simulated (12)	4	12	100	24–100
Simulated (48)	16	48	100	440–950

Notes: Fifty gene trees were subsampled from the Skink and Spider data sets. The simulated data sets were directly sampled from the model specification used during Bayesian inference (described in Section S3 of Supplementary material available on Dryad). In the final column, we estimate the time required for the MCMC chain to converge using StarBeast3 with 16 threads (min–max across 5 MCMC replicates). These terms were approximated as the time to achieve an effective sample size of 200 for the posterior density $p(\theta|D)$, with a 20% burn-in.

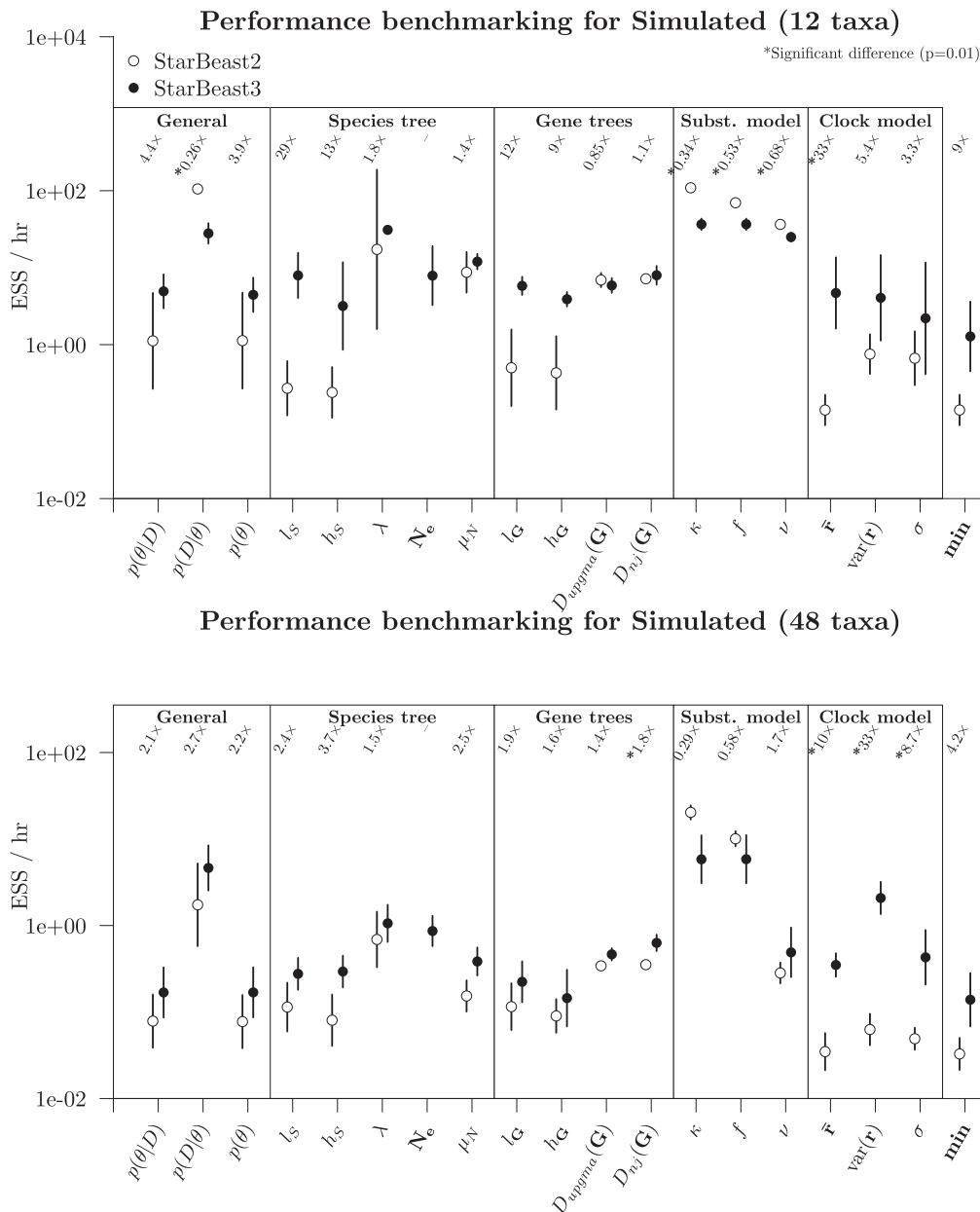


FIGURE 6. Performance benchmarking the two simulated data sets. Each point is the geometric-mean ESS/h across five replicates, for either StarBeast2, or StarBeast3 with 16 threads. The geometric-mean relative performance of StarBeast3, compared with StarBeast2, is indicated above each term, and a * is present if the difference across five replicates is significant according to a Student's t -test. Note that the y -axis is in log-space.

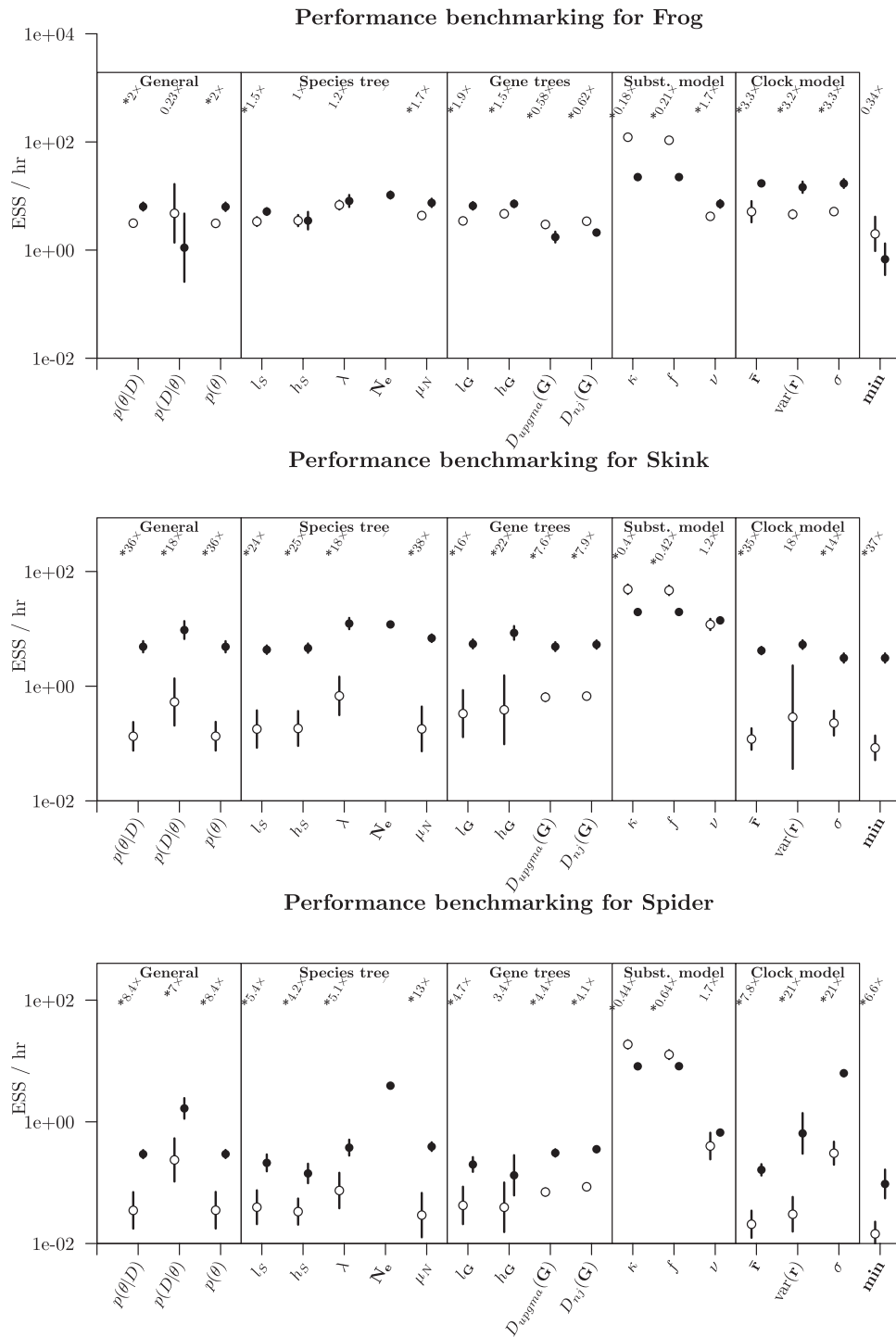


FIGURE 7. Performance benchmarking the two biological data sets. See Figure 6 caption for figure notation.

(Baele et al. 2017). Indeed, these operators have yielded a software package which outperforms StarBeast2 by up to one-and-a-half orders of magnitude, depending on the data set and the parameter.

While StarBeast3 provides a clear advancement to the problem, Bayesian MCMC is still lagging behind the volumes of next-generation genomic data. Therefore,

the continued development of efficient, meticulous, and adaptive MCMC operators is essential.

Efficient Parallelized Bayesian Inference under the MSC

As genomic data becomes increasingly available, concatenating genomic sequences and inferring the phylogeny of the species as that of the genes can

TABLE 4. Learned weights of the suboperators of `AdaptableOperatorSampler` (t_S), averaged across five replicates

Data set	Uniform	Interval	ConstantDistanceMSC	TreeScale	CUnif	CExp	UpDown
Frog	0.06	0.0078	0.34 ^a	9.8e-05	0.04	0.22	0.33
Simulated (12)	1.1e-05	0.00013	0.99 ^a	2.6e-05	0.00043	8e-04	0.0061
Simulated (48)	4.3e-05	0.00055	0.98 ^a	5.1e-06	0.0011	0.00029	0.016
Skink	0.008	0.0087	0.34	4.8e-05	0.013	0.04	0.59 ^a
Spider	0.0019	0.0034	0.84 ^a	1.5e-05	0.0063	0.0025	0.15

Notes: The operator which attained the highest proposal probability is indicated by ^a.

TABLE 5. Average species tree RNNI distance between before and after each proposal/operator runtime for the suboperators of `AdaptableOperatorSampler` (T_S), averaged across five replicates

Data set	NE	NER	CNE	CNER
Frog	0.0091/0.29 ms	0.0091/0.29 ms ^a	0.0091/0.4 ms	0.0091/0.39 ms
Simulated (12)	0.003/0.091 ms	0.0032/0.094 ms ^a	0.0028/0.19 ms	0.0028/0.19 ms
Simulated (48)	0.00043/0.77 ms	0.00043/0.62 ms ^a	0.00047/1 ms	0.00047/0.83 ms
Skink	0.021/0.3 ms	0.021/0.3 ms ^a	0.021/0.5 ms	0.021/0.48 ms
Spider	0.019/1.6 ms	0.019/1.2 ms ^a	0.019/1.8 ms	0.019/1.3 ms

Notes: Note that the timer starts at the beginning of the proposal and ends when the proposal has accepted or rejected. NE = narrow exchange; NER = narrow exchange rates; CNE = coordinated narrow exchange; CNER = coordinated narrow exchange rates. The operator which was rewarded the highest proposal probability for each data set is indicated by ^a.

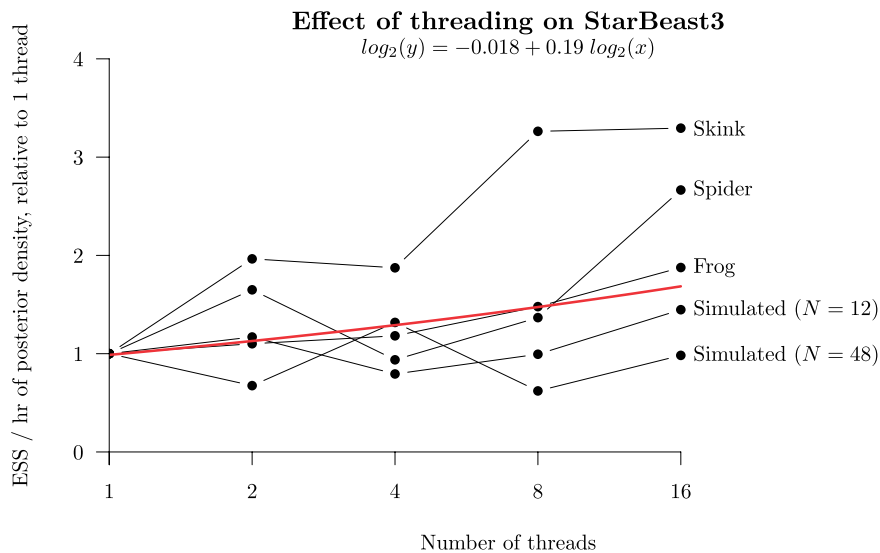


FIGURE 8. Effect of threading on StarBeast3 performance. Each point represents the ESS/h of the posterior density $P(\theta|D)$ (averaged across five replicates), for the indicated thread count and data set. These terms are normalized to enable comparison across data sets, by dividing it by that of one thread. A linear model was fitted to the ESS/h and number of threads, each in \log_2 space, and is reported at the top of the plot. The positive coefficient of the slope indicates that performance increased with the number of threads, across the range of threads considered. Parallel MCMC chain lengths were optimized using the adaptive scheme presented in Figure 3.

become enticing. However, this approach makes for an inconsistent estimator of topology when divergence times are small (Pamilo and Nei 1988), and a biased estimator of species divergence times and substitution rates when ILS is present (Arbogast et al. 2002; Mendes and Hahn 2016; Ogilvie et al. 2016). MSC methods address these issues, but at the drawback of their demanding computational runtimes.

Therefore, as multithreading technologies become increasingly affordable, the appeal in parallelizing multispecies inference becomes clear. StarBeast3 exploits the assumption of conditional independence between gene trees, by doing Bayesian inference on gene trees

in parallel, and therefore it scales with the size of the problem. StarBeast3 can handle large data sets (with hundreds of genes) and achieve convergence several times faster than its predecessors.

A Balanced Traversal Through Parameter Space

All areas of parameter space should be explored approximately evenly during MCMC. If one area of parameter space is being explored more rapidly than another, then computational resources allotted to the former should be diverted to the latter. This is best exemplified by the phylogenetic substitution model which,

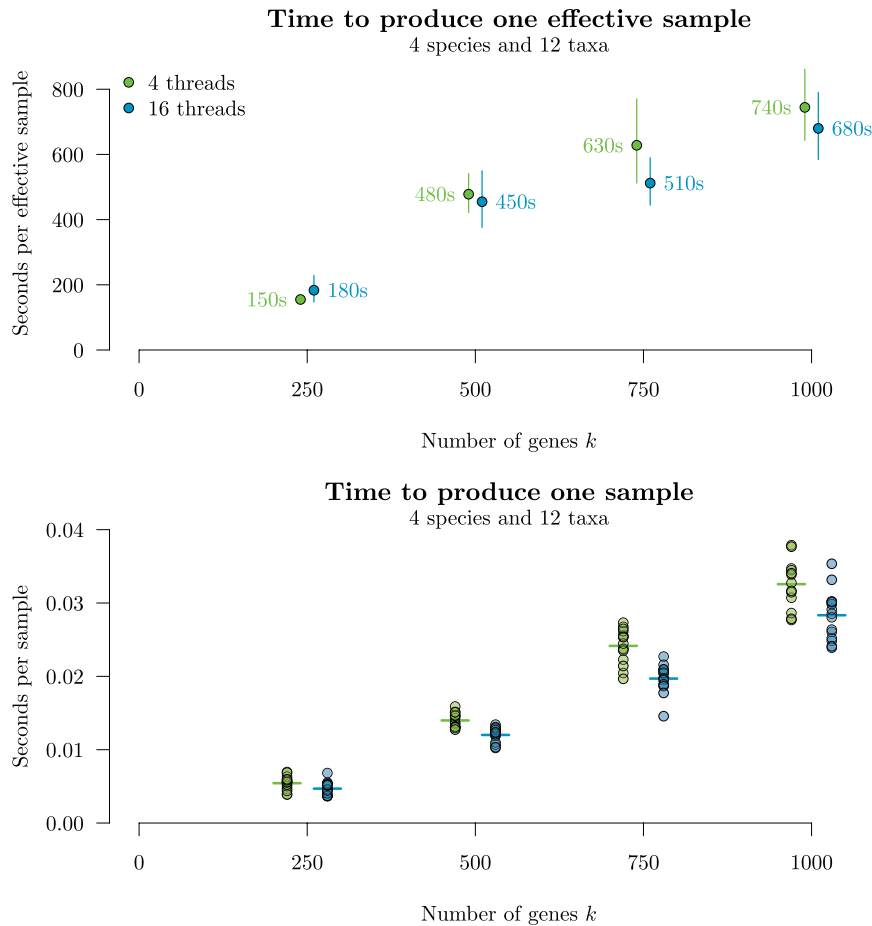


FIGURE 9. Performance of StarBeast3, across varying gene-tree sizes k and varying thread counts. Fifteen replicates of MCMC were run under each setting. Top: mean time taken to produce one effective sample (averaged across the ESSes of the following terms: $p(\theta|D)$, $p(D|\theta)$, $p(\theta)$, h_S , l_S , λ , N_e , μ_N , l_G , h_G , $D_{\text{upgma}}(\mathbf{G})$, $D_{ij}(\mathbf{G})$, κ , f , and v), ± 1.96 se. Means and standard errors were computed in log space. Bottom: time required to produce one state in the MCMC chain.

despite requiring relatively little attention to converge, still requires full recalculation of the tree likelihood upon every proposal (Felsenstein 1981). Conversely, tree topologies often converge rather poorly and can require significant attention to be rescued from local optima. By fine tuning our MCMC operator proposal probabilities, we have achieved a balanced traversal through all areas of the MSC parameter space. Although some parameters converge slower for StarBeast3 than they do for StarBeast2 (such as those in the substitution model), the slowest parameters converge significantly faster for the former; up to $37\times$ as fast (see the min term in Figs. 6 and 7).

For StarBeast3, we employed adaptable operators which are able to learn the proposal probabilities of other operators based on their ability to explore a single area of parameter space (Douglas et al. 2021b). However, there would be a great benefit in an adaptable operator scheme which learns and applies a balanced exploration across different areas of parameter space on a problem-by-problem basis.

CONCLUSION

Here we introduce StarBeast3—a software package for performing efficient Bayesian inference on genomic data under the MSC model. We verified StarBeast3’s correctness and we benchmarked its performance against StarBeast2, which is an order of magnitude faster than its still popular predecessor *BEAST. We showed that StarBeast3 is significantly faster than StarBeast2. Notably, relaxed clock parameters converged between 3 and $30\times$ faster, but most importantly even the “slowest” parameters converged up to $36\times$ faster. Our adaptive operator scheme allows proposal probabilities to be learned on a problem-by-problem basis, making StarBeast3 suitable for a range of data sets. By estimating effective population sizes (instead of analytically integrating the term out), we were able to parallelize gene-tree proposals and demonstrated that doubling the number of allotted threads was associated with an increase in performance by around 14%. StarBeast3 is highly effective at performing fast Bayesian inference on large data sets with over 100 genes.

SOFTWARE AVAILABILITY

StarBeast3 is available as an open-source BEAST 2 package with an easy-to-use graphical user interface. Instructions for downloading and running StarBeast3 can be found at <https://github.com/rbouckaert/starbeast3>.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.f1vhhmgzk>.

FUNDING

This study was supported by a Marsden grant 18-UOA-096 from the Royal Society of New Zealand. Software packages were benchmarked using the New Zealand eScience Infrastructure (NeSI) cluster, funded by the New Zealand Ministry of Business, Innovation, and Employment.

REFERENCES

- Arbogast B.S., Edwards S.V., Wakeley J., Beerli P., Slowinski J.B. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu. Rev. Ecol. Syst.* 33:707–740.
- Baele G., Lemey P., Rambaut A., Suchard M.A. 2017. Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in *beast*. *Bioinformatics* 33:1798–1805.
- Ballesteros J.A., Sharma P.P. 2019. A critical appraisal of the placement of xiphosura (chelicerata) with account of known sources of phylogenetic error. *Syst. Biol.* 68:896–917.
- Barrow L.N., Ralicki H.F., Emme S.A., Lemmon E.M. 2014. Species tree estimation of North American chorus frogs (hylidae: Pseudacris) with parallel tagged amplicon sequencing. *Mol. Phylogenet. Evol.* 75:78–90.
- Belfiore N.M., Liu L., Moritz C. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Syst. Biol.* 57:294–310.
- Benson A., Friel N. 2018. Adaptive MCMC for multiple changepoint analysis with applications to large datasets. *Electron. J. Stat.* 12:3365–3396.
- Bickford D., Lohman D.J., Sodhi N.S., Ng P.K., Meier R., Winker K., Ingram K.K., Das I. 2007. Cryptic species as a window on diversity and conservation. *Trends Ecol. & Evol.* 22:148–155.
- Blom M., Bragg J.G., Potter S., Moritz C. 2016a. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. *Syst. Biol.* 66:352–366.
- Blom M., Horner P., Moritz C. 2016b. Convergence across a continent: adaptive diversification in a recent radiation of Australian lizards. *Proc. R. Soc. B* 283:20160181.
- Bouckaert R. 2016. Phylogeography by diffusion on a sphere: whole world phylogeography. *PeerJ* 4:e2406.
- Bouckaert R. 2021. An efficient coalescent epoch model for Bayesian phylogenetic inference. *bioRxiv*.
- Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., du Plessis L., Poppinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu C.H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. *Beast 2.5*: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 15:e1006650.
- Bouckaert R., Alvarado-Mora M.V., Pinho J.R. 2013. Evolutionary rates and HBV: issues of rate estimation with Bayesian molecular methods. *Antivir. Ther.* 18:497–503.
- Bouckaert R.R. 2020. Obama: Obama for Bayesian amino-acid model averaging. *PeerJ* 8:e9460.
- Bragg J.G., Potter S., Bi K., Catullo R., Donnellan S.C., Eldridge M.D.B., Joseph L., Keogh J.S., Oliver P., Rowe K.C., Moritz C. 2017. Resources for phylogenomic analyses of Australian terrestrial vertebrates. *Mol. Ecol. Resour.* 17:869–876.
- Bryson Jr R.W., Linkem C.W., Pavón-Vázquez C.J., Nieto-Montes de Oca A., Klicka J., McCormack J.E. 2017. A phylogenomic perspective on the biogeography of skinks in the plestiodon brevirostris group inferred from target enrichment of ultraconserved elements. *J. Biogeogr.* 44:2033–2044.
- Cadena C.D., Kozak K.H., Gómez J.P., Parra J.L., McCain C.M., Bowie R.C., Carnaval A.C., Moritz C., Rahbek C., Roberts T.E., Sanders N.J., Schneider C.J., VanDerWal J., Zamudio K.R., Graham C. 2011. Latitude, elevational climatic zonation and speciation in New World vertebrates. *Proc. R. Soc. Lond. [Biol]* 279:194–201.
- Collienne L., Gavryushkin A. 2021. Computing nearest neighbour interchange distances between ranked phylogenetic trees. *J. Math. Biol.* 82:1–19.
- Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24:332–340.
- Douglas J. 2020. Uglytrees: a browser-based multispecies coalescent tree visualiser. *Bioinformatics* 37:268–269.
- Douglas J., Geoghegan J.L., Hadfield J., Bouckaert R., Storey M., Ren X., de Ligt J., French N., Welch D. 2021a. Real-time genomics for tracking severe acute respiratory syndrome coronavirus 2 border incursions after virus elimination, New Zealand. *Emerg. Infect. Dis.* 27:2361.
- Douglas J., Zhang R., Bouckaert R. 2021b. Adaptive dating and fast proposals: revisiting the phylogenetic relaxed clock model. *PLoS Comput. Biol.* 17:e1008322.
- Drummond A.J., Bouckaert R.R. 2015. *Bayesian evolutionary analysis with BEAST*. Cambridge: Cambridge University Press.
- Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.
- Drummond A.J., Nicholls G.K., Rodrigo A.G., Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.
- Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19.
- Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* 94:447–462.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.
- Fujita M.K., Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends Ecol. Evol.* 27:480–488.
- Gelman A., Gilks W.R., Roberts G.O. 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab.* 7:110–120.
- Geman S., Geman D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6:721–741.
- Grummer J.A., Bryson R.W. Jr, Reeder T.W. 2013. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group. *Syst. Biol.* 63:119–133.
- Haario H., Saksman E., Tamminen J. 2001. An adaptive metropolis algorithm. *Bernoulli* 7:223–242.
- Hamilton C.A., Lemmon A.R., Lemmon E.M., Bond J.E. 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evol. Biol.* 16:1–20.
- Hasegawa, M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160–174.
- Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27:570–580.

- Heled J., Drummond A.J. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* 61:138–149.
- Hohna S., Defoin-Platel M., Drummond A.J. 2008. Clock-constrained tree proposal operators in Bayesian phylogenetic inference. 2008 8th IEEE International Conference on Bioinformatics and BioEngineering IEEE. p. 1–7.
- Höhna S., Drummond A.J. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. *Syst. Biol.* 61:1–11.
- Jones G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *J. Math. Biol.* 74:447–467.
- Jones G., Aydin Z., Oxelman B. 2015. Dissect: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics* 31:991–998.
- Kang Y.J., Kim S.K., Kim M.Y., Lestari P., Kim K.H., Ha B.K., Jun T.H., Hwang W.J., Lee T., Lee J., Shim S., Yoon M.Y., Jang Y.E., Han K.S., Taepayoon P., Yoon N., Somta P., Tanya P., Kim K.S., Gwag J.G., Moon J.K., Lee Y.H., Park B.S., Bombarely A., Doyle J.J., Jackson S.A., Schafleitner R., Srinives P., Varshney R.K., Lee S.H. 2014. Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* 5:1–9.
- Kubatko L.S., Gibbs H.L., Bloomquist E.W. 2011. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *sistrurus* rattlesnakes. *Syst. Biol.* 60:393–409.
- Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014. Species delimitation using genome-wide SNP data. *Syst. Biol.* 63:534–542.
- Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5:1798–1805.
- Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. *Evolution* 62:2080–2091.
- Liu L., Wu S., Yu L. 2015. Coalescent methods for estimating species trees from phylogenomic data. *J. Syst. Evol.* 53:380–390.
- Liu L., Yu L., Edwards S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* 10:302.
- Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* 53:320–328.
- Maddison W.P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.
- Maliet O., Hartig F., Morlon H. 2019. A model with many small shifts for estimating species-specific diversification rates. *Nat. Ecol. Evol.* 3:1086–1092.
- Mendes F.K., Hahn M.W. 2016. Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* 65:711–721.
- Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.
- Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31:i44–i52.
- Mitchell K.J., Llamas B., Soubrier J., Rawlence N.J., Worthy T.H., Wood J., Lee M.S., Cooper A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science* 344:898–900.
- Nee S., May R.M., Harvey P.H. 1994. The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. B* 344:305–311.
- Ogilvie H., Bouckaert R., Drummond A. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* 34:2101–2114.
- Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. *Syst. Biol.* 65:381–396.
- Oliver J.C. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* 67:1823–1830.
- Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–583.
- Pepper M., Doughty P., Fujita M.K., Moritz C., Keogh J.S. 2013. Speciation on the rocks: integrated systematics of the *Heteronotia spelea* species complex (Gekkota; Reptilia). *PLoS One* 8:e78110.
- Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* 67:901.
- Rannala B., Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.* 66:823–842.
- Rowe K.C., Aplin K.P., Baverstock P.R., Moritz C. 2011. Recent and rapid speciation with limited morphological disparity in the genus *Rattus*. *Syst. Biol.* 60:188–203.
- Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.
- Sauquet H., Ho S.Y.W., Gandolfo M.A., Jordan G.J., Wilf P., Cantrill D.J., Bayly M.J., Bromham L., Brown G.K., Carpenter R.J., Lee D.M., Murphy D.J., Sniderman J.M.K., Udovicic F. 2011. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of *Nothofagus* (Fagales). *Syst. Biol.* 61:289–313.
- Sokal R.R. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* 38:1409–1438.
- Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. USA* 109:14942–14947.
- Stenson P.D., Mort M., Ball E.V., Evans K., Hayden M., Heywood S., Hussain M., Phillips A.D., Cooper D.N. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genetics* 136:665–677.
- Thawornwattana Y., Dalquen D., Yang Z. 2018. Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Anal.* 13:1037–1063.
- Vihola M. 2012. Robust adaptive metropolis algorithm with coerced acceptance rate. *Stat. Comput.* 22:997–1008.
- Whelan S., Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691–699.
- Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Mol. Phylogenet. Evol.* 92:63–71.
- Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA* 107:9264–9269.
- Yang Z., Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.* 31:3125–3135.
- Yang Z., Rodríguez C.E. 2013. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc. Natl. Acad. Sci. USA* 110:19307–19312.
- Yule G.U. 1925. II. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philos. Trans. R. Soc. Lond. B* 213:21–87.
- Zhang C., Huelsenbeck J.P., Ronquist F. 2020. Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. *Syst. Biol.*
- Zhang R., Drummond A. 2020. Improving the performance of Bayesian phylogenetic inference under relaxed clock models. *BMC Evol. Biol.* 20:1–28.