*Syst. Biol.* 71(2):334–352, 2022

DOI:10.1093/sysbio/syab047
Advance Access publication June 18, 2021

# Phase Resolution of Heterozygous Sites in Diploid Genomes is Important to Phylogenomic Analysis under the Multispecies Coalescent Model

JUN HUANG[1,2], JEREMY BENNETT[1,3], TOMÁŠ FLOURI[1], ADAM D. LEACHÉ[4], AND ZIHENG YANG[1,*]

[1]*Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK;* [2]*Department of Mathematics, Beijing Jiaotong University, No. 3 Shangyuancun, Haidian District, Beijing, 100044, P.R. China;* [3]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269-3043, USA; and* [4]*Department of Biology & Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195-1800, USA*

**Correspondence to be sent to: Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk*

*Jun Huang and Jeremy Bennett contributed equally to this article.*

*Received 26 October 2020; reviews returned 03 June 2021; accepted 21 June 2021*
*Associate Editor: Robert Thomson*

*Abstract*.—Genome sequencing projects routinely generate haploid consensus sequences from diploid genomes, which are effectively chimeric sequences with the phase at heterozygous sites resolved at random. The impact of phasing errors on phylogenomic analyses under the multispecies coalescent (MSC) model is largely unknown. Here, we conduct a computer simulation to evaluate the performance of four phase-resolution strategies (the true phase resolution, the diploid analytical integration algorithm which averages over all phase resolutions, computational phase resolution using the program PHASE, and random resolution) on estimation of the species tree and evolutionary parameters in analysis of multilocus genomic data under the MSC model. We found that species tree estimation is robust to phasing errors when species divergences were much older than average coalescent times but may be affected by phasing errors when the species tree is shallow. Estimation of parameters under the MSC model with and without introgression is affected by phasing errors. In particular, random phase resolution causes serious overestimation of population sizes for modern species and biased estimation of cross-species introgression probability. In general, the impact of phasing errors is greater when the mutation rate is higher, the data include more samples per species, and the species tree is shallower with recent divergences. Use of phased sequences inferred by the PHASE program produced small biases in parameter estimates. We analyze two real data sets, one of East Asian brown frogs and another of Rocky Mountains chipmunks, to demonstrate that heterozygote phase-resolution strategies have similar impacts on practical data analyses. We suggest that genome sequencing projects should produce unphased diploid genotype sequences if fully phased data are too challenging to generate, and avoid haploid consensus sequences, which have heterozygous sites phased at random. In case the analytical integration algorithm is computationally unfeasible, computational phasing prior to population genomic analyses is an acceptable alternative. [BPP; introgression; multispecies coalescent; phase; species tree.]

## INTRODUCTION

Next-generation sequencing technologies have revolutionized population genetics and phylogenetics by making it affordable to sequence whole genomes or large portions of the genome, even for nonmodel organisms. Many phylogenomic studies use the approach of reduced representation library to maximize their DNA sequencing efforts on a small subset of the genome. These strategies can generate thousands of genomic segments (called loci in this article irrespective of whether they are protein-coding) with high coverage, and target sequences can be assembled with confidence. Examples include restriction site-associated DNA sequencing (RADseq), which is used frequently to identify single nucleotide polymorphisms (SNPs) for population genetic and phylogeographic studies (Andrews et al. 2016; Leaché and Oaks 2017), although, it has also been applied to address phylogenetic questions at deeper timescales (Eaton et al. 2017). A more common approach for phylogenomic studies is targeted sequence capture, generating so-called reduced-representation data sets, with typically longer sequences for distantly related species than with RADseq data. Examples include exome sequencing, ultraconserved elements (UCEs, Faircloth et al. 2012), anchored hybrid enrichment (AHE, Lemmon et al. 2012), conserved nonexonic elements (CNEEs, Edwards et al. 2017), or rapidly evolving long exon capture (RELEC, Karin et al. 2020).

Typical sequencing technologies produce short fragments of sequenced DNA called "reads" that are either *de novo* assembled or mapped to a pre-existing reference genome. This leads to chromosomal positions being sequenced a variable number of times across the genome (usually referred to as the sequencing depth). A common practice in genome sequencing projects has been to produce the so-called "haploid consensus sequence" for a diploid individual, which uses the most common nucleotide at any heterozygous site to produce one genomic sequence. Assemblers like Velvet (Zerbino and Birney 2008), ABySS (Simpson et al. 2009), and Trinity (Grabherr et al. 2011), pick up only one of the two nucleotide bases at any heterozygous site and essentially resolve the phase of heterozygous sites at random, producing chimeric sequence that may not exist in nature. Suppose a diploid individual is heterozygous at two sites in a genomic region, so that the diploid genotype may be represented Y...R, with two heterozygous sites Y (for T/C) and R (for A/G) (Fig. 1). Suppose the reads are 14×T and 6×C at the first site, and 7×A, 10×G, and 1×T at the second (with the single

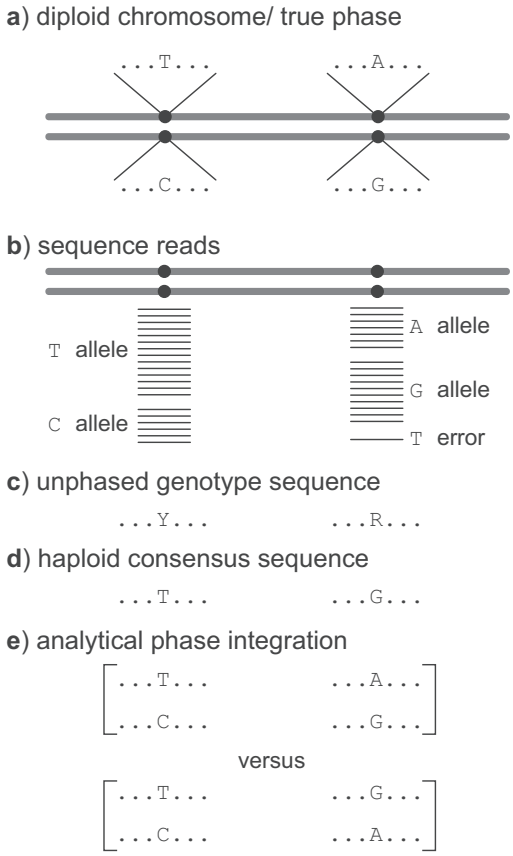


FIGURE 1. Example of heterozygote phase resolution. a) A hypothetical diploid chromosome with two heterozygous sites (T/C and A/G). The true haploid genotypes are T...A and C...G. b) Sequence reads around the two heterozygous sites, assuming that they are far apart on the chromosome so that they are not present on any single read (in which case phase would be determined) while they are close enough to be on one locus. In this case, genome assemblers should produce the unphased genotype sequence (c), using the IUPAC ambiguity codes to represent heterozygote sites, but instead they produce the so-called "haploid consensus sequence" (d), picking up the most common nucleotide at each heterozygote site (T...G since T and G are by chance the most common sequence reads at the two sites), which may not match either of the true haploid sequences. e) Analytical integration of phase resolution takes the unphased genotype sequences as data and averages over all possible phase resolutions, weighting each one appropriately according to their relative likelihood based on the whole sequence alignment at the locus.

T to be most likely a sequencing error). The haploid consensus sequence is constructed as T...G. In effect a heterozygote site with high quality scores for the two nucleotides is represented as one consensus nucleotide with a low quality score. Because it is largely pure chance which of the two nucleotides at a heterozygous site has the greater number of reads, this strategy is equivalent to resolving the phase at random and using only one of the constructed sequences. The resulting haploid consensus sequence may not be a real biological sequence and may not represent the biology of the diploid individual. Besides loss of information, a more serious problem is that the artifactual phased haploid sequence may be unusually divergent from other sequences in the sample, potentially introducing systematic biases in downstream inference. Currently, constructing true diploid *de novo* assemblies is expensive. A sequencing platform has been developed in combination with bioinformatic algorithms to determine the true diploid genome sequence but the strategy still involves high cost (Weisenfeld et al. 2017). If a read is long and fully covers a locus, multiple heterozygous sites in the same locus will be naturally phased. However, if the reads are short, and the two heterozygous sites do not occur in the same read, their genotypic phase resolution will become an issue.

How the heterozygote phase is resolved may have a significant impact on population genomic and phylogenomic inference using genomic sequence data. Phase information is well-known to be important for relating genotype to phenotype in human disease mapping (Tewhey et al. 2011). Similarly, Gronau et al. (2011) found that use of an analytical integration method (which averages over all possible phase resolutions) leads to nearly identical performance as the use of true phase resolutions for estimating population parameters, and that random phase resolution produced unreliable estimates. Andermann et al. (2019) developed a bioinformatics pipeline to recover allelic sequences from sequence capture data and found it to produce more accurate estimation of species divergence times under the MSC model (Rannala and Yang 2003) than other strategies such as use of consensus haploid sequences, random phasing, or ambiguity encoding. Overall little is known about the effects of heterozygote phase resolution on many inference problems using multilocus genomic sequence data under the MSC model, including species tree estimation, estimation of population sizes and species divergence times, and inference of cross-species introgression/hybridization.

We have implemented in BPP (Flouri et al. 2018) an analytical integration algorithm to handle unphased diploid sequences, developed by Gronau et al. (2011) in their G-PhoCS program, which is an orthogonal extension of an earlier version of BPP (Rannala and Yang 2003; Burgess and Yang 2008). Previously, Kuhner and Felsenstein (2000) implemented an Markov chain Monte Carlo (MCMC) algorithm to average over different phase resolutions in the likelihood calculation for estimating $\theta$ under the single-population coalescent. The algorithm was found to mix slowly even for small data sets. The analytical integration algorithm uses a data-augmentation strategy, in which the unknown fully resolved haploid sequences constitute the complete data or latent variables and enumerates and averages over all possible phase resolutions, weighting them according to their likelihoods based on the whole sequence alignment. For example, if a diploid sequence has two heterozygous sites, Y...R, the approach will average over both phased genotypic resolutions: (i) T...A and C...G versus (ii) T...G and C...A (Fig. 1). Note that there may be rich information about the phase resolution of any unphased sequence in an alignment of many sequences, either from the same species or from different but closely related species. Consider for example the phase

resolutions for a human diploid sequence Y...R (Fig. 1). If we observe in the chimpanzee fully resolved sequences T...A and C...G (e.g., in an individual homozygous at both sites, with genotypes T/T...A/A) and never observe sequences T...G and C...A, then very likely the human diploid sequence has the haploid genotypes T...A and C...G, because we assume no recombination within each locus. Our implementation of the algorithm works with all four analyses under the MSC model in BPP (Yang 2015; Flouri et al. 2018; Flouri et al. 2020b), including species tree estimation (Yang and Rannala 2014; Rannala and Yang 2017) and species delimitation through Bayesian model selection (Yang and Rannala 2010; Yang and Rannala 2014; Leaché et al. 2019). We also implemented the algorithm under the multispecies-coalescent-with-introgression (MSci) model (Flouri et al. 2020a).

Here, we use computer simulation to evaluate different phase-resolution strategies in terms of their precision and accuracy in Bayesian species tree estimation under the MSC and in parameter estimation under both the MSC and MSci models. In addition to using the true phase resolution, which is generated during the simulation and is known with certainty, we also include analytical phase integration (Gronau et al. 2011; Flouri et al. 2018), phase resolution using the program PHASE (Stephens et al. 2001; Stephens and Donnelly 2003), and random resolution. The strategy of random resolution is largely equivalent to the common method of using haploid consensus sequences. The PHASE program was developed for population data from the same species, but is here applied to unphased sequences from both within and between species. We note that a number of computational phasing algorithms are available (Andres et al. 2007; Browning and Browning 2011), such as Haplotyper (Niu et al. 2002) and fastPHASE (Scheet and Stephens 2006). These are mostly developed to improve the computational efficiency and to handle long sequences (Choi et al. 2018), and are expected to produce similar results to PHASE in analysis of short sequences.

## Materials and Methods

### *Simulation to Estimate Species Trees*

We use the program MCcoal in BPP3.4 (Yang 2015) or the `simulate` switch of BPP4.3 (Flouri et al. 2020b) to simulate gene trees and multilocus sequence data using four fixed species trees for eight species (Fig. 2a,a′,b,b′). The trees have very short branches, mimicking challenging species trees generated during radiative speciation events. In the two deep trees, species divergences are much older than average coalescent times ($\theta/2$). In the two shallow trees, species divergences are very recent relative to coalescent times, mimicking different populations of the same species. Note that in this study, we make no distinction between species and populations. The MSC model has two sets of parameters: the species divergence times ($\tau$s) and the population size parameters ($\theta$s). Both are measured by the expected number of mutations/substitutions per site. For each species/population, $\theta = 4N\mu$, where $N$ is the effective population size and $\mu$ is the mutation rate per site per generation. We consider two mutation rates, with $\theta$ = 0.001 (low rate) or 0.01 (high rate), respectively, for all populations on the tree. The species divergence times ($\tau$s) are given as multiples of $\theta$. We consider 10, 20, 50, or 100 loci, with each locus having 500 sites. On average there should be 0.5 and 5 heterozygous sites between the two sequences of any individual at the low and high rates, respectively. We sample $S = 2$ or 4 haploid sequences (or 1 or 2 diploid individuals) per species at each locus. Gene trees with branch lengths (coalescent times) are generated independently among loci using the MSC density given the species tree and parameters (Rannala and Yang 2003). The JC model (Jukes and Cantor 1969) is then used to "evolve" the sequences along the gene tree to generate the sequence alignments at the tips of the tree. Analysis of this full data set by BPP is strategy "F."

To simulate unphased diploid sequences, two sequences from the same species are combined into one diploid sequence, using the International Union of Pure and Applied Chemistry (IUPAC) ambiguity characters to represent heterozygous sites (for example, Y means a T/C heterozygote) (Fig. 1c). The data of unphased diploid sequences are analyzed using the diploid or phase option of the BPP program (strategy "D"), which analytically averages over all possible phase resolutions (Gronau et al. 2011). With strategy "P," we use the program PHASE (Stephens et al. 2001) to resolve the phase and then analyze the phased sequences using BPP (with 16 or 32 sequences in the alignment per locus for $S = 2$ and 4, respectively). Lastly, we use random phase resolution, referred to as strategy "R." The simulation program automatically generates the sequence alignments for strategies F, D, and R. For strategy P, we ran PHASE 2.1 (Stephens et al. 2001) to reconstruct the phased sequences for each locus, and used the PERL program SeqPhase (Flot 2010) to convert files.

The number of replicate data sets is 100. With four trees, two mutation rates ($\theta = 0.001$ or 0.01), two sampling configurations ($S = 2$ or 4), four numbers of loci ($L = 10, 20, 50, 100$), we generated in total $4 \times 2 \times 2 \times 4 \times 100 = 6400$ data sets, each of which is analyzed using the four strategies. The BPP program (Flouri et al. 2018) was used in the analysis. Inverse-gamma priors are assigned on parameters under the MSC model, with the shape parameter 3 so that the priors are diffuse and with the mean to be close to the true value. We use $\theta \sim$ IG(3, 0.02) with mean 0.01 and $\tau_0 \sim$ IG(3, 0.08) with mean 0.04 for the age of the root of the species tree for data simulated with the high rate ($\theta$ = 0.01). For data of the low rate ($\theta$ = 0.001), the priors are $\theta \sim$ IG(3, 0.002) with mean 0.001 and $\tau_0 \sim$ IG(3, 0.008) with mean 0.004. The prior means for $\tau_0$ are close to the true values for the deep trees but are larger than the true values for the shallow trees, although the priors are diffuse. For species tree estimation, we integrate out

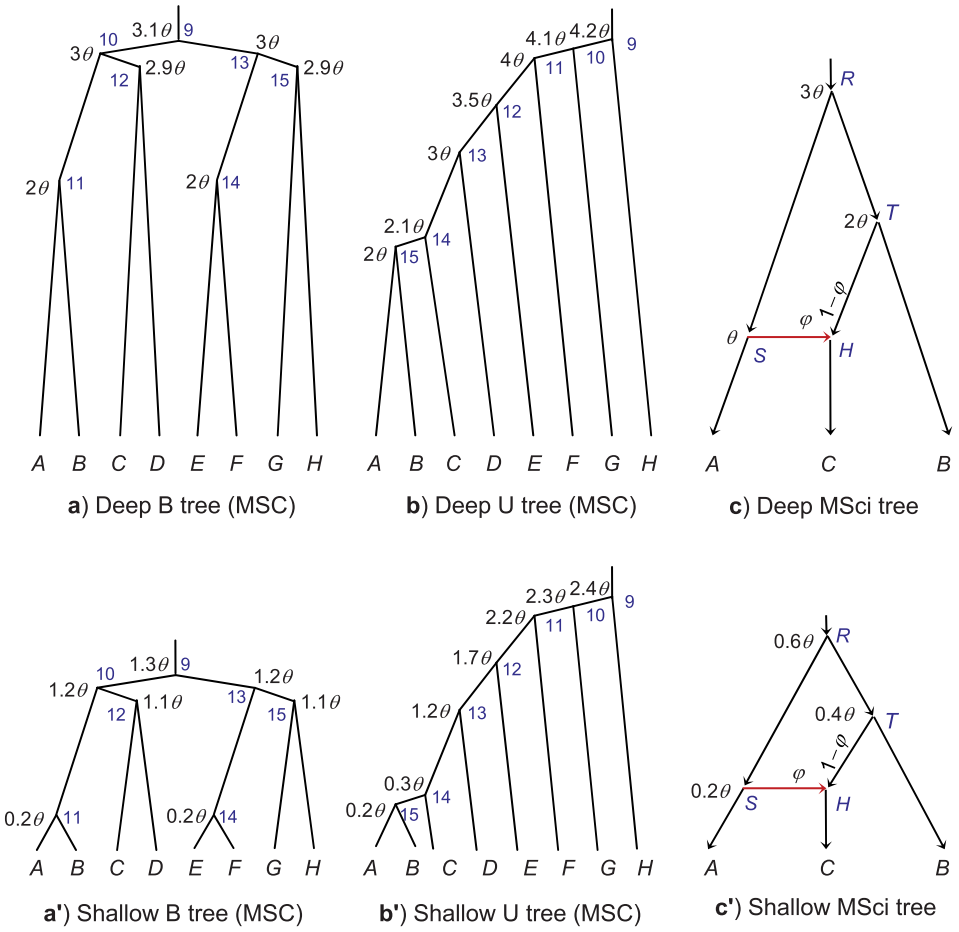


FIGURE 2. (a and a′) Deep and shallow balanced species trees and b and b′) deep and shallow unbalanced species trees for eight species used for simulating data under the MSC model. c and c′) Deep and shallow species trees with introgression used to simulate data under the MSci model. The ages of internal nodes (τs) are shown next to the nodes, with θ = 0.01 (high rate) or 0.001 (low rate). The blue indexes at internal nodes of the tree are used to identify the parameters associated with the ancestral species (e.g., $\tau_9$ is the age of the root and $\theta_9$ is the population size for the root population in a and b).

θs analytically through the use of the conjugate inverse-gamma priors. We conducted pilot runs to determine the chain lengths needed for convergence. The final settings for the MCMC are 20,000 iterations for burn-in, then taking $2\times10^5$ samples, sampling every two iterations.

Strategy P requires running the Bayesian MCMC program PHASE *L* times if there are *L* loci in the data set, to generate the fully resolved sequence alignments at the loci. This is somewhat expensive if there is a large number of loci and the mutation rate is high resulting in many heterozygous sites at each locus. After the data sets are generated, the BPP analysis of each data set by strategies F, P, and R involves about the same amount of computation. Strategy D is more expensive as the method averages over all possible phase resolutions, which may involve likelihood calculation for many site patterns, especially if there are many sequences per locus with many heterozygous sites.

For species tree estimation (A01 analysis in Yang 2015), we calculated the proportion (among the 100 replicates) with which each node on the true species tree is found in the *maximum a posteriori* (MAP) species tree in the BPP analysis. This is a measure of accuracy since the MAP tree is the best "point estimate" of the species tree (Rannala and Yang 1996). We examined the size and coverage probability of the 95% credibility set of species trees. The coverage probability is the proportion among the 100 replicate data sets in which the credibility set includes the true species tree. The size of the set indicates the precision or power of the method, but the method is considered reliable only if the coverage probability exceeds the nominal 95%.

### *Simulation to Estimate Parameters under the MSC Model*

The same data simulated under the MSC model for species tree estimation are analyzed using the four phase-resolution strategies to estimate parameters in the MSC model (θs and τs), with the species tree fixed. This is the A00 analysis in Yang (2015). We calculated the posterior means and the 95% HPD CI intervals for each parameter and examine the relative root mean

square error (rRMSE), using the posterior means as point estimates. This is defined as

$$\text{rRMSE} = \frac{1}{\omega}\left[\frac{1}{R}\sum_{i=1}^{R}(\hat{\omega}_i - \omega)^2\right]^{\frac{1}{2}}, \quad (1)$$

where $\omega$ is the true value of any parameter, and $\hat{\omega}_i$ its estimate (posterior mean) in the $i$th replicate data set, with $i = 1, \ldots, R$ over the $R = 100$ replicates. For example, rRMSE = 0.1 means that the mean square error is 10% of the true value. The rRMSE is a combined measure of bias and variance.

### *Simulation to Estimate Parameters under the MSci Model*

The MSci models for three species of Figure 2c and c′ are assumed to generate gene trees and sequence alignments using the `simulate` option of BPP4.3 (Flouri et al. 2020a). The three species have the phylogeny $(A,(C,B))$, but there was introgression from $A$ to $C$ at the time $\tau_H = \tau_S$, with the introgression probability $\varphi = 0.1$ and 0.3. Other settings are the same as above for the simulation under the MSC model. We consider two mutation rates (with $\theta$ = 0.001 and 0.01) and four datasizes (with $L = 10, 20, 50$, and 100 loci), with each locus having 500 sites. We sample either $S = 2$ or 4 sequences per species per locus. The JC model is used both to simulate and to analyze the data.

For data simulated at the high rate ($\theta = 0.01$), the priors are $\theta \sim$ IG(3, 0.02) and $\tau_0 \sim$ IG(3, 0.06) for the root age. At the low rate ($\theta = 0.001$), the priors are $\theta \sim$ IG(3, 0.002) and $\tau_0 \sim$ IG(3, 0.006). A $\mathbb{U}(0,1)$ prior is used for the introgression probability $\varphi$.

### *Analyses of Two Real Data Sets*

We applied different phase-resolution strategies (D, P, and R) to analyze two previously published data sets, one of East Asian brown frogs (Zhou et al. 2012) and another of Rocky Mountains chipmunks (Sarver et al. 2021), to demonstrate that the effects discovered in the simulations apply to real data analysis. With real data, the option of true phase resolution (F) is unavailable, and the analytical phase resolution (D) is expected to perform the best. In addition, we include an approach of treating heterozygote sites in the alignment as ambiguity characters in the likelihood calculation, and refer to it as strategy "A" (for ambiguity). As heterozygotes (with, e.g., Y meaning both T and C) are not ambiguities (with Y meaning either T or C), this is a mistaken approach of handling the data, and has the obvious effect of underestimating the heterozygosity or $\theta$ parameters. It was thus not included in our simulation, but we use it in the real data analysis to illustrate its effects.

We reanalyzed a data set of five nuclear loci from the East Asia brown frogs in the *Rana chensinensis* species complex (Zhou et al. 2012) to infer the species tree (the A01 analysis) and to estimate the parameters under the MSC on the MAP tree (the A00 analysis). There are three morphologically recognized species or four populations: *R. chensinensis* (clades C and L), *R. kukunoris* (K) and *R. huanrensis* (H) (Fig. 3a). The data set was previously analyzed by Yang (2015), treating heterozygotes as ambiguities (strategy A). Each locus has 20–30 sequences, with sequence lengths to be 285–498 sites. We assign inverse-gamma priors on parameters: $\theta \sim$ IG(3, 0.002) with mean 0.001 and $\tau_0 \sim$ IG(3, 0.004) with mean 0.002 for the root age. We used a burnin of 8000 iterations, then taking $10^5$ samples, sampling every two iterations. The same analysis was run at least twice to confirm consistency between runs. This is a small data set and the MCMC algorithm mixes well.

The second data set consist of nuclear loci from six species of Rocky Mountains chipmunks in the *Tamias quadrivittatus* group: *Tamias canipes* (C), *Tamias cinereicollis* (I), *Tamias dorsalis* (D), *Tamias quadrivittatus* (Q), *Tamias rufus* (R), and *Tamias umbrinus* (U) (Fig. 3b). Sarver et al. (2021) used a targeted sequence-capture approach to sequence 51 Rocky Mountains chipmunks from those six species. As a reference genome assembly was lacking, reads were assembled iteratively into contigs using an approach called "assembly by reduced complexity." A data set of 1060 nuclear loci was compiled for molecular phylogenomic and introgression analyses, including three individuals from an outgroup species, *Tamias striatus*. Each locus consists of 54 sequences, with the sequence length ranging from 14 to 1026 sites. High-quality heterozygotes, judged by mapping quality and read depth, are represented in the alignments using the IUPAC ambiguity codes. The filters applied by the authors suggest that the loci may be mostly coding exons or conserved parts of the genome. The majority of loci have $\leq 5$ variable sites (including the outgroup). We used the first 500 loci in our analyses to infer the species tree and to estimate parameters under the MSC model. We assigned inverse-gamma priors on parameters: $\theta \sim$ IG(3, 0.002) with mean 0.001 and $\tau_0 \sim$ IG(3, 0.01) with mean 0.005 for the root age. In the A01 analysis (species tree estimation), we used a burnin of 16,000 iterations, then taking $2 \times 10^5$ samples, sampling every two iterations. The A00 analysis (parameter estimation on the MAP tree) used the same settings except that only $10^5$ samples were collected. The same analysis was run at least twice to confirm consistency between runs.

## RESULTS

### *Species Tree Estimation under the MSC Model*

Bayesian analysis of each replicate data set using each of the four strategies produced a sample from the posterior distribution of the species trees, which we summarized to identify the maximum *a posteriori* probability (MAP) tree, and construct the 95% credibility set of species trees. The proportion, among the 100 replicates, with which the clades represented by those short branches were recovered

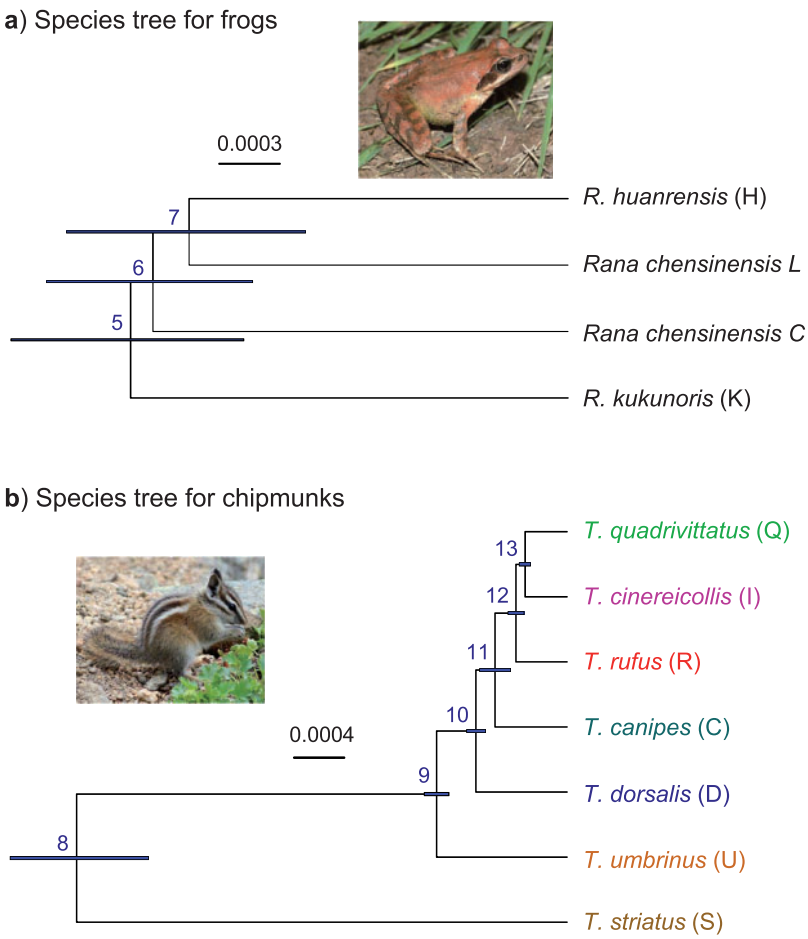


FIGURE 3. Inferred species trees (a) for East Asian brown frogs and (b) for Rocky Mountains chipmunks. Branch lengths reflect the posterior means of divergence times, with branch bars representing the 95% HPD intervals, obtained under the MSC using the analytical phase integration algorithm (strategy D). Estimates of other parameters are in Table 6.

in the MAP tree are shown in Table 1 and Supplementary Tables S1–S3 available on Dryad at https://doi.org/10.5061/dryad.vmcvdncrd. Other clades on the trees, represented by longer branches, were recovered with probability near 100%, even for the low mutation rate and 10 loci. We also plotted the posterior probabilities for the true tree for the different phasing strategies in Figure 4, Supplementary Figures S1–S3 available on Dryad. Strategy F, the analysis of the fully resolved haploid data, is expected to have the best performance and is thus the gold standard, against which the other strategies are compared.

In data simulated using the two deep trees (Deep B and Deep U) (Fig. 2a,b), the four phase-resolution strategies produced similar probabilities for recovering the true clades, with the differences among methods not being larger than the random sampling errors due to the limited number of replicates (Supplementary Tables S1 and S3 available on Dryad). The different strategies most often produced the same MAP tree, although the posterior probability attached to the MAP tree varies somewhat among methods, but the differences are comparable to MCMC sampling errors. This can be seen in Supplementary Figures S1 and S3 available on Dryad, where the posterior for the true tree is plotted. Even random resolution (R) produced very similar results to the use of the fully resolved data (F). Note that in data simulated at the high rate, there are very likely to be two or more heterozygote sites in the diploid genotype of each individual at any locus, and the switch error rate for random phase resolution, which is the average proportion of heterozygous sites misassigned relative to the previous heterozygous site (Stephens and Donnelly 2003; Andres et al. 2007), is 50%. Even the PHASE program generates substantial errors of phase resolution at the high mutation rate (Table 2). Species tree estimation is thus robust to considerable phasing errors when species divergences are much older than average coalescent times.

For the two shallow trees (Fig. 2a′,b′), large differences were found among the four strategies (Table 1 and Supplementary Table S2 available on Dryad, Fig. 4 and Supplementary Fig. S2 available on Dryad). While strategy D produced results very similar to use of the full data (F), both strategies P and R had poorer performance, especially at the high rate, when strategy R produced

TABLE 1. (MSC A01, shallow, $S=4$) Probabilities of recovering true clades and the size and coverage of the 95% credibility set of species trees when the true species tree is Shallow B and Shallow U (Fig. 2a′,b′) and $S=4$ sequences are sampled per species

| Key | Species tree B | | | | | | | Species tree U | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $C_{10}$ | $C_{12}$ | $C_{13}$ | $C_{15}$ | tree | CI cover | CI size | $C_{10}$ | $C_{11}$ | $C_{15}$ | tree | CI cover | CI size |
| Low mutation rate | | | | | | | | | | | | | |
| F, 10L | 0.34 | 0.25 | 0.27 | 0.22 | 0.00 | 0.66 | 233.8 | 0.49 | 0.47 | 0.63 | 0.12 | 0.96 | 84.4 |
| D, 10L | 0.35 | 0.25 | 0.25 | 0.22 | 0.00 | 0.66 | 233.4 | 0.47 | 0.44 | 0.61 | 0.09 | 0.96 | 82.7 |
| P, 10L | 0.34 | 0.25 | 0.25 | 0.20 | 0.00 | 0.68 | 235.5 | 0.47 | 0.47 | 0.61 | 0.12 | 0.96 | 82.9 |
| R, 10L | 0.36 | 0.24 | 0.26 | 0.24 | 0.01 | 0.66 | 225.7 | 0.47 | 0.46 | 0.61 | 0.13 | 0.94 | 81.8 |
| F, 20L | 0.46 | 0.29 | 0.34 | 0.26 | 0.00 | 0.73 | 178.3 | 0.52 | 0.53 | 0.74 | 0.22 | 0.97 | 33.5 |
| D, 20L | 0.45 | 0.29 | 0.34 | 0.22 | 0.01 | 0.73 | 175.4 | 0.54 | 0.52 | 0.72 | 0.23 | 0.97 | 33.8 |
| P, 20L | 0.46 | 0.27 | 0.38 | 0.26 | 0.01 | 0.73 | 178.2 | 0.55 | 0.53 | 0.76 | 0.26 | 0.97 | 33.1 |
| R, 20L | 0.47 | 0.26 | 0.31 | 0.26 | 0.02 | 0.70 | 168.8 | 0.53 | 0.50 | 0.74 | 0.22 | 0.97 | 32.6 |
| F, 50L | 0.56 | 0.44 | 0.56 | 0.46 | 0.07 | 0.90 | 86.3 | 0.60 | 0.65 | 0.95 | 0.40 | 0.97 | 11.4 |
| D, 50L | 0.56 | 0.45 | 0.53 | 0.47 | 0.06 | 0.90 | 83.3 | 0.59 | 0.61 | 0.95 | 0.40 | 0.95 | 11.5 |
| P, 50L | 0.61 | 0.45 | 0.52 | 0.48 | 0.08 | 0.87 | 89.0 | 0.59 | 0.61 | 0.93 | 0.40 | 0.98 | 11.7 |
| R, 50L | 0.52 | 0.37 | 0.57 | 0.46 | 0.06 | 0.86 | 80.8 | 0.65 | 0.63 | 0.91 | 0.41 | 0.96 | 11.6 |
| F, 100L | 0.72 | 0.75 | 0.81 | 0.74 | 0.33 | 0.99 | 25.5 | 0.75 | 0.77 | 0.99 | 0.60 | 0.99 | 7.0 |
| D, 100L | 0.75 | 0.74 | 0.81 | 0.76 | 0.34 | 0.98 | 25.7 | 0.75 | 0.76 | 1.00 | 0.59 | 0.99 | 6.8 |
| P, 100L | 0.74 | 0.71 | 0.80 | 0.75 | 0.34 | 0.97 | 26.4 | 0.74 | 0.78 | 1.00 | 0.59 | 0.98 | 7.1 |
| R, 100L | 0.73 | 0.66 | 0.75 | 0.72 | 0.26 | 0.96 | 27.4 | 0.75 | 0.74 | 0.98 | 0.57 | 0.99 | 7.9 |
| High mutation rate | | | | | | | | | | | | | |
| F, 10L | 0.68 | 0.58 | 0.68 | 0.49 | 0.19 | 0.92 | 91.5 | 0.70 | 0.76 | 0.96 | 0.53 | 1.00 | 11.2 |
| D, 10L | 0.70 | 0.58 | 0.66 | 0.50 | 0.18 | 0.92 | 95.3 | 0.72 | 0.75 | 0.94 | 0.52 | 1.00 | 11.8 |
| P, 10L | 0.66 | 0.58 | 0.63 | 0.48 | 0.14 | 0.91 | 94.6 | 0.68 | 0.76 | 0.92 | 0.53 | 0.98 | 12.5 |
| R, 10L | 0.53 | 0.54 | 0.64 | 0.45 | 0.12 | 0.89 | 108.4 | 0.71 | 0.77 | 0.78 | 0.44 | 0.97 | 13.3 |
| F, 20L | 0.91 | 0.74 | 0.90 | 0.72 | 0.43 | 0.99 | 22.2 | 0.80 | 0.85 | 1.00 | 0.72 | 1.00 | 5.7 |
| D, 20L | 0.92 | 0.75 | 0.88 | 0.72 | 0.43 | 1.00 | 23.3 | 0.81 | 0.85 | 1.00 | 0.72 | 1.00 | 6.0 |
| P, 20L | 0.91 | 0.70 | 0.89 | 0.72 | 0.41 | 1.00 | 27.2 | 0.77 | 0.86 | 0.97 | 0.68 | 1.00 | 6.6 |
| R, 20L | 0.86 | 0.71 | 0.79 | 0.66 | 0.31 | 0.98 | 30.1 | 0.79 | 0.84 | 0.93 | 0.64 | 0.99 | 7.3 |
| F, 50L | 1.00 | 0.97 | 1.00 | 0.94 | 0.91 | 1.00 | 4.1 | 0.90 | 0.97 | 1.00 | 0.87 | 1.00 | 2.6 |
| D, 50L | 1.00 | 0.97 | 1.00 | 0.94 | 0.91 | 1.00 | 4.1 | 0.90 | 0.97 | 1.00 | 0.87 | 1.00 | 2.6 |
| P, 50L | 1.00 | 0.94 | 1.00 | 0.92 | 0.86 | 1.00 | 4.3 | 0.91 | 0.97 | 1.00 | 0.88 | 1.00 | 2.7 |
| R, 50L | 0.98 | 0.91 | 0.94 | 0.90 | 0.76 | 1.00 | 5.6 | 0.92 | 0.97 | 1.00 | 0.89 | 0.99 | 2.9 |
| F, 100L | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 | 1.6 | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 1.7 |
| D, 100L | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 | 1.6 | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 1.6 |
| P, 100L | 1.00 | 0.99 | 1.00 | 0.99 | 0.98 | 1.00 | 1.6 | 0.99 | 0.98 | 1.00 | 0.97 | 1.00 | 1.7 |
| R, 100L | 1.00 | 0.98 | 1.00 | 0.99 | 0.97 | 1.00 | 2.0 | 1.00 | 0.98 | 1.00 | 0.98 | 1.00 | 1.7 |

*Note*: The two mutation rates are low ($\theta = 0.001$) and high ($\theta = 0.01$), while 10L, 20L, 50L, 100L are the number of loci. $C_{10}$, $C_{12}$, etc. are probabilities of recovering the true clades on the species trees, while "tree" is the probability of recovering the whole tree. "CI size" is the number of species trees in the 95% credibility set and and "CI cover" is the probability that the set contains the true species tree. Results for other simulation settings are in Supplementary Tables S1–S3 available on Dryad.

larger CI sets, with lower coverage than strategies F and D.

Thus phasing errors have different effects on species tree estimation depending on whether the species tree is deep or shallow. We suggest that this may be explained by the probability that the sequences from the same species coalesce before they reach the time of species divergence, when one traces the genealogical history at each locus backwards in time. For example, the probability that $S=2$ sequences from species $A$ coalesce before reaching the common ancestor of $A$ and $B$ is $\mathbb{P}\{t_{\text{mrca}} < \tau_{AB}\} = 1-e^{-4} \approx 0.982$ in the two deep trees and $1-e^{-0.4} \approx 0.330$ in the two shallow trees, while the corresponding probabilities for $S=4$ sequences are 0.967 and 0.077 for the deep and shallow trees, respectively (Supplementary Fig. S4 available on Dryad). In the deep trees, there is a high chance for all sequences from the same species to coalesce before reaching species divergence, and then the problem will be similar to using the ancestral sequence for each species (which is mostly determined by the most common nucleotides at the individual sites; Yang et al. 1995) for species tree estimation, a process that is not expected to be sensitive to phasing errors. In the shallow species trees, there are high chances that sequences from the same species may not have coalesced before reaching the time of species divergence, and sequences with phasing errors will enter ancestral populations, interfering with species tree estimation.

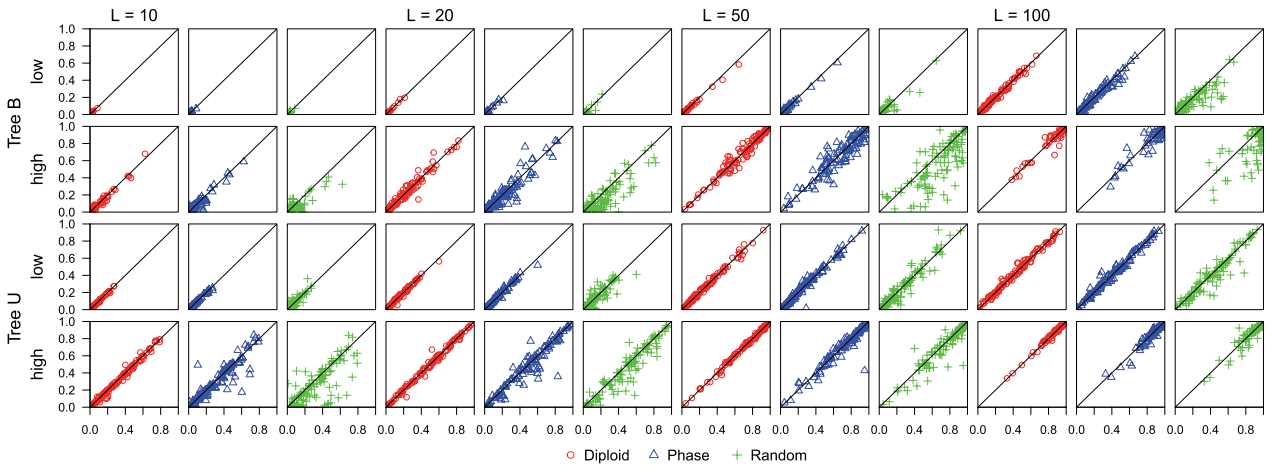


FIGURE 4. (A01 under MSC, shallow tree, $S=4$) Posterior probability for the true species tree for phase-resolution strategies D (diploid), P (PHASE), and R (random) plotted against the probability for strategy F (full data). The data are simulated under the MSC models with species trees Shallow B and Shallow U (Fig. 2a′,b′), with $S=4$ sequences sampled per species. Each plot has 100 scatter points, for the 100 replicate data sets, with the $x$-axis to be the posterior probability for strategy F while the $y$-axis is for strategies D, P, or R. "Low" ($\theta=0.001$) and "high" ($\theta=0.01$) refer to the mutation rate, and $L$ (= 10, 20, 50, 100) is the number of loci. Results for other simulation settings are in Supplementary Figures S1–S3 available on Dryad.

TABLE 2. Average switch error rate for data sets simulated under the MSC and MSci models in this study

| | PHASE (P) | | | | Random (R) | | | |
|---|---|---|---|---|---|---|---|---|
| | low | | high | | low | | high | |
| Model | $S=2$ | $S=4$ | $S=2$ | $S=4$ | $S=2$ | $S=4$ | $S=2$ | $S=4$ |
| MSC, Deep B | 0.485 | 0.327 | 0.499 | 0.371 | 0.505 | 0.504 | 0.499 | 0.501 |
| MSC, Deep U | 0.489 | 0.332 | 0.501 | 0.370 | 0.488 | 0.488 | 0.498 | 0.499 |
| MSC, Shallow B | 0.448 | 0.370 | 0.459 | 0.349 | 0.488 | 0.495 | 0.498 | 0.498 |
| MSC, Shallow U | 0.390 | 0.304 | 0.430 | 0.331 | 0.489 | 0.505 | 0.501 | 0.502 |
| MSci, Deep ($\varphi=0.1$) | 0.480 | 0.317 | 0.492 | 0.363 | 0.500 | 0.492 | 0.500 | 0.502 |
| MSci, Deep ($\varphi=0.3$) | 0.482 | 0.311 | 0.494 | 0.360 | 0.520 | 0.490 | 0.501 | 0.499 |
| MSci, Shallow ($\varphi=0.1$) | 0.402 | 0.342 | 0.461 | 0.346 | 0.496 | 0.489 | 0.492 | 0.498 |
| MSci, Shallow ($\varphi=0.3$) | 0.402 | 0.331 | 0.454 | 0.337 | 0.498 | 0.502 | 0.502 | 0.501 |

*Note*: Data of $L=100$ loci are used in the calculation although the error rate does not depend on the number of loci. The same data generated under the MSC model are used in the A01 (species tree estimation) and A00 (parameter estimation) analyses. Note that the error rate for random phase resolution (R) is expected to be 0.5.

While our main objective in this study is to evaluate the impacts of different phasing strategies, it is worth noting the effects of other major factors on species tree estimation that are obvious from our results (Fig. 4, Supplementary Figs. S1–S3 available on Dryad and Table 1, Supplementary Tables S1–S3 available on Dryad). By design species tree B is harder to recover than tree U because tree B has four short branches (for clades $C_{10}$, $C_{12}$, $C_{13}$, and $C_{15}$) while tree U has only three (for clades $C_{10}$, $C_{11}$, and $C_{15}$) (Fig. 2). Thus tree B is recovered with much lower probability than tree U by all methods in all parameter settings. We note that the individual clades in tree B are recovered with lower probabilities than those in tree U (Table 1 and Supplementary Tables S1–S3 available on Dryad). We speculate that this may be due to the fact that the four short branches in tree B are close together (so that 945 trees around them are nearly equally good) while the three short branches in tree U are far apart (so that only $3\times15=45$ trees around them are nearly equally good). Because of the symmetry in tree B, the probabilities of recovering clades $C_{10}$ and $C_{13}$ should be equal, as are those for $C_{12}$ and $C_{15}$. Differences within each pair reflect the random sampling errors due to our use of only 100 replicates. (Note that clades $C_{11}$ and $C_{14}$ were always recovered in the simulation.)

The mutation rate had a dramatic impact on the precision and accuracy of species tree estimation. At the higher rate (with $\theta$ = 0.01 vs. 0.001), the credibility set was smaller, its coverage was higher, and the MAP tree matched the true species tree with higher probability. In our species trees, species divergence times ($\tau$) are proportional to $\theta$. This allows us to compare the two values of $\theta$, mimicking the use of conserved or variable regions of the genome for species tree estimation. Our study focuses on closely related species with highly

TABLE 3. Mean and standard deviation ($\times 10^{-3}$) of estimates of $\theta$ for a single population (true value is 0.01) from a sample of $S$ sequences using BPP with different strategies of phase resolution and two summary methods

| Method | $S=2$ | $S=4$ | $S=8$ | $S=16$ | $S=32$ |
|---|---|---|---|---|---|
| BPP (F) | $10.06 \pm 1.02$ | $10.06 \pm 0.61$ | $10.03 \pm 0.52$ | $9.96 \pm 0.36$ | $10.03 \pm 0.34$ |
| BPP (D) | $10.06 \pm 1.02$ | $10.05 \pm 0.62$ | $10.17 \pm 0.53$ | $10.50 \pm 0.43$ | $11.19 \pm 0.47$ |
| BPP (P) | $10.06 \pm 1.02$ | $9.80 \pm 0.61$ | $9.84 \pm 0.51$ | $9.94 \pm 0.37$ | $10.05 \pm 0.34$ |
| BPP (R) | $10.06 \pm 1.02$ | $12.86 \pm 0.91$ | $18.13 \pm 1.32$ | $26.43 \pm 1.65$ | $41.27 \pm 3.22$ |
| Watterson ($\hat{\theta}_S$) | $9.94 \pm 1.01$ | $9.92 \pm 0.61$ | $9.85 \pm 0.55$ | $9.76 \pm 0.40$ | $9.82 \pm 0.36$ |
| Pairwise distance ($\hat{\theta}_\pi$) | $9.94 \pm 1.01$ | $9.94 \pm 0.63$ | $9.87 \pm 0.63$ | $9.78 \pm 0.50$ | $9.93 \pm 0.55$ |
| Pairwise distance($\hat{\theta}'_\pi$) | $10.01 \pm 1.03$ | $10.01 \pm 0.64$ | $9.94 \pm 0.64$ | $9.84 \pm 0.50$ | $9.99 \pm 0.56$ |

*Note*: Watterson's estimate ($\hat{\theta}_S$) and the average pairwise distance ($\hat{\theta}_\pi$) do not depend on phase resolutions. JC correction is applied in calculation of $\hat{\theta}'_\pi$.

similar sequences, and data simulated at the high rate contain more variable sites and more phylogenetic information.

The number of loci similarly had a huge impact on species tree estimation. With more loci, inference became more precise (with smaller credibility set) and more accurate (with the MAP tree matching the true tree with greater probability). Increasing the number of loci by 10 fold improves performance for all strategies more than increasing the mutation rate by the same factor.

The number of sequences sampled per species had consistent but relatively small effects on species tree estimation. Changing $S=2$ to 4 improved the probabilities of recovering the true clades in the true species tree, reduced the CI set size, and improved the coverage of the CI set, but the improvements are in general small.

It is noteworthy that the coverage of the 95% CI set was below the nominal 95% in small or uninformative data sets while above 95% in large and informative data sets. In the case of 10 loci at the low rate for tree Deep B, coverage was even below 50% (Supplementary Table S1 available on Dryad). Even though the set included nearly 500 trees, more than a half of the CI sets failed to include the true tree. In contrast, at the high mutation rate and with 50 or 100 loci, CI coverage was often 100%. The method is overconfident in small and uninformative data sets and conservative in large and informative ones. The same pattern was noted in a previous simulation examining the information content in phylogenomic data sets (Huang et al. 2020, Table 3). Note that in our simulation, the replicate data sets are generated under a fixed model (species tree) and fixed parameter values, so that we are evaluating the Frequentist properties of Bayesian model selection, and a match is not expected (Huelsenbeck and Rannala 2004; Yang and Rannala 2005). Yet the large discrepancies are striking.

### *Estimation of Divergence Times and Population Sizes under the MSC Model*

*The Impact of the Phasing Strategies.*—The same data sets simulated for species tree estimation were analyzed to estimate the parameters in the MSC model ($\theta$s and $\tau$s) with the species tree fixed (Fig. 2a,a′,b,b′). The posterior means and 95% HPD CI for the 100 replicates are plotted in Figure 5, and Supplementary Figures S5–S11 available on Dryad, while the relative root mean square errors (rRMSE) are presented in Supplementary Tables S4–S11 available on Dryad. Whereas the rRMSE reflects both biases and variances in parameter estimation, the data sets generated by the four phase-resolution strategies have about the same size in terms of the number of loci, the number of sequences per locus, and the number of sites per sequence, so that the sampling errors or variances are similar among methods and the differences in rRMSE mainly reflect differences in bias. Furthermore, we may use the symmetry of species tree B to gauge the magnitude of random sampling errors due to our use of 100 replicates: for instance, rRMSE should be equal for $\theta_A, \theta_B, \theta_E$, and $\theta_F$, and for $\tau_{10}$ and $\tau_{13}$, on the balanced trees.

The four phase-resolution strategies (F, D, P, and R) performed similarly for the Deep trees at the lower rate and when only $S=2$ sequences (or one individual) are sampled per species. We note that with $S=2$ and at the low mutation rate (with heterozygosity at $\theta=0.001$), there will be on average 0.5 heterozygous sites at the same locus, and the probability of having two or more heterozygous sites is $1-0.999^{500}-500\times 0.999^{499}\times 0.001=0.0901$. Then phase resolution will not be a serious issue, and all four strategies examined in the study will be nearly equivalent.

At the high mutation rate ($\theta=0.01$) for the Shallow trees, differences were noted among the strategies even for $S=2$ sequences (Supplementary Fig. S6 available on Dryad and Supplementary Tables S5 and S7 available on Dryad). The PHASE program produced underestimates for the youngest species divergence times ($\tau_{11}$ and $\tau_{14}$ on Shallow B and $\tau_{15}$ on Shallow U) (Fig. 2a′,b′). The biases became more pronounced when $S=4$ sequences per species are in the sample (Fig. 5 and Supplementary Tables S9–S11 available on Dryad). At the high rate, there are on average five heterozygotes per locus in the individual and the probability of having two or more heterozygotes at the locus is 96%. Two factors may be responsible for the bias. First the PHASE program may have inferred heterozygote phase incorrectly (indeed the error rate is comparable to that of random phasing

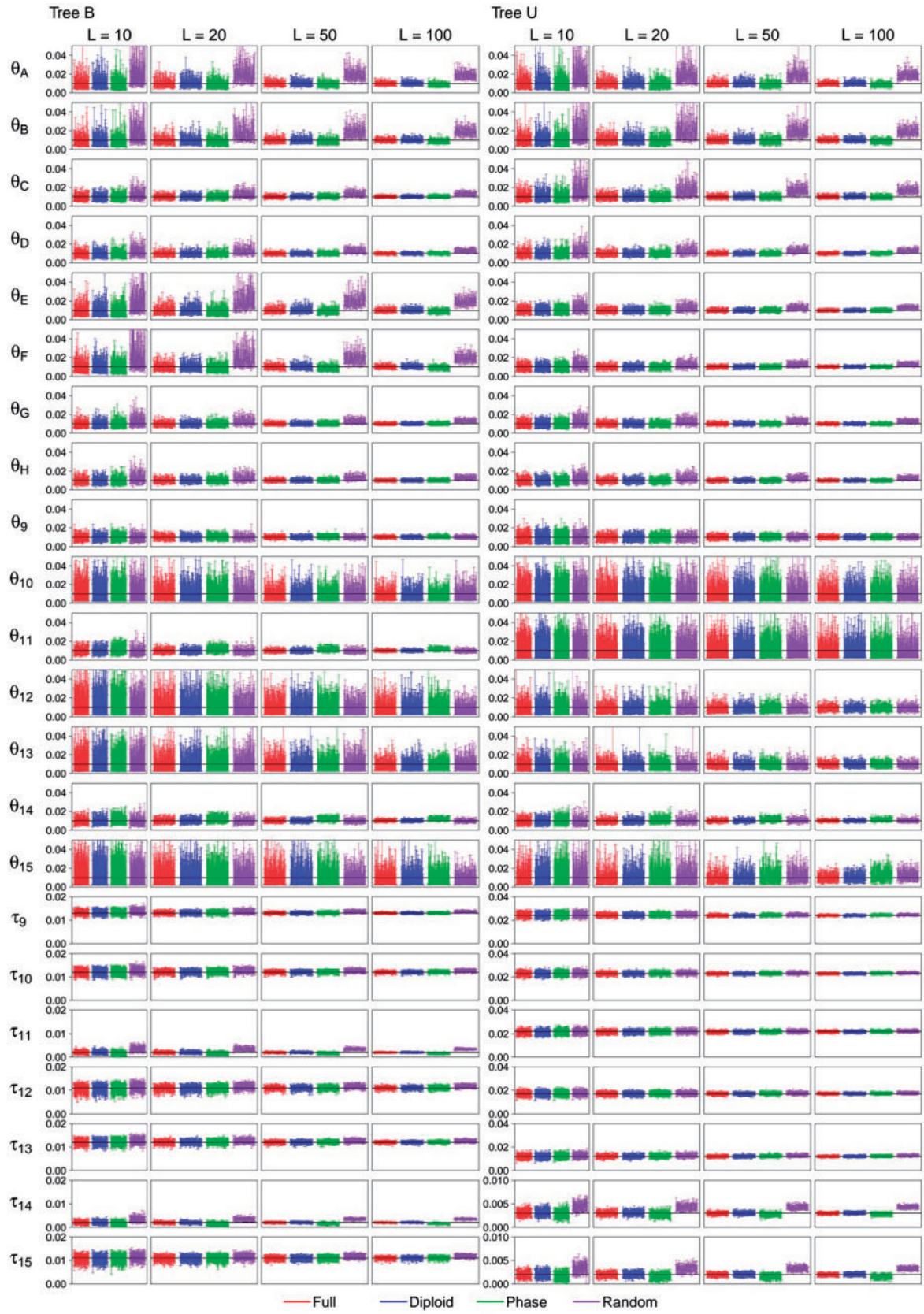


FIGURE 5. (MSC, high rate, shallow, $S=4$) The 95% HPD CIs for parameters for four phase-resolution strategies: F (the full data), D (diploid), P (PHASE), and R (random) in 100 replicate data sets simulated under MSC model trees Shallow B and Shallow U (Fig. 2a′,b′), at the high mutation rate ($\theta=0.01$) and $S=4$ sequences per species. The horizontal black lines indicate the true values. Results for other simulation settings are in Supplementary Figures S5–S11 available on Dryad.

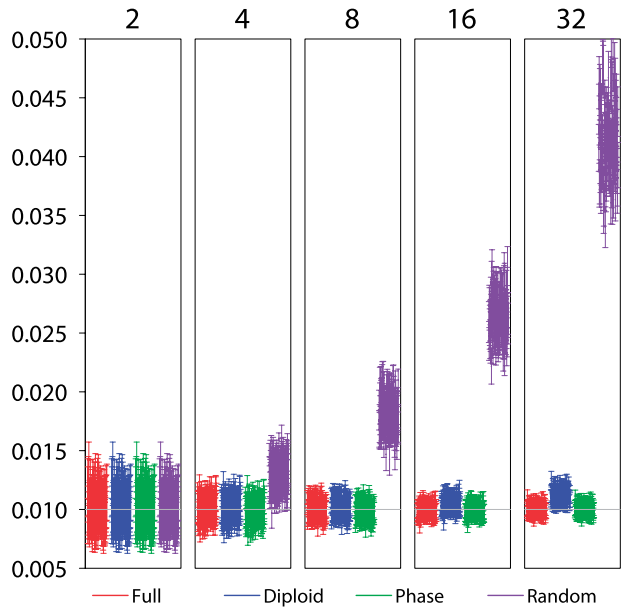


FIGURE 6. The 95% HPD CIs for parameter θ in the single-population coalescent model in 100 replicate datasets using four phase-resolution strategies: F (the full data), D (diploid), P (PHASE), and R (random). There are 100 independent loci in each data set, and at each locus there are $S$ sequences of 500 sites (or $S/2$ diploid individuals), with $S = 2, 4, 8, 16$, and 32. The true parameter value is 0.01.

with $S = 2$). Second PHASE is an MCMC program generating a posterior distribution of different phase resolutions but we used only the optimal resolution. As the optimal resolution may involve the least amount of divergence, this approach is expected to lead to underestimation of sequence divergences or of the τ and θ parameters.

At the high rate and for shallow trees, random phasing (R) also created serious biases, but the biases are in the opposite direction. Random phasing overestimated the youngest species divergence times ($\tau_{11}$ and $\tau_{14}$ on Shallow B and $\tau_{15}$ on Shallow U), and overestimated θ for all modern species. The overestimation of modern θ is most striking, and occurred for both deep and shallow species trees at the high rate and is more dramatic with more sequences ($S = 4$ rather than 2) or more loci.

We examined the number of distinct site patterns in the alignment at each locus for the high-rate data (Supplementary Fig. S12 available on Dryad). Site patterns are compressed for the JC model, so that one site pattern is constant while the others are variable (Yang 2006, p. 144), and the number is thus an indication of the level of sequence divergence. At almost every locus, the PHASE program (P) produced alignments with fewer distinct site patterns than the true phase resolution (e.g., with the mean to be 36.07 compared with the true value 38.51 on tree B), apparently because we used the optimal phase resolution inferred by the program and ignored the less likely ones. Random resolution produced about the same number of site patterns as the true number (average 38.36 vs. 38.51 for tree Deep B). The number of site patterns is thus not the reason for the poor performance of random phasing.

Note that calculation of the heterozygosity for each diploid individual, which is simply the proportion of heterozygous sites in the two sequences at the locus, does not rely on phase resolution. If we calculate the heterozygosity for each diploid individual and then average over individuals of the same species, we will get a reasonably good estimate of θ for that species. However, in the gene-tree based analysis conducted in BPP, each randomly phased haploid sequence is compared not only with the other sequence from the same individual but also with sequences from other individuals through the use of a gene tree relating all phased haploid sequences at the locus. While the true haploid sequences may all be closely related, random phase resolution may generate chimeric sequences that are very different from naturally occurring fully resolved sequences, inflating apparent coalescent times and genetic diversity in the population. This effect is expected to be more serious when more individuals are included in the sample.

*Estimation of θ for a single species.*—To explore this interpretation, we conducted a small simulation sampling independent loci from a single species to estimate the only parameter θ (Fig. 6, Table 3). With $S = 2$ sequences per locus (one diploid individual), the four phase-resolution strategies are equivalent. However, with the increase of $S$, the strategy of random phase resolution becomes increasingly biased. Previously Felsenstein (1992) examined the efficiency of two summary methods based on the number of segregating (variable) sites ($\hat{\theta}_S$; Watterson 1975) and the average pairwise distance ($\hat{\theta}_\pi$; Tajima 1983), relative to the maximum likelihood (ML) method based on gene genealogies. He found that the summary methods ($\hat{\theta}_S$ and $\hat{\theta}_\pi$) were much less efficient than the ML estimate, with orders-of-magnitude differences in the variance in large samples (Felsenstein 1992, Tables 1 and 2), indicating that there is much information about θ in the genealogical histories. The ML method should be very similar to BPP here as both are full likelihood methods. Here, we note that the number of segregating sites does not depend on phase resolutions, and similarly the average proportion of different sites, averaged over all the $S(S-1)/2$ pairwise comparisons, depends on the site configurations at each variable site (such as 10 Ts and 4 Cs) but not on the genotypic phase between different heterozygous sites. Both Waterson's estimator and the average pairwise distance are thus unaffected by phasing errors. It is also noteworthy that those two simple methods are not affected by recombination within the locus, while coalescent-based methods are (Felsenstein 2019). While it is not unexpected that a full likelihood method may be more sensitive to certain errors in the model or in the data than heuristic methods, in this case it is striking that the systematic bias is so large (with estimates to be several times larger than the true value) when the coalescent-based method is applied to randomly phased sequences.

Felsenstein's (1992) analysis, as mentioned above, assumed knowledge of the true gene trees and coalescent times (or equivalently infinitely long sequences at each locus). Here, BPP is applied to sequence alignments and accommodates uncertainties in the genealogical trees. The different methods then have much more similar performance (Table 3, $\hat{\theta}_S$, $\hat{\theta}_\pi$, and BPP strategy F), suggesting that the uncertainties in the genealogical trees due to mutational variations in the sequences have eroded much of the information in the gene trees. The summary methods (in particular, $\hat{\theta}_\pi$) have larger variances than the BPP estimates, especially in large samples of $S=32$ sequences, but the differences are relatively small. We also note that analytical phase integration (D) produced variances that are nearly identical to those for the use of the full data (F).

*Impacts of Other Factors on Parameter Estimation Under the MSC Model.*—We note that different parameters are estimated with very different precision and accuracy, reflecting the different amount of information in the data. Population size parameters ($\theta$s) for modern species are well estimated, as well as $\theta_9$ for the root population, but $\theta$s for other ancestral species, especially those represented by very short branches (e.g., $\theta_{10}, \theta_{13}, \theta_{12}, \theta_{15}$ in tree B) have large errors (Fig. 5, Supplementary Figs. S1–S3 available on Dryad). Species divergence times are all well estimated, with rRMSE to be even much smaller than those for population size parameters for modern species (Supplementary Tables S4–S11 available on Dryad).

Both the mutation rate and the number of loci had a major impact on the estimation of the parameters. For all phasing strategies increasing the number of loci by 10-fold improves performance more than increasing the mutation rate by the same factor (Fig. 5, Supplementary Figs. S1–S3, Tables S4–S11 available on Dryad).

### *Estimation of Introgression Probability under the MSci Model*

We used the MSci models of Figure 2c,c′ to simulate sequence data and used BPP to analyze them to estimate parameters in the MSci model. We are in particular interested in whether the different strategies of heterozygote phase resolution may lead to biases in the estimation of the timing ($\tau_H$) and strength of the introgression ($\varphi$). The results are summarized in Figure 7 and Supplementary Figures S13–S19 available on Dryad and Table 4 and Supplementary Tables S12–S18 available on Dryad.

As before, the diploid strategy (D) produced results almost indistinguishable from the use of the full data (F) in all parameter settings. The performance of the PHASE program (P) and random phasing (R) depends on the mutation rate and, to an lesser extent, on the number of sequences per species $S$. At the low rate, and in particular with only $S=2$ sequences per species, all four strategies have similar performance, but large differences were found at the high mutation rate. Strategy R overestimates the modern $\theta$ and the species divergence times ($\tau$) at the high rate, with the bias being more serious for $S=4$ sequences than for $S=2$. This is the same behavior as discussed earlier in the simulation under the MSC model. Strategy R also tends to overestimate $\varphi$, but the bias is small. Strategy P had the opposite bias and produced underestimates of modern $\theta$ and species divergence times when the mutation rate is high, with smaller biases than for strategy R. Strategy P also underestimates the introgression probability ($\varphi$).

An interesting question is whether each method detects introgression. We calculated the proportion of replicates in which the lower limit of the 95% HPD CI for $\varphi$ exceeds a small value, set somewhat arbitrarily at 0.001. If the CI excludes the small value, we may take it as evidence that $\varphi=0$ is ruled out so that there is significant evidence for introgression. By this measure of power of the Bayesian "test," strategies D and P had nearly identical power as the use of the full data (F), while random resolution (R) had reduced power at the high mutation rate (Table 5, Supplementary Table S19 available on Dryad). Overall, power was very high even with only 10–20 loci and at the low mutation rate. Having more sequences is noted to boost the power of the test for all phase-resolution strategies.

### *Running Time for Different Analyses*

The running time for the A01 analysis under the MSC model (species tree estimation) for the four phasing strategies (F, D, P, and R), averaged over the 100 replicates, is plotted against the number of loci in Supplementary Figure S20 available on Dryad. Running time increases nearly linearly with the number of loci, with the slope being steeper when $S=4$ sequences are sampled per species than for $S=2$. The diploid integration algorithm (D) has the longest running time. Note that the number of parameters in the MSC model, the number of loci, the number of sequences etc. are identical for the four strategies, so that their computational load is proportional to the number of site patterns. As strategy D enumerates all possible phase resolutions (including the true resolution), which may result in many distinct site patterns, it is more expensive than the other methods. The running time for each BPP analysis on a single core ranged from ~20 min for 10 loci to ~5 h for strategy D with data of 100 loci. Strategy P involves running the Bayesian MCMC program PHASE for each of the $L$ loci. At the low mutation rate with very few heterozygous sites per locus, this requires minimal computation (Supplementary Fig. S21 available on Dryad), but at the high rate and with $S=4$ sequences per species, the running time can be comparable with running the subsequent BPP analyses.

The running time for the A00 analysis (parameter estimation) under the MSC and MSci models is shown in

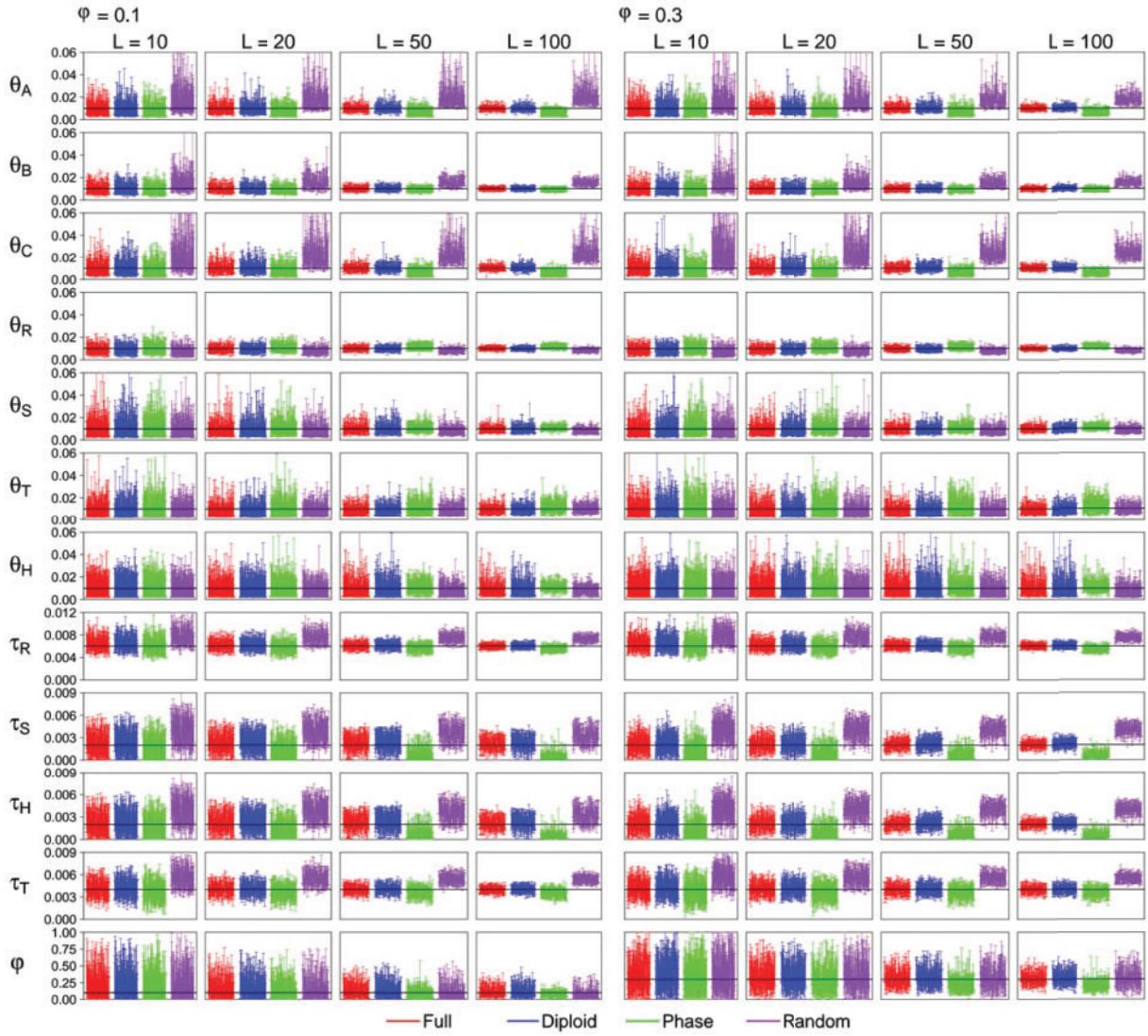


FIGURE 7. (MSci model, high rate, Shallow, $S=4$) The 95% HPD CIs for parameters under the MSci model of Figure 2c′ when $S=4$ sequences are sampled per species. Results for $S=2$ are in Supplementary Figure S14 available on Dryad. See legend to Figure 5.

Supplementary Figures S22–S25 available on Dryad. The A00 analysis under the MSC involves less computation than the A01 analysis as there is no MCMC moves to change the species tree. Overall, the same patterns are observed as discussed above for the A01 analysis. Note that the computer cluster used in this work consist of computers with different processors, so there may be random fluctuations in running time due to the different jobs being assigned to different processors. For example the differences in Supplementary Figures S21 and S23 available on Dryad reflect this random fluctuation as the data were the same.

Overall, the data sets generated in the simulation of this study are relatively small, and the running time for the best method (strategy "D") is comparable with that for strategy "P," in particular if one considers the preprocessing overhead of computational phasing.

### *Analysis of Two Real Data Sets*

We analyzed two real data sets using four different phase-resolution strategies: D (diploid), P (PHASE), R (random), and A (ambiguity). With real data, the option of true phase resolution (F) is unavailable, and the analytical phase resolution (D) is expected to have the best performance, against which we compare the other strategies.

*East Asia brown frogs.*—We reanalyzed a data set of five nuclear loci from the East Asia brown frogs in the *Rana chensinensis* species complex (Zhou et al. 2012) (Fig. 3a). This data set was previously analyzed by Yang (2015) using strategy A. The number of site patterns at each locus is 18–26 for strategy A, and 22–102 for strategy D. Running time using one thread on our server was 3 min for A, 7–8 min for P and R, and 12 min for D.

TABLE 4. (MSci A00 $S=4$, high rate, shallow) Relative root mean square error (rRMSE) for parameter estimates under the Deep MSci model (Fig. 2c′) with $\varphi=0.1$ or 0.3 at the high mutation rate and $S=4$

| | $\theta_A$ | $\theta_B$ | $\theta_C$ | $\theta_R$ | $\theta_S$ | $\theta_T$ | $\theta_H$ | $\tau_R$ | $\tau_S$ | $\tau_T$ | $\varphi$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Truth | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 2 | 0.1 | 0.3 |
| $\varphi$ = 0.1 | | | | | | | | | | | | |
| F, 10L | 0.28 | 0.24 | 0.34 | 0.24 | 0.45 | 0.36 | 0.25 | 0.15 | 0.45 | 0.17 | 0.42 | - |
| D, 10L | 0.31 | 0.25 | 0.36 | 0.24 | 0.48 | 0.36 | 0.22 | 0.16 | 0.46 | 0.18 | 0.41 | - |
| P, 10L | 0.27 | 0.22 | 0.28 | 0.27 | 0.50 | 0.51 | 0.27 | 0.15 | 0.45 | 0.22 | 0.43 | - |
| R, 10L | 0.99 | 0.86 | 1.49 | 0.31 | 0.37 | 0.25 | 0.23 | 0.36 | 1.12 | 0.46 | 0.47 | - |
| F, 20L | 0.24 | 0.21 | 0.28 | 0.18 | 0.40 | 0.35 | 0.28 | 0.10 | 0.41 | 0.13 | 0.46 | - |
| D, 20L | 0.26 | 0.22 | 0.31 | 0.18 | 0.43 | 0.33 | 0.26 | 0.11 | 0.43 | 0.14 | 0.45 | - |
| P, 20L | 0.24 | 0.20 | 0.26 | 0.26 | 0.39 | 0.49 | 0.38 | 0.11 | 0.46 | 0.17 | 0.54 | - |
| R, 20L | 1.09 | 0.73 | 1.75 | 0.27 | 0.32 | 0.25 | 0.22 | 0.30 | 1.04 | 0.41 | 0.53 | - |
| F, 50L | 0.17 | 0.12 | 0.18 | 0.13 | 0.24 | 0.27 | 0.39 | 0.07 | 0.33 | 0.08 | 0.53 | - |
| D, 50L | 0.18 | 0.13 | 0.20 | 0.13 | 0.27 | 0.27 | 0.32 | 0.07 | 0.38 | 0.09 | 0.51 | - |
| P, 50L | 0.31 | 0.13 | 0.34 | 0.22 | 0.21 | 0.41 | 0.29 | 0.09 | 0.60 | 0.14 | 0.68 | - |
| R, 50L | 1.18 | 0.63 | 1.50 | 0.22 | 0.23 | 0.26 | 0.24 | 0.26 | 0.90 | 0.38 | 0.64 | - |
| F, 100L | 0.12 | 0.08 | 0.13 | 0.09 | 0.18 | 0.23 | 0.32 | 0.04 | 0.27 | 0.06 | 0.59 | - |
| D, 100L | 0.14 | 0.08 | 0.16 | 0.09 | 0.19 | 0.24 | 0.30 | 0.04 | 0.33 | 0.06 | 0.57 | - |
| P, 100L | 0.39 | 0.11 | 0.42 | 0.21 | 0.15 | 0.42 | 0.23 | 0.09 | 0.72 | 0.14 | 0.74 | - |
| R, 100L | 1.27 | 0.59 | 1.60 | 0.19 | 0.20 | 0.21 | 0.23 | 0.23 | 0.73 | 0.35 | 0.73 | - |
| $\varphi$ = 0.3 | | | | | | | | | | | | |
| F, 10L | 0.30 | 0.27 | 0.31 | 0.21 | 0.35 | 0.35 | 0.30 | 0.15 | 0.34 | 0.16 | - | 0.48 |
| D, 10L | 0.31 | 0.29 | 0.41 | 0.22 | 0.40 | 0.35 | 0.28 | 0.16 | 0.35 | 0.17 | - | 0.49 |
| P, 10L | 0.27 | 0.24 | 0.27 | 0.22 | 0.48 | 0.55 | 0.35 | 0.15 | 0.48 | 0.21 | - | 0.44 |
| R, 10L | 1.03 | 0.93 | 1.83 | 0.32 | 0.37 | 0.28 | 0.24 | 0.36 | 1.11 | 0.49 | - | 0.57 |
| F, 20L | 0.24 | 0.18 | 0.26 | 0.18 | 0.35 | 0.28 | 0.37 | 0.09 | 0.30 | 0.12 | - | 0.39 |
| D, 20L | 0.28 | 0.20 | 0.33 | 0.19 | 0.33 | 0.27 | 0.33 | 0.10 | 0.33 | 0.13 | - | 0.40 |
| P, 20L | 0.26 | 0.17 | 0.29 | 0.21 | 0.45 | 0.47 | 0.42 | 0.10 | 0.52 | 0.18 | - | 0.37 |
| R, 20L | 1.04 | 0.69 | 1.69 | 0.30 | 0.30 | 0.26 | 0.24 | 0.31 | 1.08 | 0.44 | - | 0.51 |
| F, 50L | 0.18 | 0.12 | 0.16 | 0.11 | 0.22 | 0.28 | 0.48 | 0.05 | 0.21 | 0.10 | - | 0.27 |
| D, 50L | 0.20 | 0.12 | 0.19 | 0.11 | 0.21 | 0.29 | 0.49 | 0.06 | 0.25 | 0.10 | - | 0.30 |
| P, 50L | 0.28 | 0.13 | 0.34 | 0.20 | 0.26 | 0.53 | 0.56 | 0.09 | 0.60 | 0.17 | - | 0.33 |
| R, 50L | 0.97 | 0.61 | 1.56 | 0.24 | 0.21 | 0.24 | 0.25 | 0.29 | 1.07 | 0.41 | - | 0.38 |
| F, 100L | 0.11 | 0.09 | 0.11 | 0.08 | 0.15 | 0.20 | 0.42 | 0.04 | 0.15 | 0.08 | - | 0.20 |
| D, 100L | 0.12 | 0.10 | 0.14 | 0.08 | 0.14 | 0.20 | 0.49 | 0.04 | 0.19 | 0.08 | - | 0.21 |
| P, 100L | 0.34 | 0.12 | 0.41 | 0.20 | 0.19 | 0.47 | 0.37 | 0.09 | 0.66 | 0.15 | - | 0.34 |
| R, 100L | 0.87 | 0.59 | 1.50 | 0.22 | 0.18 | 0.17 | 0.30 | 0.27 | 1.06 | 0.39 | - | 0.32 |

*Note*: Truth represents the true parameter values used in the simulation; values for $\theta$ and $\tau$ are $\times 10^{-2}$. Results for other simulation settings are in Supplementary Tables S12–S18 available on Dryad.

In the A01 analysis (species tree estimation), the four strategies (D, P, R, and A) produced the same MAP tree (Fig. 3a): (((H, L), C), K), with the posterior to be 0.29 for D, 0.36 for P, 0.35 for R, and 0.21 for A. The analysis of Yang (2015) produced a different MAP tree, ((H, L), (C, K)). The difference is due to the use of different priors: Yang (2015) used BPP3.1, with gamma priors on the parameters ($\theta$s for all populations and $\tau$ for the root), whereas here inverse gamma priors are used in BPP4.3. Note that the species trees have low support in both analyses.

In the A00 analysis (parameter estimation under MSC with the MAP species tree fixed), the posterior means and 95% HPD intervals are shown in Table 6a. Strategy P (PHASE) produced similar results to strategy D. Strategy R (random) produced overestimates of $\theta$s for modern species, while strategy A (ambiguity) produced serious underestimates of $\theta$s for modern species and divergence times. The results are consistent with our findings from the simulation.

*Rocky Mountains Chipmunks*.—In the A01 analysis (species tree inference) of the 500 nuclear loci for Rocky Mountains chipmunks, strategies D, P, and R produced the same MAP tree, shown in Figure 3b, with the posterior for every node ~1.0. This is also the species tree inferred by Sarver et al. (2021) using summary methods, although the authors obtained lower support values even with all 1060 loci used. The difference may be due to the higher power of the BPP analysis, which uses the full data of multilocus sequence alignments rather than data summaries (e.g., Shi and Yang 2018; Kim and Degnan 2020; Zhu and Yang 2021). Strategy A (ambiguity) produced a different MAP species tree from the other strategies (Fig. 3b), with the relationship (C,

TABLE 5. (MSci test, shallow) Power of the Bayesian test for introgression (measured by the proportion of replicates in which the lower limit of the 95% HPD CI for $\varphi$ is $>0.001$) when the true model is the Shallow MSci tree

| | low | | | | high | | | |
|---|---|---|---|---|---|---|---|---|
| | 10L | 20L | 50L | 100L | 10L | 20L | 50L | 100L |
| $\varphi = 0.1$ | | | | | | | | |
| $S=2$ seqs per species | | | | | | | | |
| F | 0.42 | 0.41 | 0.42 | 0.49 | 0.48 | 0.52 | 0.64 | 0.89 |
| D | 0.48 | 0.40 | 0.45 | 0.46 | 0.38 | 0.35 | 0.61 | 0.80 |
| P | 0.49 | 0.45 | 0.49 | 0.49 | 0.55 | 0.60 | 0.83 | 0.99 |
| R | 0.51 | 0.44 | 0.47 | 0.48 | 0.27 | 0.25 | 0.40 | 0.57 |
| $S=4$ seqs per species | | | | | | | | |
| F | 0.58 | 0.47 | 0.55 | 0.58 | 0.59 | 0.71 | 0.98 | 0.99 |
| D | 0.56 | 0.44 | 0.54 | 0.60 | 0.56 | 0.66 | 0.94 | 0.99 |
| P | 0.57 | 0.43 | 0.44 | 0.49 | 0.60 | 0.65 | 0.94 | 1.00 |
| R | 0.56 | 0.45 | 0.50 | 0.57 | 0.44 | 0.46 | 0.71 | 0.81 |
| $\varphi = 0.3$ | | | | | | | | |
| $S=2$ seqs per species | | | | | | | | |
| F | 0.68 | 0.74 | 0.87 | 0.94 | 0.86 | 0.97 | 1.00 | 1.00 |
| D | 0.64 | 0.81 | 0.89 | 0.93 | 0.84 | 0.91 | 1.00 | 1.00 |
| P | 0.68 | 0.78 | 0.85 | 0.89 | 0.86 | 0.94 | 0.99 | 1.00 |
| R | 0.66 | 0.72 | 0.87 | 0.92 | 0.74 | 0.82 | 0.96 | 0.99 |
| $S=4$ seqs per species | | | | | | | | |
| F | 0.79 | 0.88 | 0.97 | 1.00 | 0.95 | 1.00 | 1.00 | 1.00 |
| D | 0.83 | 0.86 | 0.95 | 1.00 | 0.92 | 0.99 | 1.00 | 1.00 |
| P | 0.84 | 0.84 | 0.95 | 1.00 | 0.94 | 0.97 | 1.00 | 1.00 |
| R | 0.84 | 0.82 | 0.94 | 0.99 | 0.75 | 0.89 | 1.00 | 1.00 |

*Note*: Results for the Deep MSci model are in Supplementary Table S19 available on Dryad.

(D, (IQR))) instead of (D, (C, (IQR))), with the posterior at 0.94. The running time for the A01 analysis, using eight cores on a server with Intel Xeon Gold 6154 3.0GHz processors, was 9 hours for strategy A, and 16–17 h for strategies D, P, and R, with strategy D having slightly longer running time. The number of site patterns at the 1060 loci for strategy D is shown in Supplementary Fig. S26 available on Dryad. Strategy P also needed the additional time for running the PHASE program, which was 33 min to phase all 1060 loci using one thread on the server.

In the A00 analysis (parameter estimation), strategy P (PHASE) produced nearly identical results to strategy D (diploid) (Table 6). Compared with strategy D, strategy R (random) produced overestimates of $\theta$s for modern species, while divergence times for recent nodes were also overestimated very slightly. Strategy A (ambiguity) produced serious underestimates of $\theta$s for modern species, with divergence times, especially of recent nodes, to be underestimated as well. Those results mimic our findings about the relative performance of the different strategies in the simulated data. Running time for the A00 analysis was 2.5 h for strategy A, and 5–6 h for strategies D, P, and R. Note that in the A00 analysis the chain is only half as long as in the A01 analysis.

## DISCUSSION

### *The Impact of Phasing Errors Depends on the Inference Problem*

We have used simulation to examine the performance of four different strategies for handling heterozygote phase in genomic sequence data: F (full phased data), D (diploid analytical phase integration), P (PHASE), and R (random). Inference problems examined have included species tree estimation under the MSC model and parameter estimation under the MSC and MSci models. We found that the different strategies, including random phase resolution (or equivalently the use of haploid consensus sequences), did not affect species tree estimation when the species divergences are much older than the coalescent times. The different phasing strategies may be expected to have even less impact on inference of deep phylogenies, where within-species polymorphism is much lower than between-species divergence. However, species tree estimation is affected by phasing errors if the species tree is shallow and between-species divergence is similar to within-species polymorphism, if the mutation rate is high so that there are many heterozygote sites in the sequence, and if many sequences are sampled from each species. Phasing errors are clearly important when genomic data are used to infer the divergence history of populations of the same species. Previously, Kates et al. (2018) used several summary methods (including astral and concatenation/maximum likelihood) to analyze a data set of breadfruit, jackfruit, and relatives (333 loci from 23 species of the genus *Artocarpus*) to infer the species phylogeny, evaluating several phase-resolution strategies such as haploid consensus sequences, ambiguity encoding, and a bioinformatics procedure to incorporate read information to phase alleles. The authors found that heterozygote phase resolution had minimal impact on phylogeny reconstruction. The results are consistent with our simulation, which also found that phasing errors do not influence estimation of deep species phylogenies. However, our simulation also suggests that the authors' conclusion based on the deep species tree of *Artocarpus*, with an estimated age of 40 Myr for the crown age of the ingroup, may not apply to inference of shallow phylogenies, in which species divergence times are similar to average coalescent times.

We found that estimation of parameters in the MSC and MSci models is more sensitive to phasing errors than is species tree estimation. In particular, population sizes for modern species are seriously overestimated under the MSC and MSci models when random phasing or haploid consensus sequences are used. Our analysis of the simple case of estimating $\theta$ under the single-population coalescent suggests that the bias is caused mainly by the unusual sequences generated by random phase resolution (Fig. 6 and Table 3). Estimates of the introgression probability and introgression time under the MSci model may also be biased by errors in random phasing. The biases are more serious when the mutation

TABLE 6. Posterior means and 95% HPD CIs for parameters under the MSC model for the east Asian brown frogs and for the chipmunks

| | Diploid (D) | PHASE (P) | Random (R) | Ambiguity coding (A) |
|---|---|---|---|---|
| (a) East Asian brown frogs (Fig. 3a) | | | | |
| $\theta_K$ | 4.94 (2.62, 7.65) | 6.35 (3.32, 9.80) | 6.81 (3.64, 10.48) | 2.78 (1.06, 4.99) |
| $\theta_C$ | 20.57 (11.78, 30.66) | 22.84 (12.98, 34.29) | 32.00 (18.11, 48.37) | 5.65 (2.41, 9.71) |
| $\theta_L$ | 10.82 (6.32, 15.94) | 10.12 (5.98, 14.76) | 12.33 (7.45, 17.79) | 6.73 (2.37, 12.29) |
| $\theta_H$ | 3.73 (1.42, 6.73) | 3.35 (1.29, 5.96) | 5.39 (1.83, 10.16) | 1.18 (0.29, 2.51) |
| $\theta_5$ | 5.13 (2.20, 8.43) | 5.49 (2.54, 8.82) | 4.41 (1.50, 7.49) | 4.56 (1.74, 7.84) |
| $\theta_6$ | 2.21 (0.21, 6.53) | 2.00 (0.20, 5.76) | 2.65 (0.22, 7.87) | 1.85 (0.21, 5.32) |
| $\theta_7$ | 1.72 (0.20, 4.61) | 1.43 (0.21, 3.62) | 1.54 (0.23, 3.94) | 1.28 (0.19, 3.15) |
| $\tau_5$ | 2.14 (1.57, 2.73) | 2.00 (1.50, 2.53) | 2.49 (1.89, 3.17) | 1.37 (0.86, 1.93) |
| $\tau_6$ | 2.03 (1.53, 2.54) | 1.91 (1.45, 2.38) | 2.30 (1.79, 2.83) | 1.23 (0.78, 1.70) |
| $\tau_7$ | 1.85 (1.28, 2.44) | 1.77 (1.27, 2.27) | 2.15 (1.60, 2.72) | 1.11 (0.65, 1.61) |
| (b) Rocky Mountains chipmunks (Fig. 3b) | | | | |
| $\theta_Q$ | 0.81 (0.70, 0.93) | 0.83 (0.72, 0.94) | 0.93 (0.81, 1.05) | 0.39 (0.31, 0.47) |
| $\theta_I$ | 0.78 (0.67, 0.89) | 0.81 (0.69, 0.91) | 0.94 (0.81, 1.07) | 0.26 (0.21, 0.32) |
| $\theta_R$ | 0.36 (0.30, 0.41) | 0.36 (0.30, 0.41) | 0.37 (0.31, 0.42) | 0.32 (0.25, 0.39) |
| $\theta_C$ | 0.47 (0.38, 0.55) | 0.47 (0.39, 0.55) | 0.50 (0.42, 0.58) | 0.48 (0.39, 0.56) |
| $\theta_D$ | 1.79 (1.61, 1.98) | 1.79 (1.60, 1.97) | 2.05 (1.84, 2.26) | 0.67 (0.57, 0.77) |
| $\theta_U$ | 1.04 (0.93, 1.15) | 1.04 (0.93, 1.15) | 1.06 (0.95, 1.17) | 0.83 (0.73, 0.94) |
| $\theta_S$ | 0.79 (0.67, 0.90) | 0.79 (0.67, 0.90) | 0.84 (0.71, 0.96) | 0.34 (0.25, 0.43) |
| $\theta_8$ | 9.94 (8.31, 11.54) | 10.03 (8.61, 11.45) | 9.95 (8.32, 11.55) | 10.05 (8.62, 11.43) |
| $\theta_9$ | 1.24 (1.02, 1.47) | 1.24 (1.02, 1.45) | 1.24 (1.01, 1.46) | 1.24 (1.01, 1.46) |
| $\theta_{10}$ | 1.01 (0.65, 1.39) | 1.06 (0.68, 1.44) | 0.99 (0.64, 1.34) | 1.06 (0.63, 1.50) |
| $\theta_{11}$ | 4.33 (0.33, 9.43) | 5.13 (0.77, 10.38) | 2.87 (0.35, 5.87) | 2.08 (0.20, 5.90) |
| $\theta_{12}$ | 2.43 (0.50, 4.62) | 1.84 (0.34, 3.68) | 2.16 (0.57, 3.74) | 2.45 (0.69, 4.38) |
| $\theta_{13}$ | 0.51 (0.21, 0.86) | 0.54 (0.19, 0.91) | 0.49 (0.21, 0.80) | 0.90 (0.34, 1.54) |
| $\tau_8$ | 3.83 (3.30, 4.50) | 3.80 (3.30, 4.24) | 3.85 (3.30, 4.48) | 3.70 (3.19, 4.23) |
| $\tau_9$ | 1.04 (0.95, 1.14) | 1.04 (0.95, 1.13) | 1.04 (0.95, 1.13) | 0.92 (0.82, 1.01) |
| $\tau_{10}$ | 0.74 (0.67, 0.81) | 0.72 (0.65, 0.79) | 0.75 (0.68, 0.81) | 0.59 (0.52, 0.67) |
| $\tau_{11}$ | 0.58 (0.45, 0.71) | 0.52 (0.42, 0.63) | 0.60 (0.49, 0.71) | 0.55 (0.44, 0.64) |
| $\tau_{12}$ | 0.41 (0.35, 0.46) | 0.41 (0.35, 0.46) | 0.43 (0.38, 0.49) | 0.34 (0.26, 0.43) |
| $\tau_{13}$ | 0.33 (0.29, 0.38) | 0.34 (0.29, 0.37) | 0.37 (0.33, 0.41) | 0.21 (0.16, 0.26) |

*Note*: All values are multiplied by 1000.

rate is high so that there are multiple heterozygote sites at each locus and when multiple sequences are sampled per species. Those results are consistent with Gronau et al. (2011), who also found that random phase resolution affected parameter estimation in their analysis of genomic sequence data from different human populations.

### *Limitations of our Simulation and Implications to Practical Data Analysis*

Here, we note a few limitations of our study. First, we have examined only one inference method, the Bayesian method implemented in the BPP program. Our results may be expected to apply to other full likelihood implementations such as StarBeast (Ogilvie et al. 2017; Zhang et al. 2018). or PhyloNet-Seq (Wen and Nakhleh 2018), but may not apply to summary methods. Similarly, we considered only a few inference problems under the MSC and MSci models using genomic sequence data. We have not examined the impact of phasing errors on inference of population demographic changes or on inference of migration/introgression histories (our simulation under the MSci model assumed a fixed introgression event).

Given those caveats, we discuss the implications of our simulation results to practical data analysis. First, our simulation as well as those of Gronau et al. (2011) and Andermann et al. (2019) suggest that random phase resolution or the use of haploid consensus sequences should be avoided. Strategy R never performed better than computational phasing (strategy P) in our simulations. Similarly strategy A (ambiguity) should not be recommended. Virtually all phylogenetic likelihood programs accommodate ambiguities in a sequence alignment representing undetermined nucleotides using a data-augmentation algorithm in the likelihood calculation (Felsenstein 2004, pp. 255–6; Yang 2014, pp. 110–112). As this approach misinterprets heterozygotes as ambiguities, it leads to underestimation of polymorphism or $\theta$ for the modern species. Bias may also be introduced into estimates of other parameters, such as underestimation of divergence times (Andermann et al. 2019). The approach also underestimates the information content in the data, as it in effect treats two sequences (although unphased) as only one. This mistake in the treatment of the data was made by

Rannala and Yang (2003) in the analysis of three human noncoding loci of Zhao et al. (2000), Yu et al. (2001), and Makova et al. (2001), and by Yang (2015) in the analysis of the five nuclear loci from East Asian brown frogs (Zhou et al. 2012). The mistake is easy to see from the occurrence of the same ambiguity character (such as Y) in multiple sequences at the same site in the alignment.

Strategy D (diploid analytical integration) produced results that are extremely similar to the use of the full data (F) in all simulation settings of this study (see also Gronau et al. 2011). As the algorithm averages over all possible phase resolutions and constitutes a full likelihood approach to handling missing data, it is the optimal statistical approach when the data consist of unphased diploid sequences, and may thus be recommended in general, even for inference problems that are not examined in our simulation study. As a statistical inference method, strategy D is equivalent to the approach of sampling phase resolutions in a Markov chain Monte Carlo (MCMC) algorithm (Kuhner and Felsenstein 2000), but the approach of analytical phase integration appears to have a computational advantage.

The computational effort required by strategy F is proportional to the number of distinct site patterns at each locus that may result from enumerating all possible phase resolutions. For the data sets simulated in this article as well as the two real data sets analyzed here, the computational load for strategy D is comparable to (if not better than) that for strategy P, in particular when one considers the preprocessing of computational phasing required by strategy P. In such cases, strategy D should be recommended. However, when there are many long sequences of high heterozygosity at a locus, enumeration of all phase resolutions may lead to a huge number of site patterns. For example, the three noncoding regions of human DNA analyzed by Rannala and Yang (2003) have about 60 sequences per locus, with $\sim 10^4$ sites. The number of site patterns is 50–73 in the unphased alignments (strategy A), but reaches 1.2–4.4 million for strategy D, rendering the analysis unfeasible. Note that those loci are long genomic segments, which may be affected by recombination, whereas data sets suitable for analysis under the MSC typically involve much shorter genomic segments (e.g., Burgess and Yang 2008).

We suggest that computational phasing (strategy P) should be an acceptable alternative when strategy D is computationally unfeasible. In our analyses of the simulated and real data sets, strategy P produced similar results to the use of full data (F) or the analytical phase integration approach (D), with very small biases. Note that the Bayesian program PHASE assumes a population genetics model and is designed for sequence or allelic data from the same species. However, our use of it to analyze sequence data from multiple species produced relatively small biases in parameter estimation in both simulated data and in the two real data sets, much better than random phase resolution or haploid consensus sequences. We also note that phasing based on reads combined with bioinformatic analysis shows great promise (Andermann et al. 2019). In particular, exciting developments in sequencing technology to provide longer reads, combined with computational algorithms (Porubsky et al. 2020; Zhou et al. 2020; Cheng et al. 2021), may soon make it practical to produce routinely fully phased diploid genomes.

## Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.vmcvdncrd.

## Acknowledgments

We thank two anonymous reviewers, the AEs (Laura Kubatko and Robert Thomson) and the EiC (Bryan Carstens) for constructive comments. We thank Jiayi Ji for converting the chipmunk alignments of Sarver et al. (2021) from the Nexus format to the BPP format.

## Funding

This work was supported by a Biotechnology and Biological Sciences Research Council grant [BB/P006493/1 to Z.Y.]; and a BBSRC equipment grant [BB/R01356X/1]; a National Science Foundation grant [NSF-SBS-2023723 to A.D.L.]. J.H.'s visit to London is supported by China Scholarship Council (CSC).

## References

Andermann T., Fernandes A.M., Olsson U., Topel M., Pfeil B., Oxelman B., Aleixo A., Faircloth B.C., Antonelli A. 2019. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. Syst. Biol. 68(1):32–46.

Andres A.M., Clark A.G., Shimmin L., Boerwinkle E., Sing C.F., Hixson J.E. 2007. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. Genet. Epidemiol. 31(7):659–671.

Andrews K.R., Good J.M., Miller M.R., Luikart G., Hohenlohe P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17(2):81–92.

Browning S.R., Browning B.L. 2011. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 12(10):703—714.

Burgess R., Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25:1979–1994.

Cheng H., Concepcion G.T., Feng X., Zhang H., Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods. 18:170–175.

Choi Y., Chan A.P., Kirkness E., Telenti A., Schork N.J. 2018. Comparison of phasing strategies for whole human genomes. PLoS Genet. 14(4):e1007308.

Eaton D.A.R., Spriggs E.L., Park B., Donoghue M.J. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66(3):399–412.

Edwards S., Cloutier A., Baker A. 2017. Conserved nonexonic elements: a novel class of marker for phylogenomics. Syst. Biol. 66(6):1028–1044.

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61(5):717–726.

Felsenstein J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. 59:139–147.
Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.
Felsenstein J. 2019. Theoretical evolutionary genetics. https://evolution.genetics.washington.edu/pgbook/pgbook.html.
Flot J.F. 2010. SeqPhase: a web tool for interconverting phase input/output files and FASTA sequence alignments. Mol. Ecol. Resour. 10(1):162–166.
Flouri T., Jiao X., Rannala B., Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35(10):2585–2593.
Flouri T., Jiao X., Rannala B., Yang Z. 2020a. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Mol. Biol. Evol. 37(4):1211–1223.
Flouri T., Rannala B., Yang Z. 2020b. A tutorial on the use of bpp for species tree estimation and species delimitation. In: Scornavacca C., Delsuc F., Galtier N., editors. Phylogenetics in the genomic era, book section 5.6. No Commercial Publisher. p. 5.6.1–16.
Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29(7):644–652.
Gronau I., Hubisz M.J., Gulko B., Danko C.G., Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43:1031–1034.
Huang J., Flouri T., Yang Z. 2020. A simulation study to examine the information content in phylogenomic datasets under the multispecies coalescent model. Mol. Biol. Evol. 37:3211–3224.
Huelsenbeck J., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.
Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro H., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.
Karin B.R., Gamble T., Jackman T.R. 2020. Optimizing phylogenomics with rapidly evolving long exons: comparison with anchored hybrid enrichment and ultraconserved elements. Mol. Biol. Evol. 37(3):904–922.
Kates H.R., Johnson M.G., Gardner E.M., Zerega N.J.C., Wickett N.J. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of artocarpus. Am. J. Bot. 105(3):404—-416.
Kim A., Degnan J. 2020. Pranc: Ml species tree estimation from the ranked gene trees under coalescence. Bioinformatics. 36:4819–4821.
Kuhner M.K., Felsenstein J. 2000. Sampling among haplotype resolutions in a coalescent-based genealogy sampler. Genet. Epidemiol. 19(Suppl):S15–S21.
Leaché A.D., Oaks J.R. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Ann. Rev. Ecol. Evol. Syst. 48:69–84.
Leaché A.D., Zhu T., Rannala B., Yang Z. 2019. The spectre of too many species. Syst. Biol. 68(1):168–181.
Lemmon A.R., Emme S.A., Lemmon, E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61(5):727–744.
Makova K.D., Ramsay M., Jenkins T., Li W.H. 2001. Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics 158:1253–1268.
Niu T., Qin Z.S., Xu X., Liu J.S. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet. 70(1):157–169.
Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34(8):2101–2114.
Porubsky D., Ebert P., Audano P.A., Vollger M.R., Harvey W.T., Marijon P., Ebler J., Munson K.M., Sorensen M., Sulovari A., Haukness M., Ghareghani M., Human Genome Structural Variation Consortium, Lansdorp P.M., Paten B., Devine S.E., Sanders A.D., Lee C., Chaisson M.J.P., Korbel J.O., Eichler E.E., Marschall T. 2020. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. Nat. Biotechnol. 39:302–308.
Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Rannala B., Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Syst. Biol. 66:823–842.
Sarver B.A.J., Herrera N.D., Sneddon D., Hunter S.S., Settles M.L., Kronenberg Z., Demboski J.R., Good J.M., Sullivan J. 2021. Diversification, introgression, and rampant cytonuclear discordance in rocky mountains chipmunks (sciuridae: Tamias). Syst. Biol. 70:908–921.
Scheet P., Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78(4):629–644.
Shi C., Yang Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol. Biol. Evol. 35:159–179.
Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J., Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123.
Stephens M., Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. 73:1162–1169.
Stephens M., Smith N.J., Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68:978–989.
Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.
Tewhey R., Bansal V., Torkamani A., Topol E.J., Schork, N.J. 2011. The importance of phase information for human genomics. Nat. Rev. Genet. 12:215–223.
Watterson G. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276.
Weisenfeld N.I., Kumar V., Shah P., Church D.M., Jaffe, D.B. 2017. Direct determination of diploid genome sequences. Genome Res. 27(5):757–767.
Wen D., Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst. Biol. 67(3):439–457.
Yang Z. 2006. Computational molecular evolution. Oxford, UK: Oxford University Press.
Yang Z. 2014. Molecular evolution: a statistical approach. Oxford, England: Oxford University Press.
Yang Z. 2015. The BPP program for species tree estimation and species delimitation. Curr. Zool. 61(5):854–865.
Yang Z., Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.
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(12):3125–3135.
Yang Z., Kumar S., Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650.
Yu N., Zhao Z., Fu Y.X., Sambuughin N., Ramsay M., Jenkins T., Leskinen E., Patthy L., Jorde L.B., Kuromori T., Li W.H. 2001. Global patterns of human dna sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. 18:214–222.
Zerbino D.R., Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.
Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018. Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35:504–517.
Zhao Z., Jin L., Fu Y.X., Ramsay M., Jenkins T., Leskinen E., Pamilo P., Trexler M., Patthy L., Jorde L.B., Ramos-Onsins S., Yu N., Li W.H. 2000. Worldwide DNA sequence variation in a 10-kilobase

noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA 97:11354–11358.

Zhou Q., Tang D., Huang W., Yang Z., Zhang Y., Hamilton J.P., Visser R.G.F., Bachem C.W.B., Robin Buell C., Zhang Z., Zhang C., Huang S. 2020. Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat. Genet. 52(10):1018–1023.

Zhou W.W., Wen Y., Fu J., Xu Y.B., Jin J.Q., Ding L., Min M.S., Che J., Zhang Y.P. 2012. Speciation in the *Rana chensinensis* species complex and its relationship to the uplift of the Qinghai-Tibetan plateau. Mol. Ecol. 21(4):960–973.

Zhu T., Yang Z. 2021. Complexity of the simplest species tree problem. Mol. Biol. Evol. doi:10.1093/molbev/msab009.