MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 193–205 doi: 10.1111/1755-0998.12435

Posterior predictive checks of coalescent models: P2C2M, an R package

MICHAEL GRUENSTAEUDL,*‡ NOAH M. REID,† GREGORY L. WHEELER* and BRYAN C. CARSTENS*

**Department of Evolution, Ecology & Organismal Biology, Ohio State University, Columbus, OH 43210, USA†Department of Environmental Toxicology, University of California, Davis, CA 95616, USA*

Abstract

Bayesian inference operates under the assumption that the empirical data are a good statistical fit to the analytical model, but this assumption can be challenging to evaluate. Here, we introduce a novel R package that utilizes posterior predictive simulation to evaluate the fit of the multispecies coalescent model used to estimate species trees. We conduct a simulation study to evaluate the consistency of different summary statistics in comparing posterior and posterior predictive distributions, the use of simulation replication in reducing error rates and the utility of parallel process invocation towards improving computation times. We also test P2C2M on two empirical data sets in which hybridization and gene flow are suspected of contributing to shared polymorphism, which is in violation with the coalescent model: *Tamias* chipmunks and *Myotis* bats. Our results indicate that (i) probability-based summary statistics display the lowest error rates, (ii) the implementation of simulation replication decreases the rate of type II errors, and (iii) our R package displays improved statistical power compared to previous implementations of this approach. When probabilistic summary statistics are used, P2C2M corroborates the assumption that genealogies collected from *Tamias* and *Myotis* are not a good fit to the multispecies coalescent model. Taken as a whole, our findings argue that an assessment of the fit of the multispecies coalescent model should accompany any phylogenetic analysis that estimates a species tree.

Keywords: multispecies coalescent, phylogeography, posterior predictive simulation, species trees

Received 24 November 2014; revision received 22 May 2015; accepted 26 May 2015

Introduction

In the presence of rapid speciation, the stochastic process of allele coalescence produces discordant genealogies across neutral loci (Hudson & Turelli 2003). Instead of taking such genealogies at face value, researchers can apply a coalescent model (Kingman 1982) to estimate different parameters. For phylogeographers interested in the pattern and timing of population divergence, inference of the species tree is particularly useful (Maddison 1997). While there are several variations of the species tree model (Maddison & Knowles 2006; Ané *et al.* 2007; Liu & Pearl 2007; Kubatko *et al.* 2009; Than & Nakhleh 2009; Heled & Drummond 2010), most operate under the simplifying assumption that genetic polymorphism shared among terminal populations results from incompletely sorted ancestral polymorphism. Species trees have been integral to hundreds of empirical studies and have become an essential model for phylogeographic investigations (Carstens *et al.* 2013). However, stochasticity in the segregation of loci is not the only factor that influences genealogical sorting (Maddison 1997). In addition to the migration of individuals among populations, natural processes (collectively referred to as gene flow) can also affect the pattern of allele coalescence, including horizontal gene transfer (Chung & Ané 2011) and hybridization (Mallet 2007).

Researchers using simulated data have demonstrated that some methods of species tree inference have reduced accuracy in the presence of gene flow. Moderate levels of gene flow can lead to an underestimation of lineage divergence times (Leaché *et al.* 2014), while gene flow at higher levels can lead to inaccuracies in the estimation of the topology (Eckert & Carstens 2008). Some coalescent-based approaches, such as STEM-Hy

Correspondence: Michael Gruenstaeudl, Fax: +1 614 292 2030; E-mail: gruenstaeudl.1@osu.edu
‡Present address for Michael Gruenstaeudl: Institut für Biologie-Botanik, Dahlem Centre of Plant Sciences, Freie Universität Berlin, Altensteinstraße 6, 14195 Berlin, Germany.

(Kubatko 2009), estimate species trees in the face of hybrid taxa, but no current implementation can coestimate the phylogeny (consisting of topology, branch lengths and population size) and the rate of gene flow (but see Pickrell & Pritchard 2012). Methods that estimate gene flow require either a defined phylogeny (e.g. IMa2; Hey 2010) or utilize an *n*-island model that does not account for temporal divergence (e.g. Migrate-n; Beerli & Felsenstein 2001). Researchers who suspect that gene flow is responsible for some of the shared polymorphism evident in their data are thus confronted with a difficult choice: they can use a coalescent method to estimate gene flow and not attempt to estimate the topology of divergence, or, alternatively, they can estimate a species tree while ignoring gene flow and hope that this process, if present, does not influence their phylogeny estimate. Because we were unhappy with either of these options, we develop here a third approach, based on posterior predictive simulation, that allows users to check whether their data are a good fit to the multispecies coalescent model (MSCM).

Model checking should be an essential component of phylogenetic inference (Goldman 1993), and posterior predictive approaches enable such exploratory data analyses in the Bayesian framework (Gelman 2003). In essence, posterior predictive simulation is a Bayesian version of the parametric bootstrap, which has long been applied in the phylogeographic context (e.g. Sullivan *et al.* 2000; Knowles 2001). Such an evaluation allows researchers to assess model adequacy and may allow them to learn how a model does not fit the data (Gelman & Shalizi 2012). In the case of the multispecies coalescent, researchers can discover when patterns inherent to the data (the observed genealogies) are inconsistent with model assumptions (e.g. that shared polymorphism results from incomplete lineage sorting).

Posterior predictive checks in Bayesian phylogenetics were introduced by Huelsenbeck *et al.* (2001) in the context of assessing the adequacy of the models of sequence evolution, which are essential to the calculation of the posterior distribution in Bayesian inference. Work on assessing the fit of sequence evolution models has continued, with recent authors introducing posterior predictive approaches to evaluating the fit of models of sequence evolution in Bayesian phylogenetic inference (Lewis *et al.* 2014), and the development of new statistics for detecting cases where model misspecification negatively influences phylogeny estimation (Brown 2014). Other authors have applied these approaches to assessing the fit of more complex phylogenetic models. For example, Reid *et al.* (2014) used posterior predictive simulation (PPS) to demonstrate that in many cases empirical data do not fit the MSCM. They developed a method to measure the fit of genealogies estimated from DNA sequence data to the MSCM as implemented by *BEAST (Heled & Drummond 2010). Twenty-five empirical data sets were evaluated using PPS, of which the majority displayed poor fit to the MSCM. Reid *et al.* (2014) published several R scripts as part of their investigation, but did not conduct full simulation testing due to difficulties inherent in automating *BEAST analyses.

Here, we introduce a novel R package based on the original scripts by Reid *et al.* (2014) and present results from analyses on simulated and empirical data that illustrate the importance of several package features. The package was developed in the statistical language R (R Development Core Team 2013) for the purpose of a comprehensive evaluation of model fit to the MSCM in a Bayesian context. It is titled ‘Posterior Predictive Checks of Coalescent Models’ (P2C2M) and allows users to evaluate the fit of empirical or simulated data to the MSCM. The package was designed to be modular and as such is easily customizable. It features several enhancements over the previous scripts, including a greater number of available summary statistics, a replication method designed to approximate the full range of coalescent stochasticity, the ability to speed up calculations by invoking parallel processes using multiple computer CPUs and data parsers to allow input files from several different versions of *BEAST. P2C2M is supplemented with a software script (‘BEAUTiAutomator.py’) that automates the set-up of input files. Using these novel features, we conduct simulation testing to evaluate (i) the applicability and consistency of different summary statistics, (ii) the use of simulation replication to control for high levels of coalescence stochasticity in order to reduce error rates and (iii) the speed improvements achieved using multiple, linked computer processors. Finally, we evaluate the fit of two empirical data sets to the MSCM via P2C2M and demonstrate that our software offers improved statistical power compared to the original scripts by Reid *et al.* (2014).

Methods

Evaluation of model fit via PPS

In order to evaluate the fit of data to the MSCM in a posterior predictive framework, a posterior distribution of gene genealogies, estimated from empirical data, is compared to a posterior predictive distribution of genealogies produced via simulation. P2C2M constructs the posterior predictive distribution by sampling species trees from the Markov chain generated by *BEAST and transferring their information on topology, branch lengths and population size to a software able to simulate new genealogies under the MSCM (MS, Hudson 2002; Fig. 1a). The comparison of the posterior and the

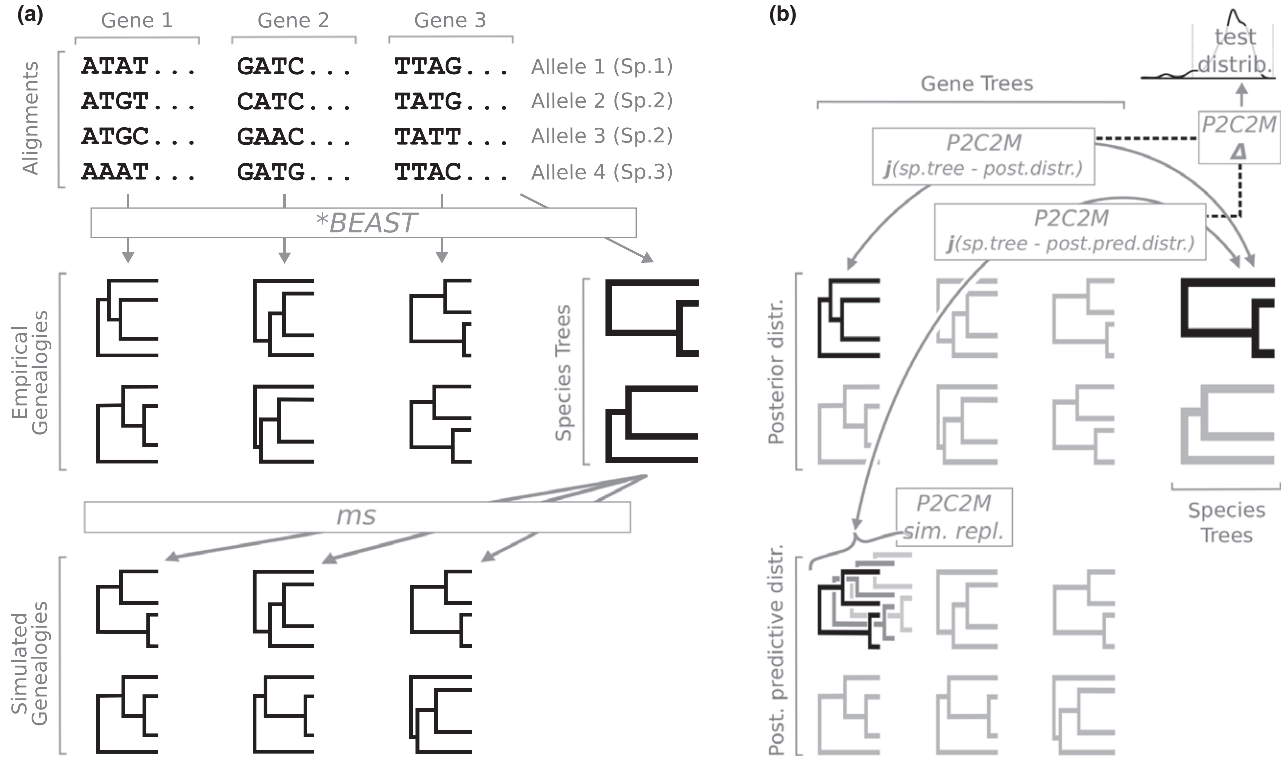

Fig. 1 Schematic of the analysis procedure prior to and during the application of P2C2M. (a) Overview of the combined inference of genealogies and species trees in *BEAST and the simulation of genealogies under the MSCM with MS. (b) Overview of the comparison of genealogies from the posterior distribution to the species trees and from the posterior predictive distribution to the species trees, respectively, via a hypothetical summary statistic *j*. Simulation replication is conducted during the generation of the posterior predictive distribution. The test distribution is formed by the difference values resulting from the comparison of the summary statistics generated on the posterior distribution to those generated on the posterior predictive distribution.

posterior predictive distributions is conducted by sampling from each of these distributions, inferring summary statistics from the samples and generating a test distribution by calculating the difference of the paired summary statistic values (Fig. 1b). When samples are drawn from data with a good fit to the MSCM, the summary statistics from each distribution should be approximately equal and their difference near zero (Reid *et al.* 2014).

A central aspect of evaluating model fit for multilocus data in a posterior predictive framework is the joint sampling of gene and species trees. In *BEAST, the MSCM is implemented as a parameter-rich, hierarchical model that can be used to simultaneously estimate gene genealogies and species trees from multilocus DNA sequence data. Ultrametric coalescent genealogies are estimated from aligned sequence data in a combined inference of tree topologies, branch lengths and coalescence events using models of sequence evolution and molecular clock models (Heled & Drummond 2010). Species trees are estimated concurrently by MCMC sampling from the joint parameter distribution (Fig. 1a). To maintain model congruence across the inference and simulation of genealogies, a model of piecewise constant rate change is used for the estimation of effective population sizes (Heled & Drummond 2008). Upon species tree inference, P2C2M can draw values from the joint posterior distribution while simulating a posterior predictive distribution for each locus (Fig. 1b).

Markov chain convergence and independent sampling

Bias in the initial stages of MCMC sampling is not automatically accounted for in a posterior predictive framework. Therefore, it is advisable to (i) remove a sufficient number of initial MCMC generations (burn-in) to ensure that the posterior tree distribution has reached stationarity (Felsenstein 2004) and (ii) subsample from the remaining generations (Reid *et al.* 2014). Based on the results of preliminary analyses, we selected a burn-in of 20% of all MCMC generations and then subsampled every fourth generation to generate a set of 1000 genealogies per posterior distribution. Markov chain convergence and independent sampling of generations were evaluated with summary statistics in TRACER v.1.6 (Rambaut *et al.* 2014).

Summary and test statistics in P2C2M

Various test statistics have been applied by previous investigations to compare the posterior and the posterior predictive distributions in the PPS framework (Bollback 2002; Joly 2012; Brown 2014; Reid *et al.* 2014), and P2C2M includes a set of four descriptive summary statistics. Each measures a certain aspect of the gene tree/species tree relationship and, depending on the data under study, may have different rates of error. The four statistics are the number of deep coalescences (Maddison 1997), the genealogical sorting index (Cummings *et al.* 2008) and two probability-based statistics based on the 'coalescent likelihood' (i.e. the product of individual gene tree densities conditional on species tree branches calculated across the species tree; Rannala & Yang 2003; Liu *et al.* 2009). The number of deep coalescences (*ndc*) is a count of the alleles that fail to coalesce within a given branch of the species tree; its values are always positive integers and are positively correlated with the number of tips of the gene trees. The genealogical sorting index (*gsi*) quantifies the degree of exclusive ancestry among of a set of tips in a rooted tree; it is normalized to a unity range between 0 and 1 and calculated as the minimum number of nodes needed to unite a group of given size divided by the number of nodes actually uniting the group. The *gsi* reaches its maximum when the observed group is monophyletic and its minimum when all nodes on a tree are required to unite a group. The coalescent likelihood measures the product of the branch-specific probability densities of a gene tree given the species tree across all branches of the species tree (Rannala & Yang 2003). Several R packages can calculate the coalescent likelihood or specific aspects thereof (e.g. Liu & Yu 2010; Paradis 2013; Reid *et al.* 2014; Volz *et al.* 2014). The package by Liu & Yu (2010) implements the joint probability distribution of a gene tree given a species tree, along with the coalescent times, in a formal calculation of the coalescent likelihood. The package by Reid *et al.* (2014) estimates the likelihood of the coalescent waiting times of a gene tree, while the package by Volz *et al.* (2014) can calculate the log-likelihood of a gene tree given a demographic history under different population demographic processes. Although these implementations result in slightly different gene tree likelihoods, they are all derivations on the probability density function of gene trees (Rannala & Yang 2003). P2C2M includes two of these calculations as summary statistics: the probability distribution of a gene tree given a species tree as calculated with the R package PHYBASE v.1.3.1 (Liu & Yu 2010) and the likelihood of the coalescent waiting times of a gene tree given a species tree as implemented by Reid *et al.* (2014). The implementation by Liu & Yu (2010) conducts computations following equations 1–9 of Rannala & Yang (2003) and is abbreviated with '*coal*' to reflect its correspondence to the coalescent likelihood. The implementation by Reid *et al.* (2014) conducts computations following equation 1–7 of Rannala & Yang (2003) and is abbreviated with '*lcwt*' (*l*ikelihood of the *c*oalescent *w*aiting *t*imes). The results from these two probability-based statistics are strongly correlated, and both are also correlated with the full species coalescent calculated by *BEAST (Fig. S1, Supporting information). The inclusion of similar probabilistic summary statistics allows researchers to compare their utility in different applications (e.g. Nakhleh 2013) and is here also necessitated by policy of the R package repository CRAN (which requires that default functions must be free of non-CRAN dependencies, such as package 'PHYBASE').

P2C2M measures the amount of discrepancy between the posterior and the posterior predictive distributions by computing the difference between the summary statistics calculated from each (Fig. 1b). To identify data with substantial deviations from the expectation of no difference between these distributions, we follow Gelman *et al.* (2009) and use quantiles conditioned on the distribution of differences as test statistics. Data that are a poor fit to the MSCM are recognized when deviation from the expectation of a difference distribution that is centred on zero is encountered above a specified quantile level. For practical purposes, these quantiles serve a similar role as alpha-values in a parametric bootstrap.

Evaluation of P2C2M via simulation testing

To assess the performance of P2C2M under different allele, gene and species numbers and to evaluate the statistical behaviour of the four summary statistics implemented in P2C2M, a series of simulations were performed. The simulated data sets differed in four important aspects: the number of species, the number of alleles per species, the number of loci per data set and the level of the DNA substitution rate. To facilitate the interpretation of our results, we grouped the analyses into two conceptual sets: one that simulated a low number of species and a high number of alleles per species similar to many *phylogeographic* investigations, and one with a higher number of species but with fewer alleles per species. The later simulation is similar to a low-level *phylogenetic* investigation, and we use it to evaluate how P2C2M performs when using different numbers of loci.

For the phylogeographic set of simulations, a total of 80 data sets were simulated. Each of these data sets was designed to comprise three species, each represented by approximately 15 alleles, and a total of 10 loci. Specifically, alleles from three populations ($n_A = 12$, $n_B = 14$, $n_C = 20$) and one outgroup ($n_O = 1$) were simulated, genetic diversity ($\theta = 4\mathrm{Ne}\mu$) was treated as constant throughout the species tree, and the topology of the species tree was constrained to (*O*(*A*(*B*,*C*))). The divergence between populations *B* and *C* was drawn from a uniform distribution bounded by 1N and 3N generations, divergence between populations *A* and *BC* was set to twice the divergence between *B* and *C*, and divergence between the ingroup and outgroup was assumed to be 5 times the divergence between *B* and *C*. The simulated species tree topology was thus shallow, but does not occupy the zone of anomalous gene trees described by Degnan & Rosenberg (2006). Coalescent genealogies were simulated using MS, with nucleotide sequence data simulated on the genealogies using SEQ-GEN v.1.3.2 (Rambaut & Grassly 1997) under conditions similar to those observed in empirical data sets: an HKY model of sequence evolution ($f_A = 0.3$, $f_C = 0.2$, $f_G = 0.3$, $Ti/Tv = 3.0$), a sequence length of 658 bp and two different substitution rates ($s_1 = 0.06$, $s_2 = 0.02$). Half of the phylogeographic data sets were simulated under the MSCM, while the other half were simulated with continuous gene flow between lineages *A* and *C* that occurred after the split between *B* and *C*, drawing the proportion of each lineage with migrant ancestry from a uniform distribution between 0.01 and 0.5.

For the phylogenetic simulations, a total of 100 data sets were simulated under a Yule model of tree evolution (Yule 1924) using DENDROPY v.3.12.0 (Sukumaran & Holder 2010). These simulations were conducted under the MSCM and contained 10 species, each represented by six alleles, with a maximum population divergence of 20N generations. The number of loci per data set varied between five, 10, 15 and 20 loci, and we used the lower DNA substitution rate ($s_2 = 0.02$). Other settings matched the phylogeographic simulations. The topologies of the species trees are provided as Fig. S2a–d (Supporting information).

Simulation replication

Coalescent stochasticity is one source of variability that must be accounted for in species tree inference (Rosenberg & Nordborg 2002). In the context of P2C2M, coalescent stochasticity in the posterior gene tree distributions is particularly visible in the form of varying branch lengths (Fig. S3, Supporting information). In preliminary analyses, we noticed occasional extreme outliers in some posterior predictive tree distributions and speculated that these values may lead to false-negative results, particularly in comparisons involving data that does not fit the MSCM. To explore this issue, we included an optional replication strategy in P2C2M that, if selected by the user, acts to decrease the variance of the test distributions. The strategy is designed to improve the detection of violations of the MSCM when using finite samples from the posterior distribution of gene trees. To evaluate the impact of simulation replication on the identification of poor model fit with P2C2M, a comparison between 0, 10 and 100 replicates per draw from the posterior distribution was conducted. Based on the results of this evaluation, all analyses with P2C2M in this investigation were executed with a setting of 100 simulation replicates, unless noted otherwise.

Demonstration of P2C2M with empirical data

The performance of P2C2M is illustrated using two empirical data sets: the first consists of 22 species (five loci) of chipmunks in the genus *Tamias* (Reid *et al.* 2012; also used by Reid *et al.* 2014). Previous research suggests that introgressive hybridization exists in *Tamias* chipmunks (Good *et al.* 2003, 2008), and a series of coalescent simulations support this interpretation (Reid *et al.* 2012). The second consists of data from four subspecies of *Myotis lucifugus* that may infrequently exchange alleles (Carstens & Dewey 2010). Between four and 10 individuals per subspecies, sampled across seven loci were collected for the second data set. The empirical data represent two types of gene flow: introgression between clearly separated species (*Tamias*) and possible gene flow among groups within a nominal species (*Myotis*).

Initiating a run of P2C2M

Complete data analysis in P2C2M is conducted using a single command. A user must provide a directory with three different types of input files: a species tree file, a gene tree file for each gene under study and an XML file generated by BEAUTi (i.e. the input script of BEAST; Drummond *et al.* 2012). To start the P2C2M run, the user then enters the command 'p2c2m.complete', followed by a minimum of two input parameters in parentheses: the name of the input directory and the name of the BEAUTi-generated file in that directory. Additional parameters such as the desired number of simulation replicates, the instruction to use multiple CPUs to speed up calculation times, or a specific *BEAST version number to ensure correct data parsing, can optionally be supplied, with the default values derived from our preliminary analyses. The results of a P2C2M run comprise test statistics, measures of data dispersion and deviations marked at several quantile levels for each gene under study and of the sum of all genes. Details on file formats and the name requirements of input files, a list of optional parameters and instructions on result visualization are available in the package manual of P2C2M. Our software was tested on UBUNTU v.12.04, ARCHLINUX v.3.16 and MAC OSX 10.8.5 in both single- and multi-processor environments. Supplemental to P2C2M is a software script that automates the generation of the XML input files of BEAST and *BEAST. This script, named 'BEAUTiAutomator.py', was written in PYTHON v.2.7 (Python Software Foundation 2012) and is provided so that users of P2C2M can evaluate the software under conditions specific to a given empirical system (e.g. specific number of lineages or alleles/lineage).

Results and discussion

Markov chain convergence

Markov chains that have not reached convergence can lead to false-positive results in PPS because the posterior distribution of gene trees may not contain estimates that have a high probability given the data. We found that for the data sets analysed here, approximately 50 million generations were required during species tree inference for chain convergence; shorter chains generally resulted in a far greater error rate (Table S1, Supporting information), which we attribute to poor sampling of the posterior distribution of gene and species trees.

Error rates under different summary statistics

Using the four summary statistics included in P2C2M, different levels of poor model fit were detected in both types of simulations (i.e. phylogeographic and phylogenetically inspired) and across several probability quantile thresholds (0.1, 0.05, 0.01, 0.001). Unless noted otherwise, the results described hereafter are based on a probability threshold of 0.01; we observed that this value was most appropriate for detecting violations of the MSCM while avoiding false-negative results. For phylogeographic data that fit the MSCM (Fig. 2a), false positives were not detected using either probabilistic summary statistic, in only one of 400 loci under the *ndc*, but in multiple data sets (s_1: 19/200 loci; s_2: 3/200 loci) under the *gsi*. When the same simulations were evaluated using the sum of test distributions from each locus, false positives were absent under the *lcwt* and the *ndc*, occurred in four of 40 sums under *coal*, and were present in several data sets (s_1: 9/20 sums; s_2: 3/20 sums) under the *gsi*. For data simulated in the presence of gene flow, where P2C2M should identify violations of the MSCM, the two

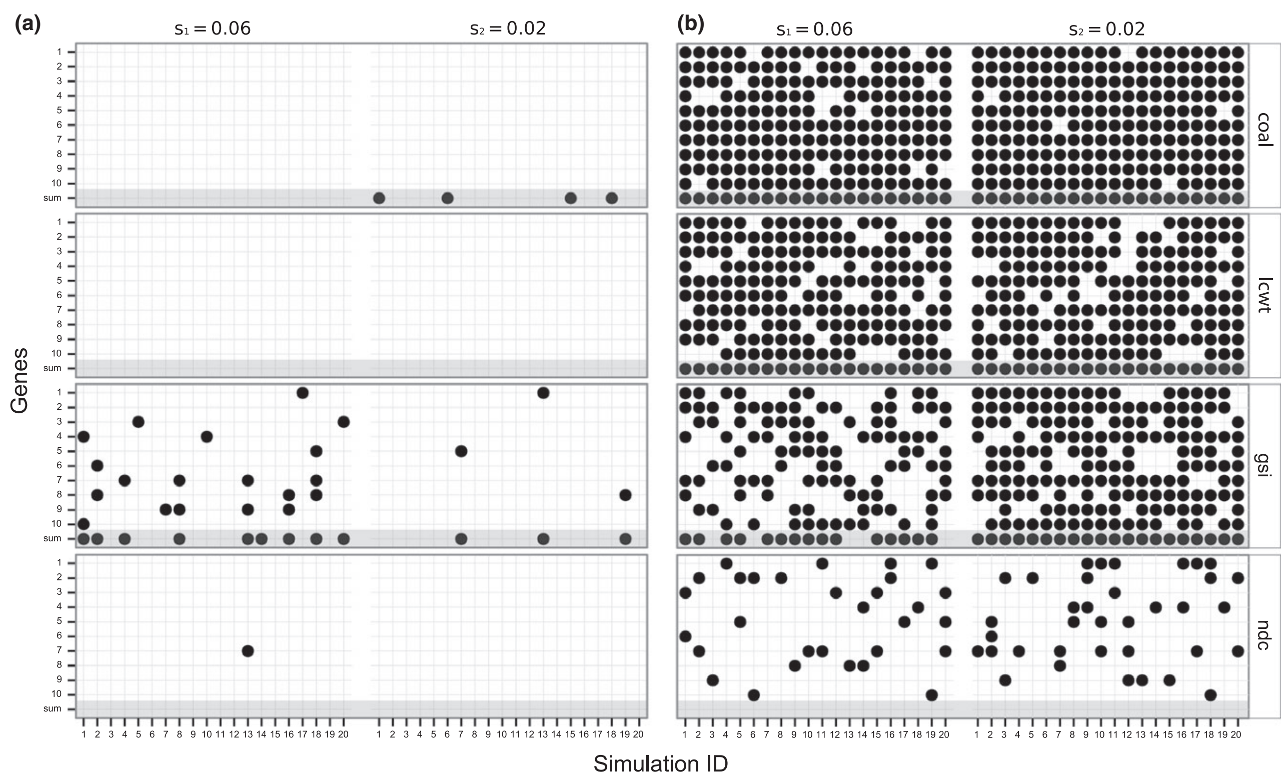

Fig. 2 Summary of cases of poor model fit among the data sets simulated to be representative of phylogeographic studies. Poor model fit was identified at a probability threshold of 0.01. Results for each gene individually, for the sum of all genes, for two different substitution rates and for each summary statistic under study are displayed. Cases of poor model fit identified on the sum of all genes are highlighted in grey. (a) Data sets simulated under the MSCM; these data are expected to fit the MSCM. (b) Data sets simulated in the presence of migration between species; these data are not expected to fit the MCSM.

probabilistic summary statistics also displayed a high success rate (Fig. 2b): *Coal* correctly identified a high rate of poor model fit across individual loci, failing in only a small proportion (rate of false negatives for s_1: 0.04; s_2: 0.025). Similarly, the *lcwt* was generally effective at detecting violations of model fit, but had a slightly higher failure rate (s_1: 0.165; s_2: 0.14). In contrast, the *gsi* exhibited a substantially higher rate of false-negative results (s_1: 0.485; s_2: 0.205), while the *ndc* was unable to identify even half of all cases of poor model fit under either substitution rate. When these simulations were evaluated using the sum of test distributions from each locus, false negatives were absent under both probabilistic summary statistics, but occurred in four of 40 sums in the *gsi* and all simulations using the *ndc*. Similar error rate differences between summary statistics were also identified under a probability threshold of 0.05 (Fig. S4, Supporting information).

The high rate of false-negative results detected under the *ndc* among the phylogeographic simulations is probably a consequence of the low number of species or the relatively short internal branches of the species trees: every species tree comprises only four species under these simulations, thus deep coalescent events can only occur on three nodes. It remains unclear what other characteristics of the data may have led to these false-negative results, as the topology of the species tree is generally being estimated accurately. The two probabilistic summary statistics, by contrast, exhibited the least number of false-positive and false-negative results under all simulated data sets under study, which correlates with our results on their comparability with the speciation coalescent inferred by *BEAST (Fig. S1, Supporting information). Moreover, summary statistic *coal* displayed the least amount of variance among the test distributions when measured by the mean coefficient of variation averaged across all genes under study (Fig. S5, Supporting information).

Error rates at different data set sizes

For data simulated under the MSCM and designed to resemble low-level phylogenetic investigations, the proportion of false positives at individual loci was positively correlated with the number of loci (Fig. 3). However, differences in the relative proportion of false positives were detected depending on the summary statistic used. While the rates of false positives were low under the two probabilistic summary statistics, we observed an even lower rate when using the *ndc*. The error rates of *coal* and *lcwt* were similar at the individual gene level, but when the same simulations were evaluated upon the summation of gene-wise test distributions, the *lcwt* displayed a much lower error rate than *coal*. The *gsi* displayed high rates of false positives at individual loci and when summed across loci.

We suggest that users of P2C2M consider both probabilistic summary statistics as well as the *ndc* when interpreting the results of their model checking analysis. The probabilistic statistics perform well, particularly at deeper levels of divergence and when the ratio of lineages to alleles sampled is high (and thus there are fewer opportunities for incomplete lineage sorting to be observed). For phylogeographic investigations, where a greater number of alleles/lineage are typically sampled, the *ndc* performs slightly better than the probabilistic statistics, as it is slightly less prone to false-positive results, but suffers from reduced statistical power when detecting false negatives. While more research is needed to completely understand why this is the case, the *ndc* has proven to be very useful for phylogeographic research since its introduction by Maddison (1997). The slight increase in computation time required to calculate *ndc* in addition to the probability-based summary statistics is trivial compared to the added benefit. Finally, in the majority of our simulations, results from *ndc* and the probabilistic statistics are consistent, suggesting that P2C2M is an effective addition to the analytical toolbox available to phylogeography and molecular systematics.

Summary of simulation tests

Upon comparison of the four summary statistics implemented in P2C2M, we selected the two probabilistic summary statistics to interpret the results of the simulation tests. Using these statistics, both of which are based on the calculation of the coalescent likelihood, tests of good model fit exhibit only few instances of false-positive results, particularly on data sets with higher numbers of loci. Given that these data had been simulated under species trees exhibiting very shallow levels of divergence, the substitution rate selected during data simulation had only a limited effect on the rate of false-positive results, suggesting that P2C2M will be effective across a wide range of empirical data sets. A lower substitution rate was generally found to coincide with a slight decrease in the rate of false-positive as well as the rate of false-negative results in both data sets.

Simulation replication

A comparison of the test distribution quantiles generated under different replication levels indicated that simulation replication and subsequent averaging of summary statistic values were useful for the correct identification of poor fit to the MSCM. When data fit the MSCM, simulation replication had little or no effect on the test distributions (Fig. 4a). However, for the data that were

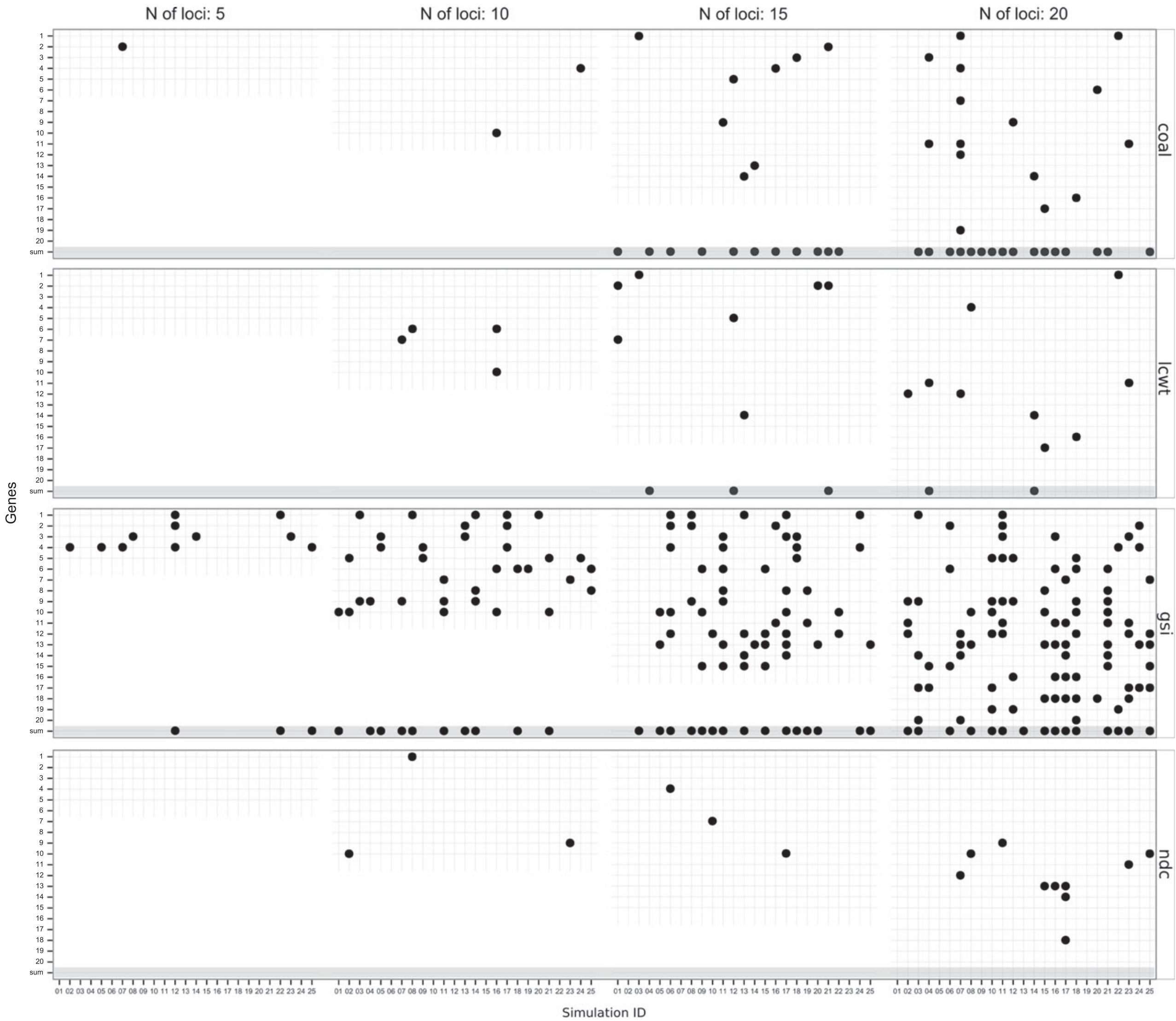

Fig. 3 Summary of cases of poor model fit among the data sets that were simulated to be representative of phylogenetic studies. Results for each gene individually, for the sum of all genes, for different numbers of loci per data set and for each summary statistic under study are displayed. All data sets were simulated under the MSCM using substitution rate s_2; other settings are identical to Fig. 2.

simulated in violation to the MSCM (i.e. by including gene flow), moderate levels of simulation replication were found to improve the rate of correct identification of poor model fit (Fig. 4b). Increasing the number of replicates from 0 to 10 caused the strongest relative reduction of test distribution variance, but some improvement to the variance and, in particular, the modality of the test distributions was also detected at higher levels of replication. Simulation replication was particularly effective in reducing error rates in evaluations where low probability thresholds (i.e. 0.01 and below) were applied, because here even a small number of outliers within the posterior predictive tree distribution can alter the significance assignment. We encourage users of P2C2M to explore this issue using simulation testing under conditions that reflect their empirical data.

Computation times

Several code improvements pertaining to the faster calculation of summary statistics were made in P2C2M compared to the original software scripts of Reid *et al.* (2014). In particular, the implementation of parallel process invocation via the message-passing interface OpenMPI (Gabriel *et al*. 2004) helped to offset the increase in computation time caused by simulation replication. We found that if the number of genes, taxa and MCMC generations is held constant, the number of simulation

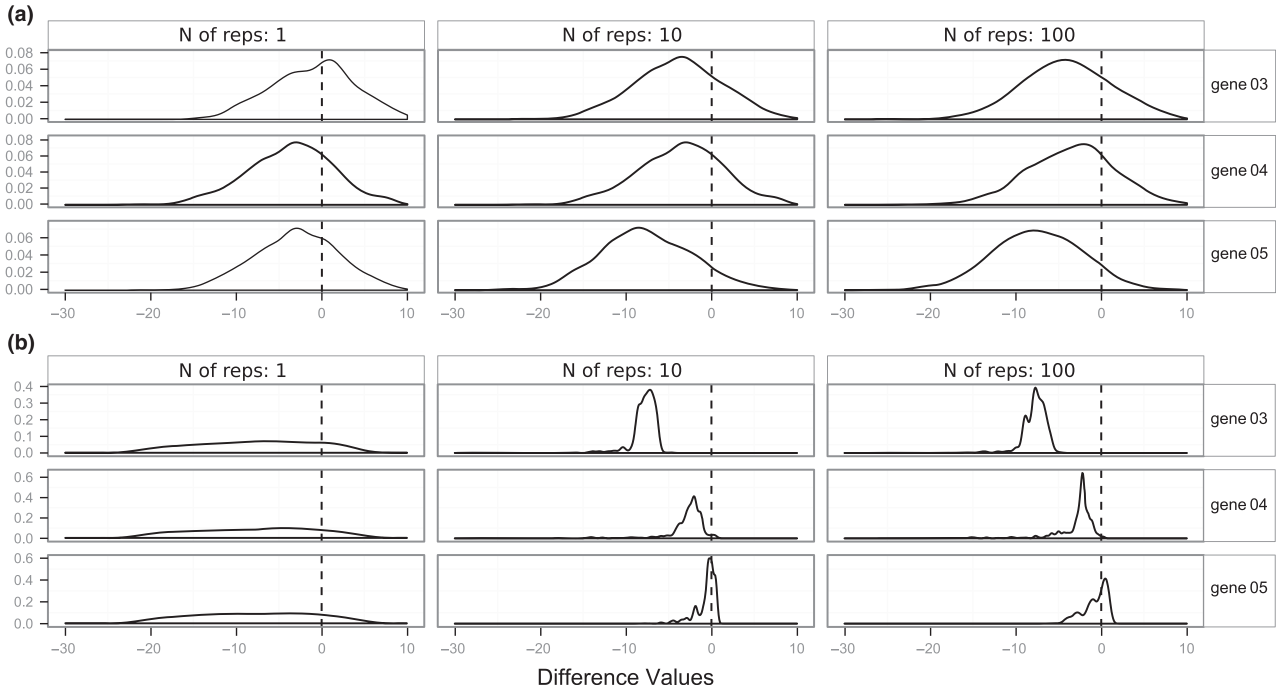

Fig. 4 The effect of simulation replication on the test distributions. The plots display the test distributions (i.e. the distributions of differences between the posterior and the posterior predictive distributions) under summary statistic *coal*. The expected centre of a distribution with good model fit is indicated by the dashed vertical line. Only, the results for the substitution rate s_1 and for genes 3, 4 and 5 of data set 11 of the phylogeographically representative simulations are displayed. (a) Data simulated under the MSCM; (b) data simulated with migration between species.

replicates increases computation times approximately linearly. Parallel process invocation hereby displayed a lower rate of increase than calculations without it (Fig. 5). Utilizing multiple processors came at the price of an initial slowdown, however, which constitutes the time that OpenMPI requires to initialize and coordinate the individual processes: for data sets with 10 or fewer loci and a setting of less than five simulation replicates, the time required to initialize multiple processes was found to outweigh the subsequent speed gains. Apart from the base time required for the actual simulation of genealogies, computation times were positively correlated with the size of the analysed data sets and the number of summary statistics selected. Under parallel process invocation and a setting of 100 simulation replicates, the simulated phylogeographic data sets required on average 6.2 h of computation time when calculating all summary statistics (Fig. 5). The simulated phylogenetic data sets required an average of 21.7 h when using 20 loci per data set, with proportionally less time for fewer loci. Given the diminishing returns in reducing the variance of the test distributions with higher numbers of simulation replicates on the one hand, and the associated increase in computation time on the other, 10–50 simulation replicates with concurrent parallel process

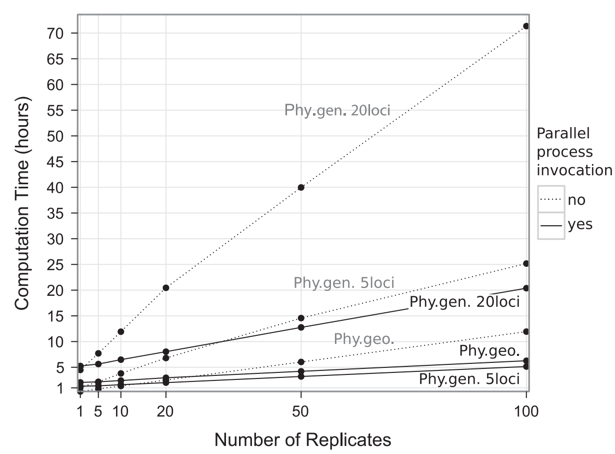

Fig. 5 Average computation times of P2C2M by data set size, number of replicates and parallel process invocation. Computations performed under parallel process invocation are indicated by solid, such without by dotted lines. Type and size of the data sets evaluated are indicated by the graph annotations. All values constitute the averages of three separate data sets with identical dimensions. Computation times were measured under the inference of all four summary statistics on a machine with an i7-4770S 3.1 GHz Intel Quad-Core processor, eight GB of RAM and the operating system UBUNTU 14.04.

invocation are recommended for an accurate and efficient analysis.

Demonstration with empirical data

As a consequence of the enhancements described above, P2C2M offers improved statistical power compared to the original scripts of Reid *et al*. (2014). This improvement is particularly evident on empirical data sets where gene flow is suspected. For example, previous work suggested that introgressive hybridization has occurred in western *Tamias* chipmunks, and we found (as did Reid *et al.* 2014) that the MSCM is not a good fit to this system (Table 1). The original software scripts by Reid *et al.* (2014) inferred poor fit for one of four nuclear as well as one mitochondrial gene under study. Analysis via P2C2M, by contrast, identified poor fit to the MSCM in at least three of the four nuclear genes as well as the mitochondrial marker under all summary statistics (Table 1).

P2C2M also suggested a poor fit to the MSCM by the data collected from *Myotis lucifugus*. Previous work used these data to delimit four *M. lucifugus* subspecies as independent, but cautioned that there was also evidence for divergence with gene flow (Carstens & Dewey 2010).

Table 1 Results of the empirical data sets analysed under P2C2M in comparison with the results generated by Reid *et al.* (2014). All loci under study are of nuclear origin, except where indicated

Genes	*coal*	*lcwt*	*gsi*	*ndc*
***Tamias* – Reid *et al.* (2014)**				
anon	n.a.		n.a.	
acr	n.a.		n.a.	*
zan	n.a.		n.a.	
zp2	n.a.		n.a.	
cyt b (mtDNA)	n.a.	*	n.a.	*
Sum of all genes	n.a.		n.a.	
***Tamias* – this study**				
anon	*	*	*	*
acr	*	*	*	*
zan	*	*	*	*
zp2	*		*	*
cyt b (mtDNA)	*	*	*	*
Sum of all genes	*		*	*
***Myotis* – this study**				
681a	*		*	*
681b	*			*
685a	*	*		
734z	*	*		*
735b	*	*		*
735f	*	*		*
cyt b (mtDNA)	*	*	*	*
Sum of all genes	*			

Asterisks indicate cases of poor model fit at a probability level of 0.01. n.a., not applicable.

These results offer a clear illustration of the importance of model checking; the data of *M. lucifugus* are not a good fit to a coalescent model that does not include gene flow, and thus, the delimitation analyses implemented under such a model are probably not appropriate.

Implications on species tree inference and other evolutionary analyses

While model checking should be an essential part of Bayesian phylogenetic inference (Goldman 1993), it is important to interpret the results of a P2C2M analysis in the context of other analyses and general information from particular systems. For example, we applied P2C2M to two empirical systems (*Tamias*, *Myotis*) where previous information (including MSAT data from hybrid zone transects; Good *et al.* 2008) led to the suspicion that the MSCM was violated due to gene flow. In the case of *Tamias*, the poor fit of the MSCM detected by P2C2M leads us to be sceptical of the branch lengths (and possibly the topology) of the species tree estimate, as theoretical work indicates that gene flow in various forms can decrease the accuracy of species tree estimates (e.g. Eckert & Carstens 2008; Leaché *et al.* 2014). We do not plan on replacing the species tree estimate with a tree generated using other methods; rather, we consider any phylogenetic tree to be an inadequate summary of the evolutionary history in this group given our results. Similarly, Carstens & Dewey (2010) suspected that gene flow had occurred in *Myotis;* our results corroborate that the MSCM is not an adequate summary for the demographic history in this system.

Evolutionary analysis is the process by which researchers identify the set of processes (e.g. divergence, gene flow, selection, population size change) that have left the most substantial imprint on the genetic diversity of the focal species. Model checking, along with model selection (e.g. Carstens *et al.* 2013), thus offers well-justified statistical methods for evaluating a variety of models. Once this evaluation is complete, inferences can be made from the results that are most appropriate to a particular system. Identifying a case where a given data set has a poor fit to the MSCM does not mean that researchers should discard the resulting species tree estimate entirely, or replace it with an estimate of phylogeny made using a different analytical approach. Rather, such a results should prompt researchers to conduct additional analyses that do not assume that all shared polymorphism results from incomplete lineage sorting.

Conclusions

Model checking is an essential component of Bayesian statistical inference (Morey *et al.* 2012), and PPS is an ideal tool for this purpose (Gelman & Shalizi 2012). By comparing statistics inferred from empirical data to statistics simulated under the very model used to estimate the empirical parameters, PPS extends data analysis beyond the estimation of parameters (Gelman 2003). It allows researchers to assess model adequacy and to learn how and why a model does not fit the data (Gelman & Shalizi 2012). In the case of the multispecies coalescent, researchers can discover whether patterns inherent to the data (the observed genealogies) are inconsistent with model assumptions (e.g. that shared polymorphism results from incomplete lineage sorting).

Several lines of argument can be offered for incorporating model checking into a species tree analysis. First, most current implementations of the MSCM do not account for gene flow, a process that is likely more common that previously suspected (Nosil 2008; Pinho & Hey 2010). Unless gene flow is accounted for, species divergence times may be considerably underestimated (Mallet 2005). Second, a potentially large number of factors can cause poor fit to the MSCM. In addition to evolutionary processes that lead to polymorphism being shared across lineages (e.g. hybridization, recombination or horizontal gene transfer), poor model fit may also be caused by unmodelled population structure and inaccurately estimated gene trees (Reid *et al.* 2014). Third, different computational implementations of the MSCM model exist (Huang *et al.* 2010), and it ultimately remains the responsibility of the user to evaluate whether their data fit a specific model implementation and analysis strategy (Knowles *et al.* 2012). All these factors indicate the importance of measuring model fit prior to the application of the MSCM during species tree inference. P2C2M enables users to easily and effectively evaluate whether a MSCM is appropriate for their data.

Acknowledgements

We would like to thank Jeremy Brown, Jordan Satler and members of the Carstens laboratory for conversations related to this work, as well as Mike Sovic, Paul Blischak and Teofil Nakov for help with testing P2C2M. We also thank Maxim Kim for assistance with the R package *RMPI* and Liang Liu for his help in maintaining compatibility between P2C2M and the R package *PHYBASE*. This work was supported in part by an allocation of computer resources from the Ohio Supercomputer Center.

References

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, **24**, 412–426.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences B*, **98**, 4563–4568.

Bollback JP (2002) Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, **19**, 1171–1180.

Bouckaert RR (2010) DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, **26**, 1372–1373.

Brown JM (2014) Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology*, **63**, 334–348.

Carstens B, Dewey T (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. *Systematic Biology*, **59**, 400–414.

Carstens BC, Brennan RS, Chua V *et al.* (2013) Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Molecular Ecology*, **22**, 4014–4028.

Chung Y, Ané C (2011) Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. *Systematic Biology*, **60**, 261–275.

Cummings MP, Neel MC, Shaw KL (2008) A genealogical approach to quantifying lineage divergence. *Evolution*, **62**, 2411–2422.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, e68.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Eckert AJ, Carstens BC (2008) Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics and Evolution*, **49**, 832–842.

Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Gabriel E, Fagg GE, Bosilca G *et al.* (2004) OpenMPI: goals, concept, and design of a next generation MPI implementation. In: *Recent Advances in Parallel Virtual Machine and Message Passing Interface* (eds Kranzlmueller D, Kacsuk P, Dongarra J), pp. 97–104. Hungary, Budapest.

Gelman A (2003) A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, **71**, 369–382.

Gelman A, Shalizi CR (2012) Philosophy and practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, **66**, 8–38.

Gelman A, Carlin JB, Stern HS, Rubin DB (2009) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/ CRC Texts in Statistical Science, Boca Raton, FL, USA.

Goldman N (1993) Statistic tests of models of DNA substitution. *Journal of Molecular Evolution*, **36**, 182–198.

Good JM, Demboski JR, Nagorsen DW, Sullivan J (2003) Phylogeography and introgressive hybridization: Chipmunks (Genus *Tamias*) in the Northern Rocky Mountains. *Evolution*, **57**, 1900–1916.

Good JM, Hird S, Reid N *et al.* (2008) Ancient hybridization and mitochondrial capture between two species of chipmunks. *Molecular Ecology*, **17**, 1313–1327.

Heled J, Drummond A (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, **8**, 289.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

Hey J (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, **27**, 905–920.

Huang H, He Q, Kubatko L, Knowles L (2010) Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, **59**, 573–583.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Hudson RR, Turelli M (2003) Stochasticity overrules that 'Three-times Rule': genetic drift, genetic draft, and coalescence times for nuclear loci vs. mitochondrial DNA. *Evolution*, **57**, 182–190.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Joly S (2012) JML: testing hybridization from species trees. *Molecular Ecology Resources*, **12**, 179–184.
Kingman JFC (1982) On the genealogy of large populations. *Journal of Applied Probability*, **19**, 27–43.
Knowles LL (2001) Did the Pleistocene glaciation promote divergence? Tests of explicit refugial models in montane grasshoppers. *Molecular Ecology*, **10**, 691–701.
Knowles LL, Lanier HC, Klimov BP, He Q (2012) Full modeling vs. summarizing gene-tree uncertainty: method choice and species-tree accuracy. *Molecular Phylogenetics and Evolution*, **65**, 501–509.
Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, **58**, 478–488.
Kubatko L, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Journal of Bioinformatics*, **25**, 971–973.
Leaché AD, Harris RB, Rannala B, Yang Z (2014) The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, **63**, 17–30.
Lewis PO, Xie W, Chen M-H, Fan Y, Kou L (2014) Posterior predictive Bayesian model selection. *Systematic Biology*, **63**, 309–321.
Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, **56**, 504–514.
Liu L, Yu L (2010) Phybase: an R package for species tree analysis. *Bioinformatics*, **26**, 962–963.
Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, **53**, 320–328.
Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, **46**, 523–536.
Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55**, 21–30.
Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, **20**, 229–237.
Mallet J (2007) Hybrid speciation. *Nature*, **446**, 279–283.
Morey RD, Romeijn J-W, Rouder JN (2012) The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, **66**, 1–8.
Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology and Evolution*, **28**, 719–728.
Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103–2106.
Paradis E (2013) coalescentMCMC: MCMC Algorithms for the Coalescent. R package, version 0.4. Available at: http://CRAN.R-project.org/package=coalescentMCMC
Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.
Pinho C, Hey J (2010) Divergence with gene flow: models and data. *Annual Reviews of Ecology, Evolution and Systematics*, **41**, 215–230.
Python Software Foundation (2012) Python Language Reference, version 2.7. Available at: http://www.python.org
R Development Core Team (2013) R: a language and environment for statistical computing. (R Foundation for Statistical Computing). Vienna, Austria. Available at: http://www.r-project.org
Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computational Applications to the Biosciences*, **13**, 235–238.
Rambaut A, Suchard MA, Xie D, Drummond AJ (2014) Tracer v1.6. Available at: http://beast.bio.ed.ac.uk/Tracer
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.
Reid N, Demboski J, Sullivan J (2012) Phylogeny estimation of the radiation of western North American chipmunks (*Tamias*) in the face of introgression using reproductive protein genes. *Systematic Biology*, **61**, 44–62.
Reid NM, Brown JM, Satler JD *et al*. (2014) Poor fit to the multi-species coalescent model is widely detectable in empirical data. *Systematic Biology*, **63**, 322–333.
Rosenberg N, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380–390.
Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
Sullivan J, Arellano E, Rogers DS (2000) Comparative phylogeography of Mesoamerican Highland Rodents: concerted vs. independent response to past climate fluctuations. *American Naturalist*, **155**, 755–768.
Than C, Nakhleh L (2009) Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, **5**, e1000501.
Volz EM, Ratmann O, Severson ER (2014) rcolgem: statistical inference and modeling of genealogies generated by epidemic and ecological processes. R package, version 0.0.4. Available at: http://R-Forge.R-project.org/projects/colgem/
Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis F.R.S. *Philosophical Transactions of the Royal Society of London B*, **213**, 21–87.

N.M.R. wrote the original R scripts. B.C.C. and M.G. conceived the study design and generated the simulation data. M.G. wrote the code of P2C2M, the PYTHON script BEAUTiAutomator.py and the manuals and package vignettes. M.G. conducted the analyses of this investigation and produced all figures, tables and supporting files (except for Table S1, Supporting information). G.L.W. produced Table S1 (Supporting information). G.L.W., M.G. and N.M.R. conducted testing of P2C2M. M.G. and B.C.C. prepared and edited the manuscript, with input from all authors. Funding was provided via a start-up package from The Ohio State University to B.C.C.

Data accessibility

The package P2C2M, including a user manual, a package vignette with instructions for result visualization and a set of example files, is available from CRAN at http://cran.r-project.org/web/packages/P2C2M/. All simulated data sets are publicly available from Dryad under doi:10.5061/dryad.n715n. BEAUTiAutomator is available with documentation and example files from http://github.com/michaelgruenstaeudl/BEAUTiAutomator/.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Pairwise comparisons between summary statistic *coal*, summary statistic *lcwt*, and the full species coalescent likelihood as inferred by *BEAST via scatterplots, correlation coefficients, and goodness-of-fit tests to standard linear models.

Figure S2 a–d Overview of the species tree topologies that are used to represent contemporary phylogenetic studies.

Figure S3 Visual comparison of a posterior gene tree against 1000 corresponding posterior predictive gene trees using DENSITREE (Bouckaert 2010).

Figure S4 Summary of the cases of poor model fit among the simulated data sets representative of contemporary phylogeographic studies.

Figure S5 Comparison of the mean values of the coefficients of variation (CVs) across all four summary statistics.

Table S1 Evaluation of independent sampling under the Markov chain.