Software for Systematics and Evolution

Syst. Biol. 66(6):1045–1053, 2017

DOI:10.1093/sysbio/syx001
Advance Access publication February 16, 2017

PHRAPL: Phylogeographic Inference Using Approximate Likelihoods

NATHAN D. JACKSON[1], ARIADNA E. MORALES[2], BRYAN C. CARSTENS[2], AND BRIAN C. O'MEARA[1,*]

[1]*Department of Ecology and Evolutionary Biology, University of Tennessee, 442 Hesler Biology Building, Knoxville, TN 37996, USA and*
[2]*Department of Evolution, Ecology and Organismal Biology, Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210, USA*
**Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA; E-mail: bomeara@utk.edu.*

Received 18 May 2016; reviews returned 19 August 2016; accepted 4 January 2017
Associate Editor: David Posada

Abstract.—The demographic history of most species is complex, with multiple evolutionary processes combining to shape the observed patterns of genetic diversity. To infer this history, the discipline of phylogeography has (to date) used models that simplify the historical demography of the focal organism, for example by assuming or ignoring ongoing gene flow between populations or by requiring *a priori* specification of divergence history. Since no single model incorporates every possible evolutionary process, researchers rely on intuition to choose the models that they use to analyze their data. Here, we describe an approximate likelihood approach that reduces this reliance on intuition. PHRAPL allows users to calculate the probability of a large number of complex demographic histories given a set of gene trees, enabling them to identify the most likely underlying model and estimate parameters for a given system. Available model parameters include coalescence time among populations or species, gene flow, and population size. We describe the method and test its performance in model selection and parameter estimation using simulated data. We also compare model probabilities estimated using our approximate likelihood method to those obtained using standard analytical likelihood. The method performs well under a wide range of scenarios, although this is sometimes contingent on sampling many loci. In most scenarios, as long as there are enough loci and if divergence among populations is sufficiently deep, PHRAPL can return the true model in nearly all simulated replicates. Parameter estimates from the method are also generally accurate in most cases. PHRAPL is a valuable new method for phylogeographic model selection and will be particularly useful as a tool to more extensively explore demographic model space than is typically done or to estimate parameters for complex models that are not readily implemented using current methods. Estimating relevant parameters using the most appropriate demographic model can help to sharpen our understanding of the evolutionary processes giving rise to phylogeographic patterns. [AIC; grid search; isolation-with-migration; migration rate; multispecies coalescent; parameter optimization; population genetics; tree topologies.]

Phylogeographic research has benefited greatly from its embrace of quantitative models. By statistically fitting explicit evolutionary hypotheses to genetic data, researchers can be objectively guided by the empirical evidence to a better understanding of those processes (such as genetic drift and gene flow) that have given rise to observed patterns (Knowles and Maddison 2002). However, there remains one area in which subjectivity and intuition are still commonly relied upon: the selection of the analytical models that are used to estimate parameters that form the basis for phylogeographic inference. Researchers can choose among a wide range of software packages that implement powerful phylogeographic models (e.g., packages by Beerli and Felsenstein 2001; Heled and Drummond 2010; Hey 2010); however, due to their inherent complexity, most of these methods impose limits on the parameter space under consideration and thus are unable to jointly infer all the parameters that may be of interest (e.g., divergence time, gene flow, and the phylogeny).

Alternatively, researchers can formulate and test more complex phylogeographic hypotheses using parametric simulation (e.g., Knowles 2001) or approximate Bayesian computation (e.g., Fagundes et al. 2007), but they must still rely on their biological intuition or prior knowledge to specify the models to be tested. While this reliance on subjectivity is suitable for many questions, and can even be desirable inasmuch as it forces researchers to formulate their questions as discrete testable hypotheses, there are also many situations in which prior "knowledge" may be wrong, incomplete, or absent altogether (Templeton 2010). Quality of inference can rise no higher than the quality of the best model under consideration, and thus as long as there is reasonable uncertainty about the processes that underlie a particular dataset, there is room for a method that allows one to explore the model space in a more systematic way, to help guide one into the proverbial ballpark—or perhaps as a way to test whether one's prior assumptions are already there.

Here, we introduce a novel method (PHRAPL) that automates a framework for exploring model space by generating, testing, and comparing large sets of demographic models relatively quickly. PHRAPL calculates the approximate likelihood of each model, given the data, where models can include both divergence history (topology and branch lengths) and

migration among populations. Given the broad set of models that can be evaluated, this method shifts emphasis away from the estimation of parameters alone to optimization of models and parameters jointly. PHRAPL is specifically designed to serve a role in phylogeography that is analogous to that served by programs such as Modeltest (Posada and Crandall 1998), which select the best model of sequence evolution prior to estimating a phylogeny. When applied to empirical systems, PHRAPL can enable researchers to identify *which* parameters are important to estimate, and will thus be a useful tool for data exploration. PHRAPL works by computing the probability of observing gene tree topologies estimated from empirical data using a distribution of topologies simulated under various demographic models (O'Meara 2010). It then adopts an Akaike information criterion (AIC) framework (Burnham and Anderson 2002) to quantify the support for each model in the comparison set. Below we briefly describe the method and demonstrate that it provides a suitable framework for assessing and comparing the statistical fit of models commonly used in phylogeographic research.

How PHRAPL Works

Overview of the Phylogeographic Approximate Likelihood Method

PHRAPL inputs gene tree topologies (without branch lengths) estimated from empirical data and an association file that maps sampled individuals to user-defined populations. It operates under the standard assumptions of the coalescent model (i.e., no recombination within loci, free recombination among loci, etc.). Users first define models of demographic history, which can incorporate several types of parameters. In this paper, we focus on coalescence time (t) and migration rate (M); however, parameters that govern exponential population growth (g) and relative population size (n) can also be included. Once a model set is in hand, PHRAPL converts these models into commands for the program *ms* (Hudson 2002), which simulates gene trees under each defined model using parameter values that are optimized according to one of several strategies (see *Parameter Optimization* below). After the simulation step, PHRAPL approximates the log-likelihood of each model, given the data, by calculating the proportion of simulated topologies that match the observed topologies (O'Meara 2010). Specifically, the probability of observing a set of gene trees, G, given a particular set of demographic parameters is approximated by

$$\ln\mathbb{P}(G|t_i,M_i,g_i,n_i)\approx\sum_{k=1}^{n}\ln\left(\frac{m_k}{N}\right)$$

where m is the number of times that the kth observed tree topology occurs in a set of expected gene topologies simulated under the model, and N is the total number of trees simulated.

Strategies for Increased Efficiency

One critical challenge for PHRAPL is effectively searching the large set of possible gene trees. For just seven sampled alleles, there are 10,395 binary gene trees (Felsenstein 2004); this number increases to $>10^{21}$ trees when 20 alleles are sampled and would exceed the square of the number of atoms in the universe when more than 100 alleles are sampled. On the surface, the sheer number of possible gene tree topologies would seem to prevent PHRAPL from efficiently calculating probability based on the proportion of gene trees matching a given history. However, PHRAPL uses several strategies to circumvent this difficulty, and in combination these strategies enable PHRAPL to perform well for small to moderately sized datasets (e.g., between around 2–6 populations, with potentially hundreds of samples per population and hundreds of loci).

First, we note that although there may be millions of possible gene trees under a given demographic history, these do not occur at equal probability. For example, in cases of small populations that have been isolated for many generations, congruence between the gene tree and population tree occurs at a relatively high frequency (Hudson and Turelli 2003), so many gene trees will have the same inter-population branches, leaving only intra-population disagreements. This non-uniform distribution of gene tree probabilities (see Degnan and Salter 2005) enables a far more efficient PHRAPL inference.

Second, we implement subsampling of individuals within specified populations in order to decrease the size of tree space. Subsampling has been shown to be an effective strategy for estimating species trees from phylogeographic data (Hird et al. 2010), largely because the point of diminishing returns is quickly reached as samples from a population are added (e.g., Saunders et al. 1984). Even with fully resolved gene trees, having tens to hundreds of samples within a population may not provide much more information than having fewer samples because most coalescent events occur very recently in the tip populations (Nordborg 2001), and thus tell us little about the deeper demographic history of the populations. PHRAPL randomly samples alleles from all populations (with number of samples per population and number of replicates specified by the user) and then analyzes these subsampled gene trees against the simulated set of trees with the same number of tips. The mean likelihood across s subsample iterations is calculated such that

$$\ln\mathbb{P}(G|t_i,M_i,g_i,n_i)\approx\sum_{k=1}^{n}\ln\left(\frac{\overline{m}}{N}\right)_k.$$

Multiple sampling strategies are possible, although preliminary explorations have shown that a replicated subsampling of three or four alleles per population yields the best balance between adequate information content and computational efficiency (Supplementary Fig. S1).

Third, when comparing empirical and simulated gene trees, PHRAPL assumes that samples from within a

population are interchangeable, and then corrects for this assumption, greatly increasing the tree space over which likelihood-relevant information can be collected. For example, assume the observed gene tree is (((A1, A2), A3), (B1, C1)) and that the simulated gene tree is (((A3, A2), A1), (B1, C1)), where letters denote populations, and numbers denote the alleles sampled from that population. These two trees only fail to match due to the intra-population switching of tip labels in A. Instead of simply scoring this as a non-match, PHRAPL calculates a partial match score based on the total number of possible assignments of tip labels within populations in a given tree. To do this, PHRAPL first calculates a degeneracy weight for each subsampled tree that measures the extent to which permuting tree tip labels results in the same population tree topology. Likelihoods are then adjusted by multiplying the number of observed to expected tree matches by this weight, w, such that

$$\ln\mathbb{P}(G|t_i, M_i, g_i, n_i) \approx \sum_{k=1}^{n} \ln\left(\frac{\overline{mw}}{N}\right)_k.$$

Because there are six possible assignment permutations of labels A1, A2, and A3 in the above example, the simulated gene tree is scored as 1/3 of a match. Ignoring the arbitrary effect of intra-population labeling when quantifying matches results in a more efficient algorithm, particularly as the number of samples increases.

Fourth, given that the approximate likelihood is calculated by counting matches in a necessarily finite group of simulated trees, there is always some probability that none of the simulated genealogies will match a particular observed tree. This produces a point estimate of the likelihood equal to zero, or a log-likelihood (lnL) of negative infinity. In reality, we know that all gene topologies have a nonzero, albeit sometimes very small, probability, and having one or more negative infinities in the lnL prevents the calculation of the overall approximate lnL. Several corrections were explored to address this issue (see Supplementary Fig. S2). We found that the best way to obtain lnL values for trees yielding zero matches is to assign those trees the lowest possible probability plus a penalty constant, which is independent of the tree topology. We derived this penalty using our simulated data (Supplementary Fig. S2d) and have specified it to be that value that when added to the lnL, maximized the tightness of the linear relationship between approximate lnLs calculated from PHRAPL and actual lnLs calculated analytically using the program COAL (Degnan and Salter 2005).

Parameter Optimization

The simulation of gene tree distributions requires parameter values (e.g., coalescence times, migration rates) to be defined. PHRAPL can optimize parameters using one of two methods of nonlinear optimization—NLopt (Johnson 2008) and genoud (Mebane and Sekhon 2011), the latter of which incorporates an evolutionary search algorithm—or using a grid search, which exhaustively searches across all possible combinations of a set of specified parameter values. We have found that a grid search is both the most efficient and accurate method for parameter optimization using PHRAPL (see Supplementary Fig. S3 for a comparison of the three methods) and we have thus applied the grid approach when analyzing datasets in this study. Preliminary searches done using a coarse grid (i.e., with large increments between proposed values) may be used to construct finer grids for subsequent analyses. In addition, if the optimal value is found to be at the extreme edge of a grid, further searches should be conducted to include more extreme values. We hope to automate these grid extensions in a future release. Although the grid is composed of discrete values, parameter estimates from a given model are obtained by model averaging each value across the grid (see equation 4.1 in Burnham and Anderson 2002), and thus parameter estimates are continuous and not confined to taking on values included in the grid. Because the size of the grid exponentially expands with additional parameters, the coarseness of the grid must be balanced with desired model complexity, and it will be computationally difficult to search over a large number of parameters (e.g., >10) using either a grid or optimization.

Implementation

To run PHRAPL, one should first download the program, which is an R package currently available at https://github.com/bomeara/phrapl. We have included a tutorial in the Supplementary Materials, which provides detailed instructions for downloading, setting up, and running the program. Note that although the program is mostly written in R, it also uses Perl as a way to speed up the likelihood calculation, and thus, Perl and R must both be available prior to running PHRAPL. Once installed, there are six major steps to running a typical PHRAPL analysis, all of which are thoroughly discussed in the supplemental tutorial. First, one must input their data. This specifically consists of a set of gene trees in newick format and an assignment file that allocates tips of the trees to populations or species. Inputted trees should be rooted, either by including an outgroup sample (which is removed prior to analysis) or by midpoint rooting.

Second, one iteratively subsamples tips from each tree, which requires one to specify the number of tips to subsample per population as well as the number of subsample iterations to carry out. We recommend that at a minimum, the number of subsample iterations equal the sample size of a dataset's largest species divided by the number of individuals being subsampled. For example, if a dataset contains three species containing 23, 32, and 10 samples, respectively, and one is subsampling four tips per species, then one should at least use 32/4 = 8 subsample iterations. Because subsampling is done without replacement, increasing the number of

subsamples beyond this minimum can further reduce error in the likelihood estimate.

Third, one executes a function that calculates degeneracy weights for each subsampled tree, which allows PHRAPL to glean information from observed trees that only differ from simulated trees on an intra-population level.

Fourth, one must generate a set of models to test. This is accomplished using a PHRAPL function that generates all possible demographic models that fit a specified set of criteria. These criteria can include the number of species or populations to consider, the types of parameters to include (e.g., coalescence time, migration rate, etc.), and the maximum number of free parameters (K) to allow per parameter type or overall. In addition, *a priori* models can be created and added to a model set.

Fifth, once both the dataset and model set are ready, a PHRAPL search is conducted in which approximate likelihoods are calculated for each model in the model set, given the subsampled empirical data. This model search occurs jointly with a parameter search. When carrying out a grid search, one can either specify values or ranges of values to be considered for each parameter or use default values. The number of trees to be simulated under each model is specified using the argument 'nTrees' and should be set, at a minimum, at 10,000. Increasing nTrees can improve accuracy and precision of likelihood estimates.

Finally, after a set of models has been analyzed, PHRAPL can be used to rank models using AIC, to calculate relative probabilities of models using AIC weights, and to calculate model averaged parameter estimates.

Evaluation of PHRAPL Performance

We evaluated PHRAPL's performance using two approaches. First, we compared the approximate lnLs produced by PHRAPL to lnLs calculated analytically using the program COAL (Degnan and Salter 2005), under a variety of isolation-only histories (Fig. 1). For this, we fit both the generating species tree model as well as a false species tree model, in which labels for populations A and B were transposed (see Supplementary Methods). Second, we explored the accuracy with which approximate lnLs can be used to infer the true underlying model and parameter values generated under more complex histories (e.g., isolation-with-migration models; see Fig. 2), for which an analytical solution of the lnL is not available. Results of simulation testing using isolation-only models show that PHRAPL performs well at approximating the analytical lnL of these models. Linear correlations between lnLs calculated using PHRAPL and COAL varied across treatments, but were typically high ($R^2 > 0.8$; Fig. 3 and Supplementary Fig. S4). The lowest correlations were observed when the shortest tree lengths were combined with a small number of populations (in which case average branch lengths are at their shortest), particularly when topologies were balanced. Correlations near one were observed for the deepest trees (Fig. 3 and Supplementary Fig. S4). This pattern of poorer lnL estimation at shallower time depths likely results from increased incomplete lineage sorting, which renders the number of observed to expected tree matches less common and more stochastic. Slopes of the relationship between the two lnLs tended to be less than one at shallow tree depths and to approach one as branches lengthen. The intercept tended to decrease with increasing tree size, which should not undermine model selection given that PHRAPL only compares models for a single constant dataset. When comparing lnL values under the true model with those under a false species tree using either PHRAPL or COAL, lnLs were better for the true model in over 99% of cases. Typically, lnL distances between true and false models dwarfed lnL distances among replicate datasets generated using a particular model (see Supplementary Fig. S5), suggesting that both PHRAPL and COAL can reliably select the true model across treatments.

Although we currently lack an analytical solution for calculating lnLs for isolation-with-migration models, we can nevertheless address how closely approximate lnLs match the true lnL for these more complex models by

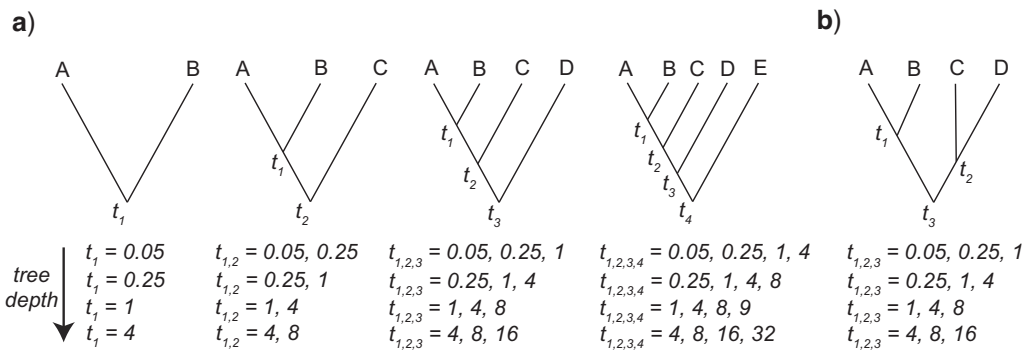

FIGURE 1. Isolation-only histories used for simulation testing. (**a**) Pectinate topologies with 2–5 populations and (**b**) a balanced topology with four populations were simulated. Each topology was simulated under four different tree depths (time parameters are given in units of $4N$) and under four different sample sizes per population (2–5). For each treatment, we simulated 30 replicate datasets containing trees for 30 loci.

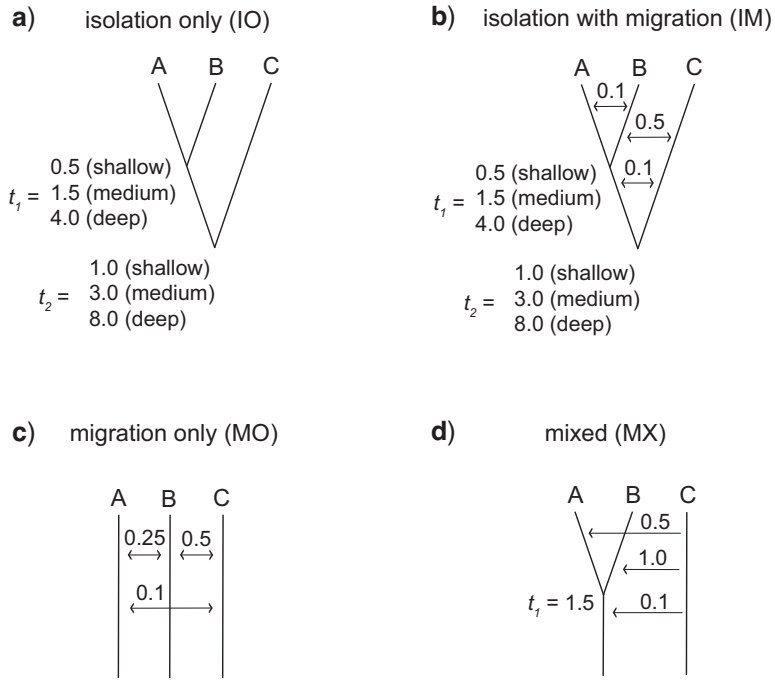

FIGURE 2. Demographic models simulated to test PHRAPL performance in model selection and parameter optimization: (**a**) isolation-only (IO) models with two coalescence events, t_1 and t_2 (times are given in units of $4N$); (**b**) isolation-with-migration (IM) models (migration rates above arrows are given in units of $4Nm$); (**c**) a migration-only (MO) model; and (**d**) a mixed model (MX), which includes one coalescence event and migration in some, but not all directions. For IO and IM models, we considered three divergence depths (shallow, medium, and deep), for a total of eight histories.

simulating data under a known model with known parameters and then observing how often PHRAPL can infer the true history. Here we report results for tests involving eight simulated histories that reflect four classes of demographic models (Fig. 2): isolation-only (IO), isolation-with-migration (IM), n-island migration-only (MO), and a mixed (MX) model, which is intermediate to IM and MO models and includes one population coalescent event, with migration between some, but not all populations. Under each history we simulated gene trees using *ms* for five different dataset sizes (1, 5, 10, 50, and 100 loci), with 50 replicates for each treatment. Parameter values spanned an order of magnitude of migration rates and divergence times and were chosen to reflect the breadth of variation commonly observed within phylogeographic datasets. To simulate the stochastic mutational process underlying empirical datasets, rather than analyzing raw genealogies from *ms*, we analyzed genealogies that were inferred from sequence datasets that we simulated for each of the original trees using Seq-Gen (Rambaut and Grassly 1997; See Supplementary Methods). All datasets were analyzed with PHRAPL using a common set of 17 models that included all four true histories as well as additional "decoy" histories containing alternative topologies and migration matrices (Supplementary Fig. S6).

PHRAPL was generally accurate at identifying the true model (i.e., the model used to generate the data), although in most cases, this was contingent on sampling many loci (Fig. 4). In cases where the generating model did not receive the best AIC score, this model was often ranked second-best and surpassed by a model with a similar set of parameters. As with many phylogeographic methods, PHRAPL accuracy improved as the size of the dataset was increased.

There were two scenarios for which accuracy did not approach 100%, even with a large number of loci: 1) IO models with shallow divergence and 2) MO models. In the former case, IO models with zero migration were difficult to distinguish from IM models with small estimated migration rates. If using PHRAPL as a tool to perform discrete model selection, this can lead to false inference of an IM model in instances of recent divergence. For this reason, it is important to also consider parameter estimates. Observing that migration rate estimates were typically low in these cases (Fig. 5) would lead one to the correct inference that migration is not a very important process underlying these genetic data, despite an IM model being inferred. For cases in which simulated datasets were generated under MO, inferred models usually included one or two divergence parameters in addition to full migration, suggesting the PHRAPL is biased towards models that include divergence, which is worrisome for systems in which migration is so important as to have swamped out the genetic signal of the underlying divergence history. That said, PHRAPL was still able to accurately estimate migration rates for these datasets (Fig. 5). Nevertheless, one important caveat is that PHRAPL tends to be biased

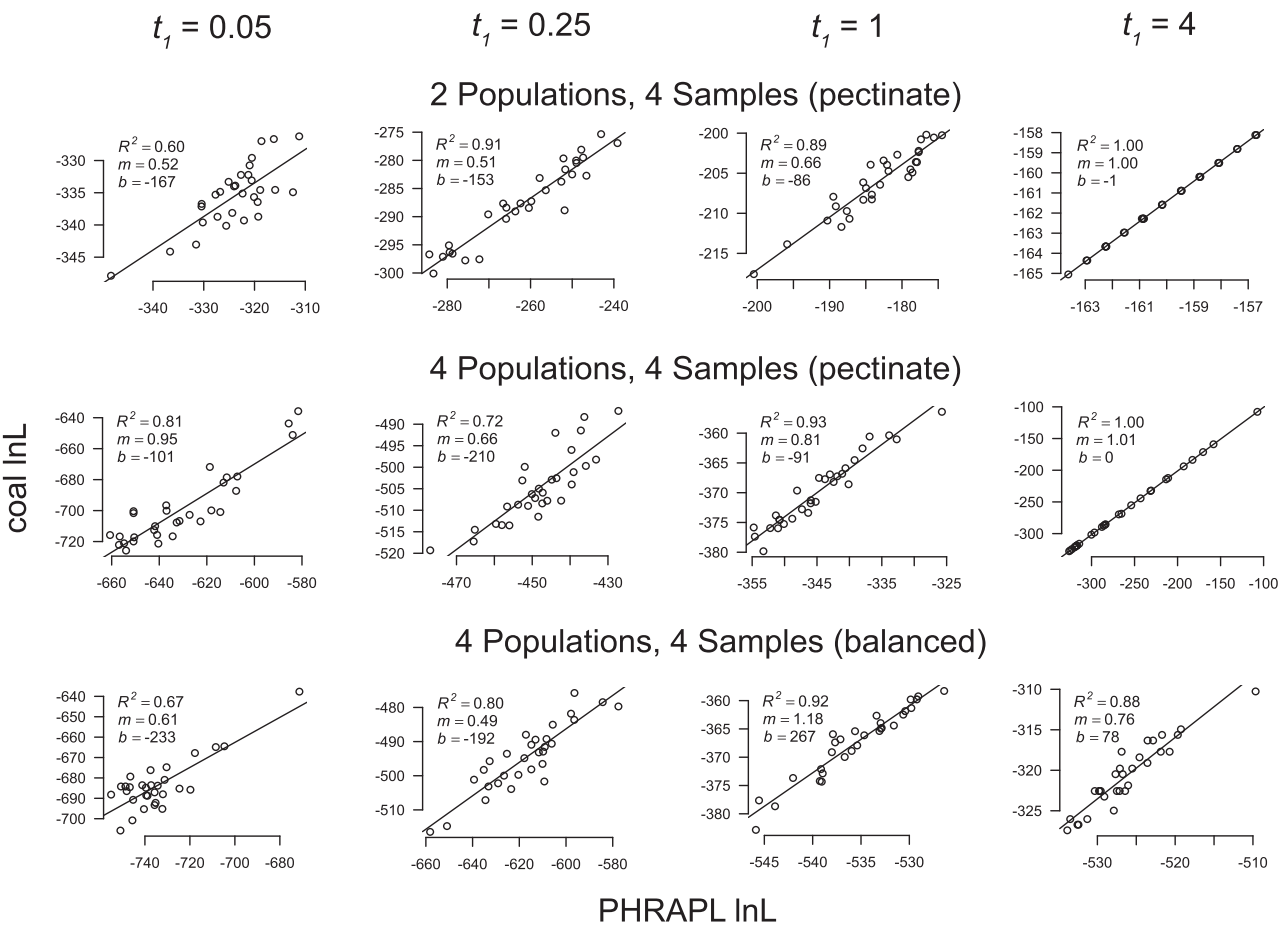

FIGURE 3. Correlations between log-likelihoods (lnL) calculated analytically using COAL and approximate lnLs calculated using PHRAPL for a subset of simulated treatments. R^2, slope (m), and intercept (b) are given for each analysis. For a few treatments, R^2 was greatly reduced due to a few outlier lnL values obtained from COAL, which likely reflect errors. Also, about one percent of COAL analyses failed; results from these trees are excluded.

toward selecting models that contain some measurable divergence among populations.

Parameter estimates were generally accurate using PHRAPL (Fig. 5), with a couple notable exceptions. First, divergence times for ancestral populations (i.e., t_2 in Fig. 2) tended to be overestimated when the true value was small or moderate. Second, for datasets generated under MO, non-zero coalescence times were estimated for populations A and B and for populations AB and C. This resulted from PHRAPL's preferential selection of the IM model that contained a migration matrix most similar to that used to generate the MO datasets.

By way of practical advice, we recommend using multiple loci, at least 10, but preferably more than that, to minimize errors in model selection. We also recommend—particularly when an IM model is inferred—that users inspect parameter values in addition to the discrete model(s) selected. In part, this is because PHRAPL is not sufficiently precise to distinguish models with very low migration from those with zero migration. Moreover, discrete model classification can be overly crude, depending on the goals of a researcher, and parameter values can help one to distinguish between migration rates that may be statistically versus biologically significant. For example, if an IM model is selected, but migration rates are very low (e.g., $M=0.01$), the most salient result might be that migration is largely unimportant. After all, an IM model in which $M=0.01$ is arguably more similar to an IO model than to an IM model accompanied by high migration (e.g., $M=2$) in respect to the influence of migration on phylogenetic patterns. By this same token, models with extremely high estimated migration rates could be interpreted as panmixia. To some extent, one can shape the definition of an "IM" model to suit one's goals by narrowing the range of migration rates that PHRAPL will consider. For example, if a researcher only cares about migration occurring at a rate greater than $M=0.5$, one should make 0.5 the minimum value in the grid.

Finally, it is important to note that when more than three populations are modeled, the number of possible models can become very large such that some restricting of model space becomes necessary. In these cases, available prior information can be used to focus the model space that is surveyed. For example, migration

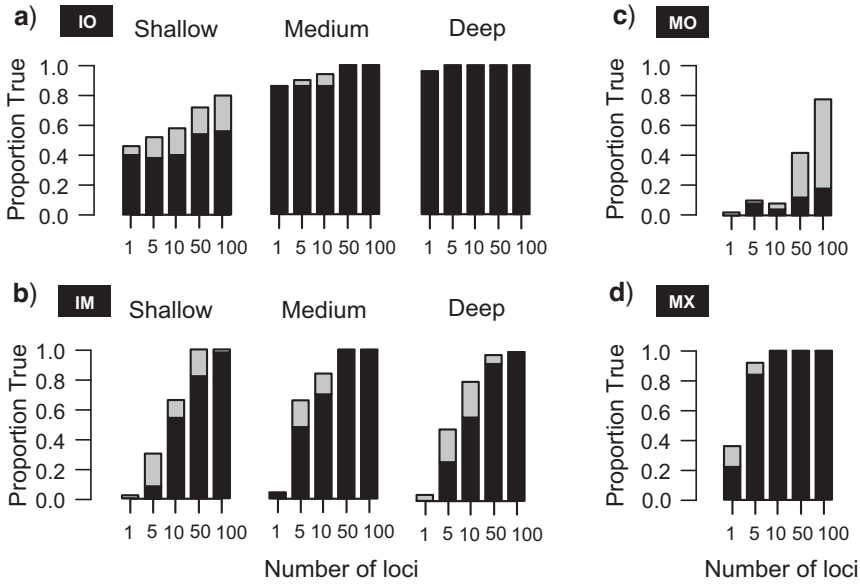

FIGURE 4. PHRAPL model selection results from analysis of the four types of simulated models depicted in Fig. 2: (**a**) isolation-only (IO), (**b**) isolation-with-migration (IM), (**c**) migration-only (MO), and (**d**) mixed (MX). Black bars give the proportion of 50 replicate analyses in which the true model garnered the highest AIC weight; gray bars give the proportion in which the true model garnered the second highest AIC weight.

can be limited to geographically adjacent populations or migration rates can be forced to be symmetrical. There are possible ways to search over a voluminous model space more objectively as well. One possible technique is to carry out analyses using a multi-step approach (e.g., Morales et al. 2016). For example, a more general model set may be analyzed initially (e.g., those with more types of parameters, but with stringent limits on the number of free parameters for any given type), followed by a subsequent analysis using a model set that is tailored to the results (e.g., one that focuses on one or two types of parameters that appear most important). Alternatively, one can initially analyze a random subsample of models in an oversized model set; a smaller model set can then be constructed and exhaustively analyzed, based on results from the random subset.

CONTRAST WITH APPROXIMATE BAYESIAN COMPUTATION

There are many similarities between our approach and that of approximate Bayesian computation (ABC) methods (Beaumont et al. 2002), which are also commonly applied to phylogeographic questions (e.g., Fagundes et al. 2007; Peter et al. 2010). ABC is a powerful approach that, like PHRAPL, can compare the relative fit of complex, customized demographic models to statistical summaries of genetic data while also estimating parameters, although it does this in a Bayesian framework. One disadvantage of ABC, which is shared by nearly all available analytical methods that carry out phylogeographic model selection, is that a relatively small set of models must typically be invoked *a priori*. PHRAPL was specifically designed for model exploration and consequently implements a method to systematically generate, test, and compare large and diverse sets of models. In contrast, performing model selection with more than a handful of models using ABC can be a laborious, time-consuming task (e.g., Pelletier and Carstens 2014).

Another difference between ABC and PHRAPL is that the latter method calculates the likelihood by counting exact matches between observed and expected tree topologies. Given the size of the parameter space, use of exact matches is only rendered feasible by the implementation of the efficiency strategies described above. For this reason, when using ABC, one typically calculates the distance between observed and expected summary statistics and accepts values that fall within a pre-specified threshold distance. While counting exact matches helps PHRAPL to more rapidly approximate the likelihood, it is possible that implementing a distance approach in PHRAPL—in which topologies that are close, but not identical, are somehow incorporated into the likelihood function—could improve model selection performance in some cases, particularly for datasets that are small or that have undergone recent divergence, while keeping the analysis computationally tractable. This could be a fruitful avenue for future research. Furthermore, the PHRAPL framework is amenable to the incorporation of other data beyond tree topologies, such as those pertaining to branch lengths, a possibility that would also be worth pursuing in a future study.

COMPUTATION TIME AND SAMPLING LIMITS

Computation time requirements of PHRAPL are similar to those of other methods commonly used by phylogeographers. Using a single core (2.66–3.33 GHz processor speed with 32–64 GB memory), individual

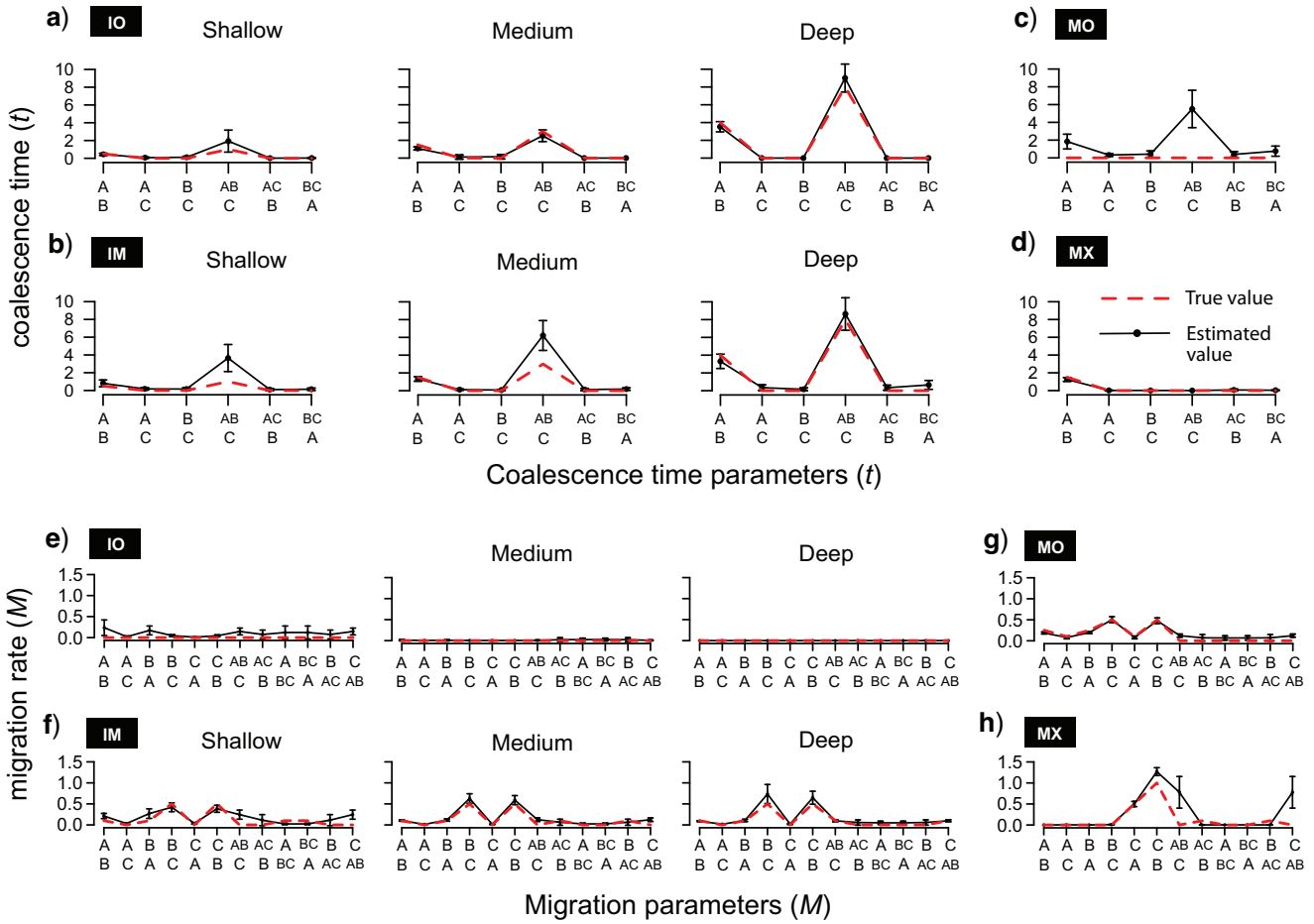

FIGURE 5. Model averaged parameter values from analyses depicted in Fig. 4 (averaged across 10, 50, and 100 loci treatments). Estimated parameter values (black dots connected by solid black lines) are shown against true parameter values (followed by the red dashed line). (**a, e**) isolation-only (IO) models (for three tree depths); (**b, f**) isolation-with-migration (IM) models (for three tree depths); (**c, g**) migration-only (MO) models; and (**d, h**) mixed (MX) models. Specific parameters (numbered on the x-axis) include all parameters present in the model set, although no single model contains all parameters. For coalescence time parameters (t), populations above coalesce with populations below, where populations with two letters represent ancestral populations. For migration rates (M), migrants from populations listed above move into populations listed below. Note that for coalescence time parameters under MO models, although we have plotted the true values as zero, they could also be plotted as infinity.

demographic models ran for anywhere between a few seconds to 26 h, depending on the number of populations and samples per population, the number of parameters modeled and loci used, and the underlying history (see Supplementary Fig. S7). The method is easily amenable to cluster computing, and we hope to expand functionality such that optimization of a single model can be distributed across multiple nodes. Under the current implementation of PHRAPL, the maximum subsampled tree size that can be effectively used appears to be ~16 tips. For trees with 20 tips, topology matches among observed and predicted trees were rare (when nTrees = 100,000), making it impossible to approximate lnLs given our computational constraints. This places a logistical limit on the number of populations and number of individuals subsampled per population that can be used. For example, if you have five populations, you should subsample no more than three individuals per population (for a total of 15 tips). For this reason, PHRAPL will likely perform best when applied to systems containing six or fewer populations. Eight or nine populations is likely the absolute maximum, but we have yet to test PHRAPL performance using such high values.

CONCLUSIONS

PHRAPL is a valuable addition to the set of analytical tools available to phylogeographers and can be used in one of several ways. First, its signature feature is to explore phylogeographic model space, enabling researchers to more objectively infer those parameters that have shaped a given dataset, and thus reducing and/or testing one's reliance on prior assumptions. Once a set of parameters of interest have been identified, PHRAPL—or another method that can estimate parameters with more rigor, if one is available—can then be used to estimate those parameters, given

the best model(s) (e.g., Morales et al. 2016). This enables PHRAPL to test phylogeographic hypotheses that are customized to a particular empirical system. PHRAPL can also be used to estimate parameters under complex models that are not implemented in existing software packages. Finally, PHRAPL may have applications to species delimitation, particularly because existing methods for species delimitation do not explicitly incorporate gene flow (e.g., Carstens et al. 2013).

Supplementary Data

Data available from Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.1414v.

Funding

This work was supported by the National Science Foundation (DEB 1257669, DEB 1257784).

Acknowledgments

We thank Jack Sullivan, Darin Rokyta, and members of the Carstens and O'Meara labs, as well as students in the first PHRAPL workshop for conversations related to this work. We also thank David Posada, Jeremy Brown, and two anonymous reviewers for comments that improved this article.

References

Beaumont M.A., Zhang W., Balding D.J. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Beerli P., Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl Acad. Sci. USA 98: 4563–4568.

Burnham K.P., Anderson D.R. 2002. Model selection and multimodel inference: a practical infromation-theoric approach. New York: Springer.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22:4369–4383.

Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Fagundes N.J.R., Ray N., Beaumont M., Neuenschwander S., Salzano F.M., Bonatto S.L., Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc. Natl Acad. Sci. USA 104:17614–17619.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–920.

Hird S., Kubatko L., Carstens B. 2010. Rapid and accurate species tree estimation for phylogeographic investigations using replicated subsampling. Mol. Phylogenet. Evol. 57:888–898.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Hudson R.R., Turelli M. 2003. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57:182–190.

Johnson S.G. 2008. The NLopt nonlinear-optimization package. Available from: URL http://ab-initio.mit.edu/nlopt.

Knowles L.L. 2001. Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshoppers. Mol. Ecol. 10:691–701.

Knowles L.L., Maddison W.P. 2002. Statistical phylogeography. Mol. Ecol. 11:2623–2635.

Mebane W.R., Sekhon J.S. 2011. Genetic optimization using derivatives: the rgenoud package for R. J. Stat. Software 42:1–26.

Morales A.E., Jackson N.D., Dewey T.A., O'Meara B.C., Carstens B.C. 2016. Speciation with gene flow in North American *Myotis* bats. Syst. Biol. 66:440–452.

Nordborg M. 2001. Coalescent theory. In: Balding D.J., Bishop M., Cannings C. editors. Handbook of statistical genetics. West Sussex, UK: John Wiley and Sons. p. 179–212.

O'Meara B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59:59–73.

Pelletier T.A., Carstens B.C. 2014. Model choice for phylogeographic inference using a large set of models. Mol. Ecol. 23:3028–3043.

Peter B.M., Wegmann D., Excoffier L. 2010. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol. Ecol. 19:4648–4660.

Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.

Rambaut A., Grassly N.C. 1997. An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, version 1.2.5. Comp. Appl. Biosci. 13:235–238.

Saunders I.W., Tavaré S., Watterson G.A. 1984. On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. 16:471–491.

Templeton A.R. 2010. Coalescent-based, maximum likelihood inference in phylogeography. Mol. Ecol. 19:431–435.