Trends in Ecology & Evolution

CellPress

Opinion

Assessing model adequacy leads to more robust phylogeographic inference

Bryan C. Carstens [1,*], Megan L. Smith,[2] Drew J. Duckett [1], Emanuel M. Fonseca [1], and M. Tereza C. Thomé[1]

Phylogeographic studies base inferences on large data sets and complex demographic models, but these models are applied in ways that could mislead researchers and compromise their inference. Researchers face three challenges associated with the use of models: (i) 'model selection', or the identification of an appropriate model for analysis; (ii) 'evaluation of analytical results', or the interpretation of the biological significance of the resulting parameter estimates, delimitations, and topologies; and (iii) 'model evaluation', or the use of statistical approaches to assess the fit of the model to the data. The field collectively invests most of its energy in point (ii) without considering the other points; we argue that attention to points (i) and (iii) is essential to phylogeographic inference.

Highlights

Phylogeography makes inferences about the evolutionary history of species by using statistical models of historical demography to analyze genetic data. These models are increasingly complex and sometimes applied in ways that can compromise the quality of phylogeographic inference.

Inferences are most often derived from estimates of evolutionary parameters made using these models. Parameter estimates are contextually dependent on the model used to estimate the parameters and are informative only if the model is a reasonable fit to the data.

A variety of approaches can be used to assess model adequacy, from simple visual examinations to statistical goodness of fit tests. The increased power and interpretability of statistical approaches justify their increased complexity.

A review of existing software packages demonstrates that, when tests for model adequacy are built into software packages by developers, users are more likely to conduct these analyses.

The rise of phylogeographic models

The past decades have seen an exponential increase in the amount of genetic data that can be collected from a given species [1,2]. For scientists conducting research into non-model systems, this transition from data poor to data rich occurred rapidly and its effects are still being felt. For example, at the time when high-throughput sequencing data became available, phylogeography was in the process of transforming from a discipline in which inferences were qualitative and based on visual patterns to one in which inferences result from the statistical analysis of **demographic models** (see Glossary). Researchers suddenly needed to develop new protocols in the wet lab, new bioinformatics resources, and, in many cases, new analytical methods. There are now dozens of 'off-the-shelf' programs that implement demographic models (e.g., [3–6]), and researchers can design models that are customized to particulars of their empirical system and either perform model selection or parameter estimation using various software (e.g., [7,8]). While the quantity of data and the widespread availability of complex demographic models are boons to researchers, they create the potential for the misuse and abuse of models when inappropriate models are applied. Model misspecification can take many forms, from incorrectly modeling the number of populations or lineages (i.e., **delimitation misspecification**), to incorrectly modeling the evolutionary relatedness among population lineages (i.e., **topology misspecification**), to incorrectly modeling the evolutionary processes that have influenced the system (i.e., **parameterization misspecification**). To complicate matters, these errors can compound one another; for example, parameterization misspecification can lead to biases in the estimates of other parameters (e.g., [9,10]). Fortunately, there are effective examples of strategies that can be used to identify and limit the misspecification of demographic models.

Researchers who intend to make inferences from model-based analyses are faced with challenges inherent to three different issues. They must first identify an appropriate model for analysis, whether by picking an available implementation of a particular model or by designing a custom model. The use of **phylogeographic model selection** is one common approach. Second, researchers must evaluate the results of the analysis; that is, interpret the biological significance

[1]Department of Evolution, Ecology, and Organismal Biology at The Ohio State University, Columbus, OH, USA
[2]Department of Biology, Indiana University, Bloomington, IN, USA

*Correspondence:
carstens.12@osu.edu (B.C. Carstens).

of the delimitations, topology, and parameter estimates. Since such inferences are often the goal of the entire research program, it can be tempting to stop here without attempting a critical assessment of the results. However, researchers who confront the third challenge by attempting to assess the statistical fit of the demographic model, for example, by conducting a **goodness-of-fit test** or an exploration of **model adequacy**, are rewarded with important contextual information about the confidence that they should place in the analysis. This third step, while essential for interpreting parameter estimates and contextualizing results, is often overlooked due to inherent difficulties in assessing the fit of the complex models often used, limited computational resources, and the lack of out-of-the box software for assessing the fit of many popular models. Here, we first discuss the role of models in phylogeographic inference. Then, we examine the potential negative effects of model misspecification and demonstrate how evaluating model fit can help to limit these effects. Finally, we discuss the limitations of current approaches and future directions.

Identifying an appropriate model for analysis

Modeling the evolutionary history of an empirical system is ideally an integrative process, in which researchers use both off-the-shelf and custom models, complete with assessment of model adequacy for these models, in combination with information from other types of data as the basis for phylogeographic inference [11]. Clues about the appropriate demographic model(s) to use might come from climate data, which can indicate regions of historical habitat stability (e.g., [12,13]); paleopollen data, which can predict the presence of a species in a given region at some point in the past (e.g., [14]); and geological data, which can suggest that diversification occurred at a particular time (e.g., [15]). In cases in which an implemented model is available (i.e., off the shelf as part of some program), researchers can estimate parameters under the model, with inferences made directly from these parameter estimates. In cases in which researchers design custom models, they often turn to phylogeographic model selection, which facilitates integration of different data types by providing the basis for decisions about which models to consider [11]. Using this approach, researchers can select the best model from a predetermined set using objective criteria (e.g., [16,17]).

The dangers of model misspecification

Regardless of how a model is chosen, all demographic models are incorrect to some extent because natural populations are complex and the processes that shape genetic variation numerous. Even though any model is necessarily a simplification of these processes, some might still be useful. However, since some model misspecifications can mislead inference, it is essential that researchers are aware of such misspecifications so that they have appropriate confidence in their results. Numerous studies have evaluated the effects of model misspecification in phylogeography and related fields. For example, failing to model gene flow when it is present can lead to underestimating divergence time and overestimating population size [10,18]. By contrast, unaccounted-for population structure within species appears to have little effect on parameter estimates under an isolation-with-migration model [19] but can mislead species delimitation [20,21]. The effects of recombination appear to be minimal when some effort is made to use nonrecombining blocks for inference [19], but misspecifying the model of nucleotide substitution can lead to substantial inference errors [19]. Finally, ignoring the role of natural selection in shaping genetic diversity can also bias parameter estimates [22]. When background selection is ignored, population growth can be inferred even when the true population history is of constant population size [23,24]. Ghost introgression (Box 1), or introgression from unsampled populations or species, is likely a common model violation and appears to have substantial effects on parameter estimates [19]. If researchers are unaware of model misspecifications in their focal system, they might place too much confidence in these results, which are potentially affected by this

Glossary

Cross-validation: simulate data under a model (or models) and analyze simulated data to assess accuracy of parameter estimation (or model selection); cross-validation is applicable to all approaches to model selection and parameter estimation; built-in to some pipelines, including 'abc' [50] and DIYABC [28].
Delimitation misspecification: type of model misspecification in which the number of populations or lineages is incorrect, or the assignment of individuals to these groups is invalid. For example, lumping individuals sampled from two lineages into a single lineage would be expected to increase estimates of $\theta = 4N_e\mu$ and could lead to inaccurate estimates of divergence time.
Demographic model: representation of the evolutionary history of a focal system, comprising sets of individual samples from populations or evolutionary lineages (i.e., the delimitation of samples), the topological relatedness of these sets (i.e., the topology), and the evolutionary processes that influence genetic diversity in the system (i.e., the parameters). Evolutionary parameters, such as divergence times (τ), population sizes (θ), and migration rates (m), are modeled based on values from empirical data.
Goodness-of-fit test: evaluation of the statistical fit of a particular model given the data. These typically proceed by summarizing any difference between observed values of statistics and the values expected under the model.
Model adequacy: evaluation of whether the model is appropriate for the empirical data, which proceeds by assessing whether the degree of observed variation in the data is expected under the assumptions of the model.
Parameterization misspecification: type of model misspecification in which the evolutionary processes that have influenced a system are not modeled correctly. For example, in systems in which high rates of gene flow have occurred, failure to model this process can lead to inaccurate estimates of divergence time [10] and, in some cases, failure in species delimitation [43].
Parametric bootstrap: method for assessing model fit by simulating data under the maximum likelihood estimates of the parameters of the model, calculating a summary statistic on the

Box 1. Power analyses

To illustrate the process and benefits of assessments of data and model adequacy, we considered a simple scenario in which researchers aimed to evaluate three models in fsc2 [21]. Briefly, three models (Figure I) were tested. We simulated ten haploid individuals per population and 10 000 independent SNPs with the following priors: N_e = U(1000, 100 000 haploid individuals), t = U(1000, 100 000 generations), m = U(0.01, 20 Nm), and tm = U(1000, 50 000)). To assess power, we simulated ten data sets under each model and then selected the best model using AIC following [30]. Results indicate that statistical power was moderate for distinguishing among the three models. However, the divergence-only model (Model 1) and divergence-with-gene-flow model (Model 2) were difficult to distinguish, and the secondary contact model was sometimes mistaken for a divergence-only model. This difficulty in distinguishing models is likely related to the wide priors on migration rates and divergence times. Knowledge of such error rates would prevent researchers from overinterpreting results.

A common critique of demographic model selection is reality is likely more complex than any of the models in the model set and, thus, the identification of a 'best' model from a set of inadequate models might not be helpful. To evaluate this scenario (i.e., when reality is more complex than the set of evaluated models), we also simulated ten data sets under a model that included gene flow from an unsampled ghost population into one of the two populations (Model 4; Figure I), with the expectation that this could lead us to select a model of gene flow between the two populations even though this was not the true history [19]. We simulated ten data sets under this model, using the same priors as used for the other models. Gene flow began in the present and ended halfway between the present and the first divergence time in the model. We repeated the power analysis (described earlier), and found that the divergence with gene flow model was selected in 90% of replicates (Table I). These results suggest that an empirical investigation that relies on model selection might incorrectly infer gene flow among two sampled populations due to the presence of gene flow into one of these populations from an unsampled outgroup. Conducting an analysis designed to evaluate the statistical fit of the model to the data (see Box 2 in the main text) is an important next step.

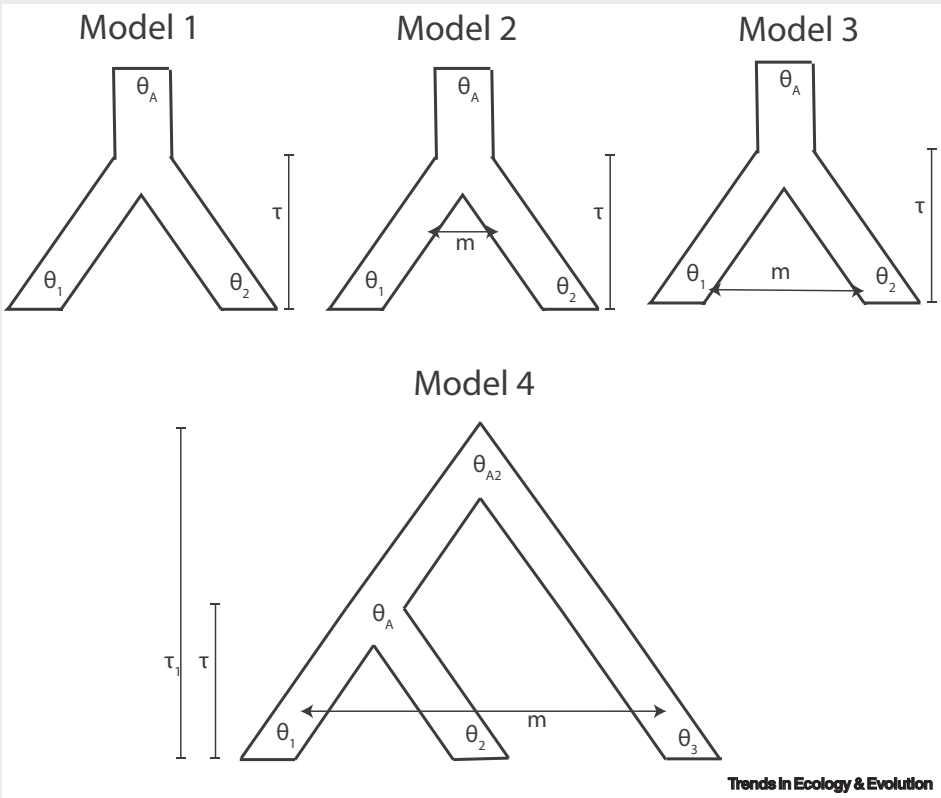

Figure I. Example models of historical demography. Evolutionary parameters, such as divergence time (τ), population sizes (θ), and gene flow (m_i), are modeled based on values from empirical data. The particular models evaluated using fsc2 in Boxes 1 (upper) and 2 (lower) in the main text are shown here.

simulated data sets, and comparing the same statistic calculated on empirical data with this null distribution. Parametric bootstraps can be applied in hypothesis testing (e.g., [51]) or model selection (e.g., [32]).

Phylogeographic model selection: objective evaluation of a set of historical demographic models, in which statistical techniques, such as Akaike information criterion (e.g., [52]), approximate Bayesian computation (e.g., [16]), or machine learning (e.g., [17]), are used to rank and identify the best models in the set.

Posterior predictive simulation (PPS): method for assessing model fit in a Bayesian context by drawing parameters from posterior distributions, simulating data under those parameters, analyzing the simulated data to generate 'posterior predictive distributions', and comparing these distributions to the original posterior distribution (or to summary statistics); for more information, see Box 3 in the main text.

Topology misspecification: type of model misspecification in which the topology of the evolutionary lineages is represented incorrectly, leading to potential errors in estimates of $\theta = 4N_e\mu$, divergence times, or other evolutionary parameters.

Table I. Example of power analysis[a]

Simulating model	Model 1	Model 2	Model 3
Model 1	0.4	0.5	0.1
Model 2	0.3	0.7	0
Model 3	0.2	0	0.8
Model 4	0.1	0.9	0

[a]Each row shows results from the fsc2 analysis of data simulated under the respective model. Values in each column show the proportion of replicates for which the generating model was selected as the best model using AIC values and information theory.

misspecification. Thus, approaches that allow researchers to assess model fit and to understand how this might affect inference are essential for robust phylogeographic inference.

Assessing model fit

Assessments of model fit enable biologists to ask how well the model they are using to analyze the data explains the pattern of variation in the genetic data. While phylogeographers already frequently use model selection to select from among a set of models the model that best fits their empirical data (e.g., [17]), assessments of absolute model fit are not always utilized (or possible) under commonly used approaches. In phylogeographic model selection, researchers can only ask which of a set of models best fits their data, and not whether the models provide a good fit to the observed data [25]. Thus, results can be positively misleading when researchers have selected the least-bad model from a set of models that all offer a poor fit to the data. Since all models are a simplification of reality, assessments of model fit that are also informative with respect to whether model violations are likely to affect inference are the most valuable to phylogeographers. An ideal approach would allow researchers not only to detect model violations, but also to infer which factors are responsible for poor model fit and how likely these factors are to affect various aspects of inference. Given that genetic data do not fit the assumptions of standard statistical distributions, such as χ^2, which are used to assess model fit in other contexts, biologists often rely on data simulation to build test distributions for assessments of model fit. Simulations are performed under the selected model and estimated parameters to construct test distributions. These test distributions can then be compared with the observed data to assess how well the model reproduces the characteristics of those data, with large differences in the distributions indicating that the model fits the data poorly. Visual approaches to model checking, **parametric bootstrapping**, or **posterior predictive simulation (PPS)** can be used to assess model fit.

Visual approaches

Visual approaches do not allow any statistical quantification of model fit, but can alert users to blatant model violations [26,27]. The most popular visual checks commonly applied to assess model fit are principal component analysis (PCA) and linear discriminant analysis (LDA). For example, in the popular software package DIYABC [28,29], users can assess whether their priors produce data that are compatible with their observed data using LDA by projecting the first two axes of their simulated and observed data and visually assessing whether observed data fall into the cloud of simulated data. This approach requires minimal computational resources, as the prior must be simulated before inference for many approaches to demographic model selection (e.g., approximate Bayesian computation; ABC). However, power to detect model violations is likely limited with these approaches because there is no formal assessment of model fit.

Parametric bootstraps

Parametric bootstraps have long been used to test hypotheses in phylogenetics and phylogeography (e.g., [11]); they utilize data simulation under the maximum likelihood parameters estimated under some model to build a test distribution that can be used to evaluate the fit of the model to the empirical data. For assessing model adequacy, a goodness-of-fit test proceeds by simulating data under the maximum-likelihood parameter estimates made from some model (e.g., [7]) before calculation of the likelihood ratio G-statistic [30] using Equation 1:

$$CLR = \log_{10}\frac{CL_O}{CL_E} \quad [1]$$

where CL_O is the observed likelihood and CL_E is the estimated likelihood for both the simulated and observed datasets. Finally, the user calculates the *P* value of the observed statistic and assesses whether the model is a poor fit to the data. Ideally, this *P* value will be close to 0.5, indicating that the model is a good fit to the data, but most practitioners choose an α level of 0.05 for a threshold that designates a significant model violation. We provide an example of the use of this test to assess model adequacy in Box 2.

Posterior predictive simulations

PPS is a Bayesian version of the parametric bootstrap used for model checking in a Bayesian framework [31]. PPS samples from the posterior distribution of an empirical analysis and simulates data using these samples under the model used to analyze the data, creating a posterior-predictive distribution. Next, test statistics calculated from both the empirical and posterior-predictive distributions are compared.

When the posterior-predictive distribution is similar to the observed data as quantified by the test statistic, the user concludes that the model is a reasonable fit to the data, but when the posterior-

Box 2. Composite likelihood ratio test

Next, we performed composite-likelihood ratio (CLR) tests following Excoffier *et al.* [7] to assess how well the data fit each model (see Figure I in Box 1). For each model, we assessed whether we could detect violations of each of the three original models by simulating 100 bootstrap site frequency spectra (SFS) from the maximum-likelihood parameter estimates under each model. We then performed parameter estimation, estimated the composite likelihood for these data sets under each model, calculated the difference in the observed and estimated likelihood for each bootstrap data set, and compared this with the same calculation for the original data sets [7]. We calculated the *P* value as the proportion of times the difference in observed and estimated likelihoods from the bootstrap replicates exceeded the value for the original data sets, and we considered $P < 0.01$ to indicate model violations. For the first three models, we never detected a model violation when evaluating the model used to simulate the data, and our ability to detect violations of the other two models varied (Table I). However, for data sets generated under the ghost introgression model, we nearly always detected model violations, except for a single replicate, which was not detected as violating the secondary contact model. This suggests that, although model selection results could have been misleading, a CLR test would have prevented researchers from overinterpreting incorrect results by identifying model violations.

Table I. Example application of CLR tests of model fit using fsc2[a]

	Model 1	Model 2	Model 3
Model 1	0	0	0.4
Model 2	0.5	0	0.5
Model 3	0.8	0.1	0
Model 4	1	1	0.9

[a]Each row shows results simulated from the 'simulating model', and each column the proportion of replicates (of ten replicates) for which a model was rejected using the CLR.

predictive distribution differs substantially from the empirical data, the user concludes that the model is a poor fit. The distributions are generally compared using posterior predictive *P* values (see [26,27] for more detailed discussion). PPS has been used in a variety of contexts in phylogenetics and phylogeography (Box 3).

Choice of test statistics

The choice of test statistics is essential to both the parametric bootstrap and PPS. Test statistics can be broadly divided into inference-based and data-based statistics [32]. Inference-based statistics require that researchers analyze posterior predictive data sets using the same inference methods and model as used for the empirical data. Then, some test statistic based on this analysis is calculated and compared between empirical and posterior predictive data sets (see [33–37] for examples). Inference-based statistics have been lauded because they allow researchers to not only assess whether a model is violated, but also to understand how that violation might affect inference [32,33]. However, such statistics can be computationally intensive to calculate since they require that full inference be performed on each posterior predictive data set. Data-based statistics are calculated directly from empirical data and posterior predictive data and, as such, are easier to calculate than are inference-based statistics. The choice of test statistic should ideally provide a balance between computational feasibility, power to detect model violations, and informativeness as to the aspect of the data that drove the model violation and the likely effects of the model violation on inference. Many approaches for assessing model fit use several complementary test statistics (e.g., [32,34,36,38–40]).

Box 3. Applications of posterior predictive simulations

PPS have been used in many evolutionary investigations. Inspired by a symposium at the 2012 Evolution Annual Meeting sponsored by the Society of Systematic Biologists, the past decade has seen an increase in software development to assist researchers in performing PPS analyses.

Phylogenetics

Some popular phylogenetics applications have built-in methods for conducting posterior predictive checks. For example, RevBayes includes P3 for checking model adequacy for phylogenetic analyses [29]. Similarly, BEAST includes TreeModelAdequacy for assessing phylodynamic models commonly used for analysis of pathogens [39]. In addition, several packages have been designed to supplement various phylogenetic analyses, including Posterior Predictive Checks of Coalescent Models (P2C2M) [40] and modadclocks [42]. Examples of these methods in application to empirical research include [43,44].

Species delimitation

PPS analyses have also been used to evaluate model adequacy for species delimitation. For example, Barley *et al.* [20] applied PPS to assess the adequacy of the multispecies coalescent for conducting species delimitation with two delimitation methods, finding that violations of the multispecies coalescent can negatively affect species delimitation, but that these violations can often be detected with PPS. Additionally, Barley and Thompson [34] showed that the substitution model can have important effects on the number of operational taxonomic units delimited when using the Automated Barcode Gap Discovery method of species delimitation. Fonseca *et al.* [35] provide an implementation of P2C2M designed to assess the model adequacy of the Generalized Mixed Yule Coalescent model.

Historical demography

PPS can be efficiently applied when examining historical demographic models with ABC because summary statistics are calculated for the empirical data set as part of the ABC process. For example, Gao *et al.* [45] used PPS to show that a model of population bottlenecks provided a good fit in Chinese mountain pines (*Pinus densata*). Additionally, Tsuda *et al.* [46], used PPS to show that a model incorporating gene flow during divergence fit their empirical data better than did a model without such gene flow. The program DIYABC includes functions for conducting PPS when performing ABC in R [47]. PPS has also been applied to assess admixture models (e.g., [48]) and to identify loci under selection given demography (e.g., [49]).

Effective strategies for the proper use of demographic models

Empirical systems represent a single replicate of an experiment of unknown design that are potentially complex in ways that are impossible to quantify. The most effective strategy for making evolutionary inferences despite these difficulties is one in which researchers are open to exploration and willing to quantify uncertainty and assess model fit. After choosing a model, researchers should use a statistical approach to assess model fit, such as PPS or parametric bootstraps. When model violations are identified, researchers should exercise appropriate caution and quantify the certainty in their inferences. Researchers might attempt to isolate which aspect of the model is violated by applying different test statistics or by considering and testing alternate models. Simulations designed to explore model adequacy and the effect of specific model violations on parameter estimates are recommended. In any event, designing custom tests for empirical systems requires researchers to critically evaluate the source of inference in their study; for example, are inferences based on parameter estimates or the model itself? Researchers with a clear understanding as to the goal of the investigation and how the model will be used to accomplish this goal can prevent themselves from being misled by model misspecification. The end result should involve inference under the best model that the researcher was able to identify, a quantification of the fit of that model to the data, and an assessment of the amount of confidence that should be placed in parameter estimates under the model.

Barriers and future directions

To better understand how often and under what conditions researchers take steps to evaluate model fit, we conducted a literature review of studies that used popular software for model selection. We identified recent publications that cited either DIYABC [28,29] or fastsimcoal2 [7] using Google Scholar, filtered out studies that did not conduct empirical model selection, and determined whether the authors of the remaining studies conducted a check of model adequacy. Users are more likely to implement some assessment of model fit if the method is built into the program and computationally feasible. For example, DIYABC offers a built-in visual assessment of model fit by allowing users to perform a PCA and plot empirical and simulated data sets, and more than half of users (51%) adopted this approach. It also includes a method for **cross-validation**. By contrast, ~6% of users adopted the goodness-of-fit test suggested by Excoffier *et al.* [7]. This test is more computationally demanding because it requires that the user perform simulations under the maximum likelihood parameter estimates and also asks the user to analyze these simulated data sets individually, a process that requires the development of custom scripts. Clearly, ease of implementation and computational tractability are huge determinants of when researchers will assess model fit. Furthermore, when software packages offer flexibility in terms of the type of inference used, it can increase use. For instance, in RevBayes [41], both inference-based and data-based test statistics are available, and this program is widely applied in phylogenetic analysis. Future work should aim to implement evaluations of model fit either as additional easy-to-use software packages or, ideally, as a component of the software packages used for model selection and parameter estimation. In the meantime, even suboptimal assessments of model adequacy are likely better than ignoring this question entirely.

Concluding remarks

Although demographic models offer a powerful approach to phylogeographic inference, appropriate caution when interpreting results is vital. Assessments of model fit can allow researchers to identify model violations that might impact inference and should be an integral part of phylogeographic investigations. Future work to evaluate the power of various test statistics, and to provide easy-to-implement assessments of model fit for popular software packages would improve the rigor of phylogeographic research. See Outstanding questions for more details.

Outstanding questions

Which summary statistics are most appropriate for assessing model fit? Simulation studies are essential to determining the appropriate statistics for detecting violations under specific models. Carefully curated statistics can provide insight into not only whether a violation has occurred, but also why and whether the violation is likely to mislead inference.

When will model violations mislead inference? Since all models are necessarily misspecified to some extent, it is essential not only to identify model violations, but also to understand the practical implications of such violations.

What types of evolutionary process are most likely to lead to model misspecification? Genetic variation in all species is influenced by a complex mix of evolutionary processes. As more researchers conduct thorough investigations into model fit in a range of systems, we are likely to better understand which evolutionary processes are likely to lead to misspecification when ignored.

How can checks of model adequacy be faster? One downside of some model checks is that they demand a large quantity of computational resources. Researchers are more likely to conduct model checks when doing so is fast and easy.

How can evaluations of model adequacy be incorporated into more software and analytical pipelines? While evaluations of model adequacy are theoretically straightforward in many Bayesian and Likelihood applications, it is less clear how model checks can be performed in other contexts. For example, machine learning has become a popular tool for phylogeographers, but is it unclear how to assess model fit in many deep learning algorithms.

Acknowledgments

We thank the National Science Foundation (NSF) for supporting this work (NSF-DBI 1661029 to B.C.C. and NSF DBI-2009989 to M.L.S.). We thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (#88881.170016/2018) for their support of E.M.F. We thank present and past members of the Carstens lab for discussions that have improved this work.

Declaration of interests

None declared by authors.

References

1. Garrick, R.C. *et al.* (2015) The evolution of phylogeographic datasets. *Mol. Ecol.* 24, 1164–1171
2. Stephens, Z.D. *et al.* (2015) Big data: astronomical or genomical? *PLoS Biol.* 13, e1002195
3. Drummond, A.J. *et al.* (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192
4. Beerli, P. and Palczewski, M. (2010) Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* 185, 313–326
5. Hey, J. *et al.* (2018) Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* 35, 2805–2818
6. Schiffels, S. and Wang, K. (2020) MSMC and MSMC2: the multiple sequentially Markovian coalescent. In *Statistical Population Genomics* (Dutheil, J.Y., ed.), pp. 147–166, Humana
7. Excoffier, L. *et al.* (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9, e1003905
8. Gutenkunst, R. *et al.* (2010) Diffusion approximations for demographic inference: *DaDi*. *Nat. Prec.* Published online June 30, 2010. https://doi.org/10.1038/npre.2010.4594.1
9. Koopman, M.M. and Carstens, B.C. (2010) Conservation genetic inferences in the carnivorous plant *Sarracenia alata* (Sarraceniaceae). *Conserv. Genet.* 11, 2027–2038
10. Leaché, A.D. *et al.* (2014) The influence of gene flow on species tree estimation: a simulation study. *Syst. Biol.* 63, 17–30
11. Knowles, L.L. and Maddison, W.P. (2002) Statistical phylogeography. *Mol. Ecol.* 11, 2623–2635
12. He, Q. *et al.* (2013) Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution* 67, 3386–3402
13. Vasconcellos, M.M. *et al.* (2019) Isolation by instability: historical climate change shapes population structure and genomic divergence of treefrogs in the Neotropical Cerrado savanna. *Mol. Ecol.* 28, 1748–1764
14. Gavin, D.G. *et al.* (2014) Climate refugia: Using fossils, genetics, and spatial modeling to explain the past and project the future of biodiversity. *New Phytol.* 204, 37–54
15. Bagley, J.C. *et al.* (2018) Testing hypotheses of diversification in Panamanian frogs and freshwater fishes using hierarchical approximate Bayesian computation with model averaging. *Diversity* 10, 120
16. Fagundes, N.J.R. *et al.* (2007) Statistical models of human evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17614–17619
17. Fonseca, E.M. *et al.* (2021) Phylogeographic model selection using convolutional neural networks. *Mol. Ecol. Resour.* 21, 2661–2675
18. Jiao, X. *et al.* (2020) The impact of cross-species gene flow on species tree estimation. *Syst. Biol.* 69, 830–847
19. Strasburg, J.L. and Rieseberg, L.H. (2010) How robust are 'isolation with migration' analyses to violations of the IM model? A simulation study. *Mol. Biol. Evol.* 27, 297–310
20. Barley, A.J. *et al.* (2018) Impact of model violations on the inference of species boundaries under the multispecies coalescent. *Syst. Biol.* 67, 269–284
21. Sukumaran, J. and Knowles, L.L. (2017) Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. U. S. A.* 114, 1607–1612
22. Johri, P. *et al.* (2021) The impact of purifying and background selection on the inference of population history: problems and prospects. *Mol. Biol. Evol.* 38, 2986–3003
23. Johri, P. *et al.* (2020) Towards an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics* 215, 173–192
24. Ewing, G.B. and Jensen, J.D. (2015) The consequences of not accounting for background selection in demographic inference. *Mol. Ecol.* 25, 135–141
25. Thomé, M.T.C. and Carstens, B.C. (2016) Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proc. Natl. Acad. Sci. U. S. A.* 113, 8010–8017
26. Gelman, A. (2003) A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Stat. Rev.* 71, 369–382
27. Gelman, A. (2004) Exploratory data analysis for complex models. *J. Comput. Graph. Stat.* 13, 755–779
28. Cornuet, J.-M. *et al.* (2014) DIYABC v2. 0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics* 30, 1187–1189
29. Collin, F. *et al.* (2021) Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Mol. Ecol. Resour.* 21, 2598–2613
30. Nielsen, R. and Wiuf, C. (2006) Composite likelihood estimation applied to single nucleotide polymorphism (SNP) data. In *ISI 2005 Final Proceedings*, ISI, www.isi-web.org/isi.cbs.nl/iamamember/CD6-Sydney2005/ISI2005_Papers/279.pdf
31. Gelman, A. *et al.* (2009) *Bayesian Data Analysis* (2nd edn), Chapman & Hall/CRC Texts in Statistical Science
32. Brown, J.M. (2014) Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Syst. Biol.* 63, 334–348
33. Bollback, J.P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19, 1171–1180
34. Barley, A.J. and Thomson, R.C. (2016) Assessing the performance of DNA barcoding using posterior predictive simulations. *Mol. Ecol.* 25, 1944–1957
35. Fonseca, E.M. *et al.* (2021) P2C2M.GMYC: An R package for assessing the utility of the Generalized Mixed Yule Coalescent model. *Methods Ecol. Evol.* 12, 487–493
36. Duckett, D.J. *et al.* (2020) Identifying model violations under the multispecies coalescent model using P2C2M. SNAPP. *PeerJ* 8, e8271
37. Pons, J. *et al.* (2006) Sequence based species delimitation for the DNA taxonomy of undescribed insects. *Syst. Biol.* 55, 595–609
38. Reid, N.M. *et al.* (2014) Poor fit to the multispecies coalescent is widely detectable in empirical data. *Syst. Biol.* 63, 322–333
39. Duchêne, S. *et al.* (2019) Phylodynamic model adequacy using posterior predictive simulations. *Syst. Biol.* 68, 358–364
40. Gruenstaeudl, M. *et al.* (2016) Posterior predictive checks of coalescent models: P2C2M, an R package. *Mol. Ecol. Resour.* 16, 193–205
41. Höhna, S. *et al.* (2018) P3: Phylogenetic posterior prediction in RevBayes. *Mol. Biol. Evol.* 35, 1028–1034
42. Duchêne, D.A. *et al.* (2015) Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Mol. Biol. Evol.* 32, 2986–2995
43. Morales, A.E. and Carstens, B.C. (2018) Evidence that *Myotis lucifugus* 'subspecies' are five nonsister species, despite gene flow. *Syst. Biol.* 67, 756–769
44. Tongo, M. *et al.* (2018) Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages. *Virus. Evolution* 4, vey003
45. Gao, J.I.E. *et al.* (2012) Demography and speciation history of the homoploid hybrid pine *Pinus densata* on the Tibetan Plateau. *Mol. Ecol.* 21, 4811–4827

46. Tsuda, Y. *et al.* (2016) The extent and meaning of hybridization and introgression between Siberian spruce (*Picea obovata*) and Norway spruce (*Picea abies*): cryptic refugia as stepping stones to the west? *Mol. Ecol.* 25, 2773–2789
47. Cornuet, J.-M. *et al.* (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1. 0). *BMC Bioinformatics* 11, 1–11
48. Mimno, D. *et al.* (2015) Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proc. Natl. Acad. Sci. U. S. A.* 112, E3441–E3450
49. Adams, R.H. *et al.* (2017) GppFst: genomic posterior predictive simulations of F_{ST} and D_{XY} for identifying outlier loci from population genomic data. *Bioinformatics* 33, 1414–1415
50. Csilléry, K. *et al.* (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* 3, 475–479
51. Knowles, L.L. *et al.* (2007) Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Curr. Biol.* 17, 940–946
52. Carstens, B.C. *et al.* (2013) Model selection as a tool for phylogeographic inference: an example from the willow *Salix melanopsis*. *Mol. Ecol.* 22, 4014–4028