MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 4581–4582

NEWS AND VIEWS

PERSPECTIVE

Navigating the unknown: model selection in phylogeography

Models of population structure: tools for thinkers

BRYAN C. CARSTENS* and L. L. KNOWLES†

**Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70808, USA,*

†Department of Ecology & Evolutionary Biology, Museum of Zoology, University of Michigan, 1109 Geddes Ave., Ann Arbor, MI 48109-1079, USA

Despite the widespread use and obvious strengths of model-based methods for phylogeographic study, a persistent concern for such analyses is related to the definition of the model itself. The study by Peter *et al.* (2010) in this issue of *Molecular Ecology* demonstrates an approach for overcoming such hurdles. The authors were motivated by a deceptively simple goal; they sought to infer whether a population has remained at a low and stable size or has undergone a decline, and certainly there is no shortage of software packages for such a task (e.g., see list of programs in Excoffier & Heckel 2006). However, each of these software packages makes basic assumptions about the underling population (e.g., is the population subdivided or panmictic); these assumptions are explicit to any model-based approach but can bias parameter estimates and produce misleading inferences if the model does not approximate the actual demographic history in a reasonable manner. Rather than guessing which model might be best for analyzing the data (microsatellite data from samples of chimpanzees), Peter *et al.* (2010) quantify the relative fit of competing models for estimating the population genetic parameters of interest. Complemented by a revealing simulation study, the authors highlight the peril inherent to model-based inferences that lack a statistical evaluation of the fit of a model to the data, while also demonstrating an approach for model selection with broad applicability to phylogeographic analysis.

Received 16 August 2010; revision received 26 August 2010; accepted 1 September 2010

Genetic data have never been easier to collect, either using Sanger sequencing or next-generation methods. As we enter this new genomic era, it is worthwhile to reflect on how we will use these data to make inferences regarding the biology of the focal taxa. Evolutionary genetics aims to identify important demographic processes using genetic data; these may be historical (phylogeography) or contemporary (landscape genetics), but in either case, the process begins with the data collection and results in one or more inferences concerning the demography of the organism. The field has devoted a great deal of attention to the development of powerful methods for parameter estimation, and these estimates often play a key role in the inference process. However, their accuracy depends on how well the empirical data fit the assumptions of the method, and to date, the process of model choice has received relatively little attention (Buckley 2008; Carstens *et al.* 2009; Knowles 2009).

Peter *et al.* (2010) provide a clear demonstration of the importance of assessing the fit of the demographic model to the data. They explore two approaches to Bayesian model choice procedures and demonstrate that both are accurate and unbiased. The simplest of these calculates the relative posterior probabilities of the competing models and is capable of identifying the best model across a wide range of sampling schemes. Intriguingly, while precise population genetic parameter estimation using Approximate Bayesian Computation (ABC) requires a large number of simulations spanning the prior distribution, their results show that it is possible to distinguish between two models without precise parameter estimates. While the generality of this finding is unknown at this time, it provides some solace to what could be construed as a daunting computational challenge. Some historical scenarios may require more data and/or additional simulations across the range of parameters specified by the prior to be distinguished statistically, and it is also possible that some scenarios may not be distinguishable. The power of the procedures for model selection described by Peter *et al.* (2010) will have to be investigated with additional detailed study on a case-by-case basis, especially in the context of analyses based on ABC (see Bertorelle *et al.* 2010).

The study also shows how model choice can dramatically effect parameter estimates and hence conclusions about the processes underlying the patterns of genetic variation. With detailed simulations, Peter *et al.* (2010) confirm that estimates of genetic diversity ($\theta = 4N_e\mu$) derived under a single, panmictic population model that allows size change can contain genetic signatures of either expansion or contraction when the data are drawn from structured populations, even if the data have evolved under history in which there was no change in population size. Inconsistencies of this sort put empiricists in a bind—the actual history of an empirical system is unknown, and

Correspondence: Bryan C. Carstens;
E-mail: carstens@lsu.edu

analyses may lead to misleading inferences when datasets are analyzed under an inappropriate model. The procedure described in Peter *et al*. (as discussed above) provides one solution to what otherwise is a primary detractor of model-based phylogeographic inference—the need for statistical procedures for model selection.

Peter *et al*. (2010) are not the first to evaluate alternative demographic models using ABC (e.g., Fagundes *et al*. 2007) or to show that model choice can have a dramatic effect on a study's conclusions, such as inferring population expansion or decline when there was none (e.g., Becquet & Przeworski 2009; Nielsen & Beaumont 2009). Their study is nonetheless notable in two specific regards. First, it demonstrates empirically the impact of model choices that had previously been limited to the theoretical realm. Namely, their work highlights the potentially grave conservation consequences if management decisions were based on conclusions about a panmictic declining population, as opposed to a stable, small subdivided population. Second, they show how to avoid being mislead by poor judgments about model choice. The flexibility of the ABC framework (Bertorelle *et al*. 2010) suggests that the model choice procedure described by Peter *et al*. (2010) could, in principle, be expanded beyond the two models considered in the study to calculating the relative posterior probability of models representing any number of demographic scenarios.

References

Becquet C, Przeworski M (2009) Learning about modes of speciation from computational approaches. *Evolution*, **63**, 2547–2562.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, **18**, 2609–2625.

Buckley D (2008) Towards an organismal, integrative, and iterative phylogeography. *Bioessays*, **31**, 784–793.

Carstens BC, Reid N, Stoute HN (2009) An information theoretical approach to phylogeography. *Molecular Ecology*, **18**, 4270–4282.

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, **7**, 745–758.

Fagundes N, Ray DA, Beaumont MA *et al*. (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the USA*, **104**, 17614–17619.

Knowles LL (2009) Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612.

Nielsen R, Beaumont MA (2009) Statistical inferences in phylogeography. *Molecular Ecology*, **18**, 1034–1047.

Peter BM, Wegman D, Excoffier L (2010) Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, **19**, 4648–4660.

doi: 10.1111/j.1365-294X.2010.04851.x