Systematic Biology Advance Access published July 16, 2009

Point of View

DOI:10.1093/sysbio/syp038

Sampling Error Does Not Invalidate the Yule-Coalescent Model for Species Delimitation. A Response to Lohse (2009)

ANNA PAPADOPOULOU[1,2], MICHAEL T. MONAGHAN[3], TIMOTHY G. BARRACLOUGH[2], AND ALFRIED P. VOGLER[1,2,*]

[1]*Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK;*
[2]*Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK; and*
[3]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Mueggelseedamm 301, 12587 Berlin, Germany;*
**Correspondence to be sent to: Department of Entomology, Natural History Museum, Cromwell Road, London SW7 5BD, UK; E-mail: apv@nhm.ac.uk.*

Lohse (2009) used a simulation study to argue that the generalized mixed Yule-coalescent (GMYC) model (Pons et al. 2006; Fontaneto et al. 2007) may overestimate species numbers. He found that incomplete sampling of demes involved in the coalescence process could artificially produce clusters that are recognized as separate GMYC groups (species). The paper also criticizes our (Papadopoulou et al. 2008) simulations of the coalescent process where we found that GMYC groups are readily formed when migration among demes drops below a particular level ($Nm < 0.01$). We interpreted these results to indicate that divergent sequence clusters form under conditions of stringent population isolation and that these clusters resemble those widely seen in empirical data from mitochondrial DNA (mtDNA) sampled across multiple populations and species. Lohse's (2009) simulations confirmed our findings but warned that additional GMYC groups are recognizable when less than about 20% of all demes are sampled. Although this is a valid point, Lohse (2009) extrapolated these findings to dismiss the utility of DNA-based approaches to species delimitation, saying that real-world samples will be composed of "essentially random clusters." We argue that these conclusions go well beyond the simulation results from partially isolated populations within a single species and that the dismissal of the GMYC model as the basis for delineating entities in taxonomic research is unjustified.

SAMPLING

Any method of species delineation is sensitive to sampling, regardless of the characters and criteria used. It is therefore not surprising that the GMYC procedure is affected by sampling. Existing quantitative methods of species delimitation that use "diagnostic" (unique to a species) characters specifically raise the issue of sampling (Davis and Nixon 1992). If within-population variation is undersampled, the subsets of haplotypes drawn from each subpopulation are more likely to show diagnostic character variation by chance. Thus, if populations (demes) with intermediate haplotype composition are left unsampled, this results in an overestimate of species numbers ("oversplitting"). In contrast, undersampling of populations from a limited distributional range may miss differentiated populations and hence underestimate the number of diagnosable groups (species). These effects of incomplete sampling are exacerbated in cases where genetic variation is substructured within a species. Whereas coalescence approaches incorporate the sampling in a more explicit way than conventional taxonomic methods (e.g., Wakeley 2000), species delimitation achieved under any approach can only ever be a preliminary hypothesis subject to further testing (Lipscomb et al. 2003). Thus, Lohse's (2009) study is a valuable assessment of the statistical properties of the GMYC method that warns against drawing conclusions when a small proportion of demes are sampled. However, it does not in any way overturn our conclusions about the effect of migration rate on the formation of GMYC groups and the congruence of these groups with the structuring of demes in geographical space (Papadopoulou et al. 2008). Because geographical coherence of demes may or may not exist in real data (Avise 2009), we argue that empirical studies are necessary to evaluate the efficacy of the model.

The empirical examples cited (Pons et al. 2006; Papadopoulou et al. 2008) are, unlike the simulations of Lohse (2009), composed of several separate species within which there might be partially isolated populations. How the GMYC method delimits clusters from empirical data depends on the typical degree of interspecific divergence as well as on intraspecific variation (which depend on multiple parameters such as speciation and extinction rates, the degree of isolation between partially isolated demes, the effective population size, and the migration rate). Through the effect simulated by Lohse (2009), the undersampling of certain populations could affect how distinct the species appear and lead to the appearance of additional clustering. However, under

a wide range of circumstances, interspecific branch lengths should still be generally longer than (and statistically distinguishable from) intraspecific branches.

The magnitude of potential error from the GMYC analysis for determining evolutionary entities in cases of low sampling density has to be assessed against external evidence, such as known biogeographic boundaries, morphological differences, or, indeed, unlinked gene markers (where the outcome of sampling error is different if demes are not geographically structured).

The findings of Lohse (2009) do not profoundly affect the conclusions of Pons et al. (2006) in the genus *Rivacindela*, which included representatives from numerous sympatric species and collections at all known sites (isolated salt flat habitats) across Western Australia. Although this sampling regime may still have left out a number of unknown sites, it is encouraging that the GMYC groups represented at multiple locations had contiguous ranges, indicating that sampling of demes in these groups was sufficient. Likewise, Papadopoulou et al. (2008) sampled *Eutagenia smyrnensis* s.l. intensively in 18 of the major Central Aegean Islands, corresponding to a total area of 2799 km^2, out of the 6064 km^2, composed of the Cyclades, Dodecanese islands, and the Ikaria-Samos complex. Thus, we have sampled extensively 46% of the central Aegean area, and the GMYC estimates should be little affected by the sampling density. As in *Rivacindela*, individual GMYC groups of *Eutagenia* showed high geographic coherence that largely matched plausible biogeographical boundaries. In addition, the single-copy nuclear marker Mp20 strongly supports the GMYC groups in *Eutagenia*, although with lower resolution as expected due to the 4x larger N_e compared with the mitochondrial genome (Papadopoulou et al. 2009). Further studies of beetles also found corroboration of GMYC groupings from unlinked nuclear loci (Monaghan et al. 2005; Ahrens et al. 2007).

The problems raised by Lohse result from the fact that offspring is produced in the vicinity of the parents, which in turn produces greater similarity of genotypes at a site compared with other sites. Undersampling therefore may cause genetic divisions being recognized where none actually exist. This effect results in difficulties in particular when distinguishing between highly structured samples (due to low migration rate) and samples from undersampled populations with intermediate migration rates (leading to random clustering). In the absence of true genetic structure, incomplete sampling would draw from a uniform pool of variation, and groups arising by the luck of the draw may be obtained anywhere across the geographical range. Even under mild genetic structuring (e.g., isolation by distance), these artifactual divisions from undersampling might arise anywhere along a continuum of variation, not primarily along known historical boundaries. The geographic distributions of the GMYC clusters in our empirical studies thus increase the confidence in the conclusions from this analysis.

More Complex Models

Lohse (2009) also recommends the implementation of more elaborate models of the coalescence process that distinguish between a scattering and a collecting phase (Wakeley 1998). More complex versions of the GMYC model could be specified that implement two classes of coalescent branches (corresponding to scattering and collection phase) in addition to the speciation branches. We agree that this might achieve a more realistic model of the assumed process and might result in improved separation of inter- and intraspecific branching rate when geographic structure is common. However, Lohse apparently used simulations very similar to those of Papadopoulou et al. (2008) and showed that the sampling issue has only a minor effect on the conclusion about the dispersal parameter itself. Therefore, the reported sampling artifact does not affect the conclusion of Papadopoulou et al. (2008): that the formation of GMYC groups is consistent with the expectations regarding migration, showing support for the GMYC model under some parameter values, but not others. Further improvements can be achieved with a recently modified GMYC model that allows for a variable transition point from coalescent to speciation across a phylogenetic tree (Monaghan et al. 2009), which is more robust when intra- and interspecies branches vary in length across the tree and hence might discriminate these types even under conditions of undersampling.

Conclusions

Although the problem of sampling raised by Lohse (2009) is a valid contribution to the test of the method, it does not diminish the possibility for DNA-based species delimitation in general. Ultimately, the applicability of all species delimitation methods needs to be assessed against empirical data of various levels of completeness and against simulations of realistic scenarios. The GMYC model provides an initial hypothesis of the number and extent of species-level groups, given a particular sampling regime and an assumed process of lineage branching. There is a possibility that these groups simply represent geographically isolated populations evolving neutrally (Fontaneto et al. 2007). However, a wide range of empirical data support the conclusion of Pons et al. (2006) that the model can effectively capture the transition from speciation to coalescent processes as GMYC groups closely correspond to species defined by independent criteria (unlinked gene loci, morphological characters, or accepted Linnean names) and track known biogeographic boundaries. The two portions of the GMYC model relate loosely to Hennig's (1966) widely accepted distinction of "tokogeny", referring to parent–offspring relationships as they exist in interbreeding populations, and "phylogeny", referring to divergent ancestor–descendant relationships. Finally, we agree with Lohse (2009) that multilocus data have greater statistical power to resolve relationships among closely related lineages. Nonetheless, mtDNA gene trees

alone provide good approximations of population histories (Avise 2009). This will remain advantageous until multilocus approaches become feasible for the broad surveys of entire clades or faunas that single-locus approaches have only recently made possible.

REFERENCES

Ahrens D., Monaghan M.T., Vogler A.P. 2007. DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). Mol. Phylogenet. Evol. 44: 436–449.

Avise J.C. 2009. Phylogeography: retrospect and prospect. J. Biogeogr. 36:3–15.

Davis J.I., Nixon K.C. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. Syst. Biol. 41:421–435.

Fontaneto D., Herniou E.A., Boschetti C., Caprioli M., Melone G., Ricci C., Barraclough T.G. 2007. Independently evolving species in asexual bdelloid rotifers. PLoS Biol. 5:914–921.

Hennig W. 1966. Phylogenetic systematics. Urbana (IL): University of Illinois Press. p. 280.

Lipscomb D., Platnick N., Wheeler Q. 2003. The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol. Evol. 18: 65–68.

Lohse K. Forthcoming 2009. Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006). Syst. Biol. doi: 10.1093/sysbio/syp039.

Monaghan M.T., Balke M., Gregory T.R., Vogler A.P. 2005. DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Phil. Trans. Roy. Soc. B 360:1925–1933.

Monaghan M.T., Wild R., Elliot M., Fujisawa T., Balke M., Inward D.J.G., Lees D.C., Ranaivosolo R., Eggleton P., Barraclough T.G., Vogler A.P. Forthcoming 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst. Biol. doi: 10.1093/sysbio/syp027.

Papadopoulou A., Anastasiou I., Keskin B., Vogler A.P. 2009. Comparative phylogeography of tenebrionid beetles in the Aegean archipelago: the effect of dispersal ability and habitat preference. Mol. Ecol. 18:2503–2517.

Papadopoulou A., Bergsten J., Fujisawa T., Monaghan M.T., Barraclough T.G., Vogler A.P. 2008. Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. Phil. Trans. Roy. Soc. B 363:2987–2996.

Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.

Wakeley J. 1998. Segregating sites in Wright's island model. Theor. Popul. Biol. 53:166–174.

Wakeley J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution. 54:1092–1101.

Received 23 February 2009; reviews returned 30 March 2009; accepted 9 June 2009
Associate Editor: Marshal Hedin