SUPPLEMENTARY METHODS FOR

PHRAPL: Phylogeographic Inference Using Approximate Likelihoods

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

[1]*Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, 442 Hesler Biology Building, Knoxville, TN, 37996*

[2]*Department of Evolution, Ecology and Organismal Biology, Ohio State University, 318 W. 12th Avenue, Columbus, OH, 43210*

COMPARING THE APPROXIMATE LIKELIHOOD WITH ANALYTICAL LIKELIHOOD

When analyzing isolation-only histories using PHRAPL, for each replicate and treatment, we calculated approximate lnLs under the generating model, setting nTrees = 100,000 and applying the following grid values for coalescence time: 0.05, 0.13, 0.32, 0.80, 2.01, 5.05, 12.71, 32. We calculated analytical lnLs under the generating specie tree using COAL (Degnan and Salter 2005). For select treatments, we also calculated approximate lnLs and analytical lnLs under a species tree in which labels for populations B and C were transposed. This allowed us to assess the degree to which error in approximate lnL estimates could be expected to reduce accuracy in model selection.

TESTING THE APPLICATION OF APPROXIMATE LIKELIHOODS TO COMPLEX MODELS THAT INCLUDE MIGRATION

To simulate the stochastic mutational process underlying empirical datasets, for all genealogies simulated for the purpose of model selection, we evolved sequences along the branches using Seq-Gen (Rambaut and Grassly 1997), assuming a HKY mutation model, 500 base pairs per locus, base pair frequencies = 0.3, 0.2, 0.2, and 0.3 (for A, C, G, and T), transition/transversion ratio = 3, and θ = 0.005 (e.g., Ence and Carstens 2011; Knowles and Carstens 2007). We then inferred gene trees from sequence data using RAxML 7.2.6 with five replicate searches, rapid hill-climbing, and the GTRGAMMA model (Stamatakis 2006). We used these inferred trees as input for all analyses.

When analyzing each dataset and model using PHRAPL, we used nTrees = 10,000 and parameter grids consisting of seven coalescence time values (t = 0.30, 0.58, 1.11, 2.12, 4.07, 7.81, and 15.00) and six migration rate values (M = 0.10, 0.22, 0.46, 1.00, 2.15, 4.64). We analyzed 10 iterative subsamples of each dataset, subsampling 4 individuals per population (resulting in 12 tip trees).

REFERENCES

Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. Evolution 59(1):24-37

Ence DD, Carstens BC (2011) SpedeSTEM: a rapid and accurate method for species delimitation. Molecular Ecology Resources 11(3):473-480

Knowles LL, Carstens BC (2007) Estimating a geographically explicit model of population divergence. Evolution 61(3):477-493

Rambaut A, Grassly NC (1997) An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, version 1.2.5. Computer Applications in the Biosciences 13:235-238

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21):2688-2690