13

SPEDE-sampler: An R Shiny application to assess how methodological choices and taxon-sampling may affect DNA-based Generalised Mixed Yule Coalescent (GMYC) species delimitation

Species delimitation tools are vital to taxonomy and the discovery of new species. These tools can make use of genetic data to estimate species boundaries, where one of the most widely-used methods is the Generalized Mixed Yule Coalescent (GMYC) model. Performance estimates of the GMYC are predominantly based on results from simulated datasets, where assumptions about the underlying statistical properties are not violated. However, GMYC performance on real-world data where model and data assumptions are likely violated remains largely unknown. In this paper, we present “SPEDE-sampler", a user-friendly R Shiny application that assesses the effect of computational and methodological choices, in combination with sampling effects, on the GMYC model using empirical datasets. The application can be used across disciplines and taxonomic groups, and its usage does not require previous coding experience due to its interactive graphical user interface. SPEDE-sampler randomly resamples a desired percentage of DNA sequences in a multiple sequence alignment file, optionally guided by user-defined groups (e.g. morphospecies or ecospecies), and produces input files for BEAST analyses. The resulting BEAST phylogenies are then used to assess the effect that (1) sample size and geographic sampling scale, (2) BEAST and GMYC parameters (e.g. prior settings and rate distributions, clock model, GMYC single vs multiple threshold approach), and (3) singletons has on GMYC output. The optional user-defined groups can be compared to GMYC species estimates to calculate percentage match scores between traditional morphological taxonomy and DNA-based taxonomic methods. Additionally, predefined groups that contribute to inflated species richness estimates are identified by SPEDE-sampler, allowing for further investigation of potential cryptic species or population structure within those groups. The application allows the user to download all generated data files, as well as associated customisable graphics in a variety of image formats. The application is open-source, and is available for download on GitHub (https://github.com/clarkevansteenderen/spede_sampler_R) with installation instructions and a fully-worked example.