The timing of speciation events in a phylogenetic tree provides insight into key evolutionary processes like speciation and extinction. Was there a mass extinction (and when)? Are rates of speciation correlated with a particular biological trait (a key innovation) or geographic area (a cradle of diversity)? Does the rate of speciation in a clade depend on the number of species? Such questions concern a process called lineage diversification. To determine if nectar spurs increase rates of speciation, an evolutionary biologist asks if the observed phylogeny can be explained by a null model, or if the null can be rejected in favor of an alternative that allows speciation rates to be higher in lineages that have nectar spurs. The results of such a comparison are only as good as the null model; if it is overly simplistic, it will always be rejected, even if the alternative models are incorrect or nonsensical.

In studies of lineage diversification, the null model is a fully-sampled, constant-rate, birthdeath process. That is, the null model describes a process where there is a single rate of speciation and a single rate of extinction for all species throughout the history of the studied clade, and assumes that every extant descendant of the ancestral species is sampled in the phylogeny. Biologically, there are reasons to be suspicious that this model does not adequately describe diversification. For example, mass extinction events and diversity-dependence cause temporal heterogeneity in speciation and extinction. Key innovations can cause variation among species in rates of diversification. Indeed, it has been shown that diversification both varies through time (Morlon et al. 2010) and across the lineages in a phylogeny (Heard and Cox 2006). Further, most phylogenies do not include every descendant of the ancestral species, and the choice of species included in the phylogeny can bias inferences of diversification (Höhna 2014). It is no surprise that Rabosky and Goldberg (2015) demonstrated that the constant-rate birthdeath process is a completely inadequate null model. Using simulation, they show that arbitrary biological traits frequently, and falsely, appear correlated with rates of diversification. The traitdependent model is not correct (the trait has no effect on diversification), but it allows diversification rates to vary, while the null model does not. Clearly, new null models of lineage diversification must be developed.

Models account for diversification-rate variation in one of two ways: either a whole tree obeys a single, time-dependent process (a whole-tree approach), or individual lineages have unique, usually constant, rates of diversification (a lineage-specific approach). Given the apparent prevalence of among-lineage rate variation, a whole-tree approach would be inadequate for a null model of lineage diversification. The most promising direction for developing new null models is to account for incomplete sampling of species and among-lineage rate variation. A good place to start developing such models is to examine relaxed clock models for nucleotide sequence evolution. These models are based on the idea that molecular evolution occurs at a consistent rate, but they allow for that rate to vary across the lineages in a phylogeny. A probability distribution is chosen that describes how rates of substitution vary within a phylogeny, that is, it describes how to relax the evolutionary clock. These relaxed-clock models are flexible largely due to the ability to choose from a wide range of probability distributions, allowing rates to cluster tightly or to be highly variable. But there is also flexibility stemming from the ability to choose whether variation is uncorrelated among all lineages, or autocorrelated such that the descendants of a lineage have substitution rates that are more like that of their ancestor. Similarly, I seek to allow the rate of lineage diversification to vary among lineages. I

will develop a flexible framework for among-lineage diversification-rate variation that allows me to compare uncorrelated and autocorrelated models using a variety of clock distributions, which will allow for empirically guided choices of distributions. I will fit relaxed-clock lineage diversification models, using a wide range of clock distributions, to a large set of empirical phylogenies. From this, I will be able to determine which clock distributions are fit for further development, and which I can safely ignore as unrealistic.

To determine which models, if any, are suitable replacements for the constant-rate birthdeath process, I will perform a simulation study. A good null model should be rejected frequently when it is not the true model—that is, when it is not the model that was used for simulation—but should not be rejected when it is the true model. This will be the most computationally intensive, and expensive, portion of the project. I will need to cover a range of biologically plausible values of model parameters, i.e. rates of speciation and extinction, the sampling fraction, and the parameters of the relaxed-clock distribution. I will specify a range of discrete values to test for each parameter, and the study will comprise every combination of parameter values in this "grid." While each cell in the grid is relatively quick, in total I will need hundreds of thousands of CPU hours. To ensure that I can run the simulations no matter where I obtain computing resources, I will implement the models in a user-friendly R package. This will also allow the models to be used by researchers of all backgrounds, anywhere in the world. When I have identified a set of statistically viable relaxed-clock birth-death models, I will be able to ask many interesting, novel questions. Do diversification rates tend to vary according to any particular relaxed-clock model or models? How does among-lineage rate variation bias whole-tree inferences of diversification? I will also be able to address current questions with more statistical power and confidence. Which biological traits are truly correlated with diversification rates? How frequently is diversification actually diversity-dependent?

BROADER IMPACTS

To complete the simulation study, I will partner with a local high school. My experience mentoring a high school student taught me that students in high school are not aware that biology can seamlessly integrate statistics and computers. I will use my simulation study as a teachable moment. As they perform the simulations (using my R package), I will explain what they are simulating. This will require me to provide an introduction to branching processes and statistical parameter inference. Impressing upon them the value of simulation studies will help showcase the importance of computational work to modern biology, and will teach them the logic required to evaluate statistical models. When all the simulations have been completed, I will show how that leads to a better understanding of the models. Then I will use those models to test an example phylogeny for trait-dependent speciation, mass extinctions, and diversity-dependence, showing them the biological relevance of the project.

References

Heard, Stephen B., and Graham H. Cox. "The shapes of phylogenetic trees of clades, faunas, and local assemblages: exploring spatial pattern in differential diversification." The American Naturalist 169.5 (2007): E107-E118. Höhna, Sebastian. "Likelihood inference of non-constant diversification rates with incomplete taxon sampling." PLoS one 9.1 (2014).

Morlon, Hélène, Matthew D. Potts, and Joshua B. Plotkin. "Inferring the dynamics of diversification: a coalescent approach." PLoS Biology 8.9 (2010): e1000493.

Rabosky, Daniel L., and Emma E. Goldberg. "Model inadequacy and mistaken inferences of trait-dependent speciation." Systematic Biology 64.2 (2015): 340-355.