Project Summary

Integrating information from the fossil record with data collected from extant taxa to infer phylogeny is a prime challenge in biology. Fossils are often our only direct observation of past biodiversity, and for the inference of dated phylogenetic trees, may be the only source of information that can be used to establish an evolutionary time scale. Newer methods for inferring dated phylogenetic trees, such as the fossilized birth-death model (FBD) model the extant and extinct data together as part of the same process of diversification. These methods typically implemented as hierarchical Bayesian models involving a model molecular and morphological character evolution, a model describing how rates of evolution are distributed across the tree, and a model of how diversification has proceeded in the focal taxa. I will explore the use of posterior predictive methods for assessing which models are most appropriate for a particular dataset. This work will provide practical guidance and research software tools for researchers to perform more complex model assessment in systematic biology.

The educational aims of this project are intended to illuminate how to incorporate computation into biology education, while improving the retention of diverse students. In this project, I will study how faculty can incorporate computation in a code-to-learn framework, in which biological information is discovered via hands-on computational and data analytic exercises. I intend to formally study if teaching lower division coursework using code-to-learn principles causes students to learn less biology. I also propose to study if early exposure to computation in the classroom can lead to improved student retention by helping students develop important research skills early in their careers.

Intellectual Merit

In the big data era, it is often possible to tease out any desired result from a dataset, depending on the assumptions used to model the data. This is true of the phylogeny of many clades. In particular, in the ants [Formicidae], a host of different phylogenies can be obtained from a joint molecular-morphological dataset depending on the assumptions made about evolution in the group. The work in this proposal will provide tools and recommendations for researchers to find models that are the best fit for their data. The results will also provide methods for researchers to figure out which assumptions of a given model may be especially bad for their data, which will provide helpful information for future methods development. My educational aims will also address key gaps in the literature on student retention and curriculum design. In particular, I will use paired-design studies to compare skills gain in students who have early exposure to computation in their studies and those who do not.

Broader Impacts

Retention of underrepresented minority students (URM) is at the heart of my educational plan. As faculty at a public institution in a poor state, I observe very clearly the issues with retention of vulnerable students. I propose to leverage existing recruitment networks for vulnerable students to identify students who are interested in scientific computing. These students a research stipend year-round to help them remain in school and do productive research with a faculty mentor. I will also study the retention of URM students as a function if they receive early exposure to important research skills, such as computational data analysis.

Results from Prior NSF Support

Co-PI Wright: PI on NSF Award 1612858, $136,000, July 1, 2016 - July 1, 2017. Postdoctoral Research Fellowship in Biology. Intellectual Merit: The project explored the use of morphological data in divergence time estimations. This included both the production of scientific software for phylogenetic estimation and theoretical work to understand the role of incomplete lineage sampling in fossil-based phylogenetic work. Publications produced: Barido-Sottani J, Justison J, Wright AM, Warnock RCM, Pett WC, Heath TA (2020), Wright AM (2019), Lanfear, R, Wright AM, Fransden PB, Senfeld T, Calcott B. (2017), Matzke NJ, Wright AM (2016), Bapst DW, Wright AM Lloyd GT, Matzke NJ (2016). One additional manuscript is accepted. Talks on this work were also presented at the Evolution Meetings in 2017, the Systematic Biology meetings in 2017, and the Entomological Society Meetings in 2018.

1 Introduction

The term 'big data' is often used to refer to the volume of data brought to bear on an analytical problem. However, the big data era poses challenges in *integrating* various data sources to answer questions. One field in which this is true is phylogenetics, and particularly the estimation of dated phylogenetic trees. Dated phylogenetic trees have branch lengths in absolute time, such as years or millions of years (Zuckerkandl and Pauling, 1965), as opposed to relative rates of evolution. In inferring dated phylogenetic trees, researchers typically include molecular data (amino acid or nucleotide), morphological character data, especially from fossils, and fossil occurrence time data (Heath et al., 2014; Warnock and Wright, 2020). Each of these data sources has its own sources of error and uncertainty (Barido-Sottani et al., 2019). I propose several research and education aims for estimating joint molecular-morphological timetrees, evaluating model adequacy for working with large and varied datasets, and training a diverse scientific workforce able to do quantitative work with complex biological data.

Without occurrence time information, either in the form of fossils or other information, a phylogenetic tree cannot be scaled to absolute time without making extremely strong assumptions about evolutionary rate (Harvey et al., 1994; Zuckerkandl and Pauling, 1962). Historically, fossil data have been incorporated through what are termed node-calibration methods (Heath et al., 2014). Fossils aren't truly data under these methods. Instead, fossils are used to constrain (or *calibrate*) the potential ages of nodes. For example, a fossil that is known to be a member of a clade can be used to tell how old that clade must minimally be. If the fossil exists, so must have the clade to which it belongs. However, the researcher must fix that fossil to be part of a specific clade, a placement which is assumed to be known without error (Donoghue and Benton, 2007). Additionally, only the oldest fossil in a clade may be used (Yang and Rannala, 2006).

Recent methods that more completely model the processes that generate fossil data have provided new ways to incorporate fossils. Fossilized birth-death (FBD) methods (Stadler, 2010; Heath et al., 2014) model the process of speciation, extinction, and sampling that leads to what is termed the 'observed tree.' The observed tree is the portion of the true tree (the true tree is unobservable due to patchy sampling in both the present

and past) that we can estimate from our data. FBD can be used to estimate the phylogeny, node ages on that phylogeny, and other macroevolutionary parameters given a set of character data (molecular and/or morphological) and occurrence times (the ages of fossils). The way fossils are treated in FBD models is vastly different than under calibration methods. Under the FBD, fossils can be placed via morphological character information, or their placement may be integrated out analytically if no character data exist (Gavryushkina et al., 2017). Multiple fossils can be used per clade, as well (Matzke and Wright, 2016; Warnock et al., 2020). Due to these differences, the FBD framework may be expected to give more robust estimates of topology and divergence times. However, FBD increases the complexity of the overall phylogenetic model. The FBD is often implemented in a tripartite framework, in which there is a model of character evolution (molecular of morphological character change), a clock model (describing the distribution of evolutionary rates over the tree) and the FBD model, describing the process of diversification and sampling leading to the tree. This tripartite model is typically implemented as a Bayesian model, in which samples of plausible parameters will be estimated for each parameter (Warnock and Wright, 2020).

The character change model describes how likely to change from one state to another a character is. Molecular character change models are fairly well-developed. As early as the late 1960s researchers began to catalog the biochemical properties of amino acids and nucleotides (Eck and Dayhoff, 1966). The first molecular sequence models (Jukes and Cantor, 1969) made fairly restrictive assumptions: that every nucleotide state is equally likely to transition to any other nucleotide state, and that character change is instantaneous along a branch. Bolstered by carefully observation of molecular evolution more complex models that describe realistic evolutionary scenarios are now available (Kimura, 1980; Felsenstein, 1981; Tavaré, 1986; Lartillot, 2006). Morphological character change models are considerably underdeveloped relative to molecular models and are roughly equivalent to the earliest nucleotide models (Lewis, 2001). Most molecular sequence models assume a nucleotide will have the same properties wherever it occurs in a sequence. This allows for a model to be applied across the entirety of a sequence. This is not true for morphological data. In discrete character data, '0' often refers to the absence of a character. That absence may be due to the character never having evolved in the clade or due to loss of that character. Characters also vary widely in their complexity, and the number of genetic changes required to generate that character state. When including morphological and molecular data together in one analysis, there is an asymmetry in how well we can model the two types of evolution (Wright et al., 2016; Wright, 2019).

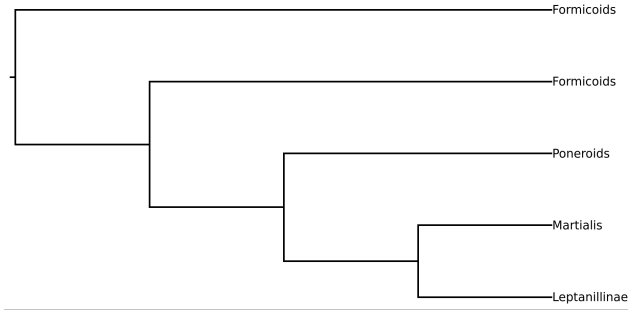
In order to make full use of the tripartite model, expanded toolkits are needed to understand the adequacy of each component of the components of the hierarchical model. In the ant group (Formicidae), using a tripartite model can yield any of the topologies and overall clade ages obtained by previous methods, depending on the precise model assumptions made (see Preliminary Work). **In this work, I would like to apply and further develop a set of methods referred to as posterior predictive model adequacy testing (Lewis et al., 2014; Brown, 2014; Slater and Pennell, 2014; Duchêne et al., 2015; Hoehna et al., 2017) to both solve an empirical challenge, the phylogeny of ants, and to more broadly assess how well FBD methods are performing in empirical data.**

This work has implications far beyond phylogenetics. In nearly every field where large datasets and multiple lines of evidence have been brought to bear, the difficulty of adequate modeling looms large. I work at a primarily undergraduate institution, and have been actively developing coursework involving quantitative and statistical thinking in biology students. This learning takes place in many of my courses, from introductory biology, to genetics, to upper division courses. My educational components for this project focus on the evaluation of the coursework I have developed. **I am interested in evaluating if early involvement in computational learning and research helps to retain students from vulnerable populations in STEM research. I am also proposing to investigate if computation can be integrated in lower-level biology coursework without losing domain knowledge.**

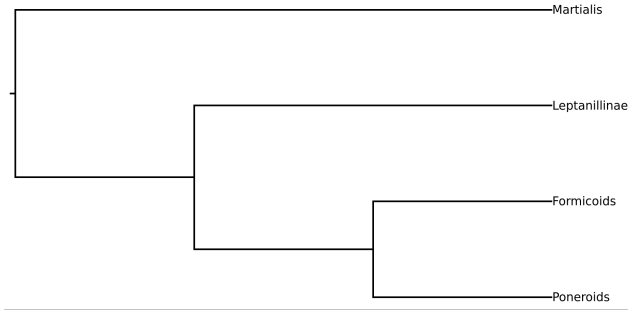
2 Preliminary Work

My preliminary work on this question can be split into three components: data collection, theoretical and software tools, and education.

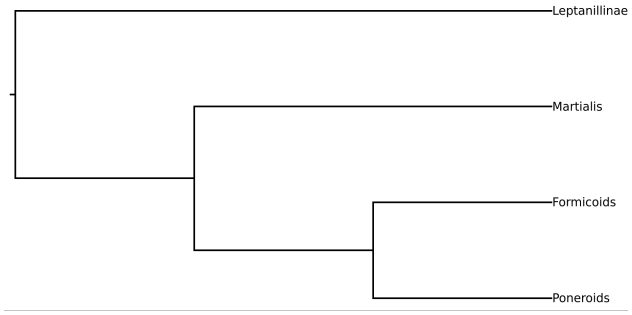
I am using previously published data for this research. The DNA data come from Blanchard and Moreau (2017), and comprise 9 nuclear and mitochondrial markers. Because errors in homology assignment have been shown to impact downstream phylogenetic estimation, I worked with an undergraduate assistant to use the software PASTA (Mirarab et al., 2015) to improve the DNA sequence alignments, yielding higher-quality alignments. Additionally, I have assembled a large morphological matrix from several sources (Keller, 2011; Barden and Grimaldi, 2016). I have also worked with two undergraduate as-



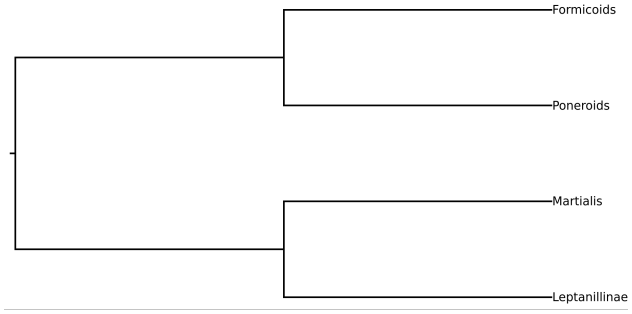
(a) Topology estimate from morphology



(b) Rabeling et al. 2008 molecular phylogeny



(c) Kuck et al. 2011 molecular phylogeny



(d) Boroweic et al 2019 molecular phylogeny

Figure 1: Molecular trees show four monophyletic clades of ants, while those estimated from morphology often do not. Preliminary analyses using the FBD with a joint molecular-morphological dataset uncovered support for all four trees, depending on assumptions made about diversification.

sistants to develop a pipeline for accessing and reproducibly cleaning fossil occurrence time data from the Paleobiology Database.

I have also made strides on the application of the FBD model to these data, as well as more broadly. As shown in Figure 1, I, along with a team of 6 undergraduate researchers, have estimated several dated phylogenetic trees using different parameterizations of the FBD and its component models. As can be seen in the figure, these have resulted in different topologies and ages. In order to better visualize uncertainty in topological estimations, I have also been applying treespace visualization tools to morphological datasets (Wright and Lloyd, 2020). This analysis has been underused, historically, but allows researchers to visually explore the distribution of solutions in their datasets. In the initial application of this technique, I used treespace visualizations (Hillis et al., 2005; Warren et al., 2016) to look at how well various solutions in the Bayesian posterior sample fit fossil record data independent to those used to estimate the tree. These initial explorations showed that this can be a promising method to understand model performance. I intend to strengthen these approaches in this proposal.

3 Proposed Work

Despite improving methods and more data, many nodes in the tree of life remain stubbornly unresolved, with different resolutions differing between trees. Some conflicts have biological causes, such as incomplete lineage sorting (Maddison and Knowles, 2006). Other sources of error can include mismatch of the phylogenetic model to the data used to generate the tree (Lemmon and Moriarty, 2004; Wright et al., 2016). Additionally, there can be conflict between data types. For example, in the Formicidae, the molecular and morphological trees are quite different (see Fig. 1; Keller (2011); Rabeling et al. (2008); Kück et al. (2011); Barden and Grimaldi (2016); Borowiec et al. (2019)). Using a joint molecular-morphological dataset, I will explore the use of new model-fitting techniques to assess which models are best capturing the generating processes that underlie the data.

This project will have three major components. The first aim is to develop biological hypotheses for how the parameters of the fossilized birth-death process vary with time, and to assess the fit of these models to the data using stepping-stone model selection (Xie et al., 2011). Stepping-stone model selection allows researchers to discriminate among candidate models. It cannot tell you if a model is capturing the process that generated the data. Therefore, the second objective will be to use posterior predictive model testing to determine if the chosen model is adequate (**?**Hoehna et al., 2017) for the data, and to detect in what aspects of the tree we may see artifactual results due to inadequacy. Lastly, I will expand on software toolkits for performing statistical model testing and assessment. Existing tools to perform these analyses involve estimating phylogenetic trees in one piece of software, and performing simulation in another. I've made preliminary steps towards interoperability of these tools, and I will expand on this progress to make these complex analyses more accessible to researchers.

3.1 Research Objective 1: Identifying plausible hypotheses of diversification in Formicidae

As in many taxa, the relationships in the Formicidae are contentious. The group is very large, with approaching 16,000 known species. Molecular data and morphological data do not produce the same tree (Figure 1). Different molecular data sources do not produce the same tree (Fig. 1). Our preliminary work shows that different hypotheses about diversification result in different topologies being supported. Therefore, the first step in this project must be to use model fitting approaches to discover the best-fit models for the molecular data, the morphological data, the clock model, and the FBD tree model.

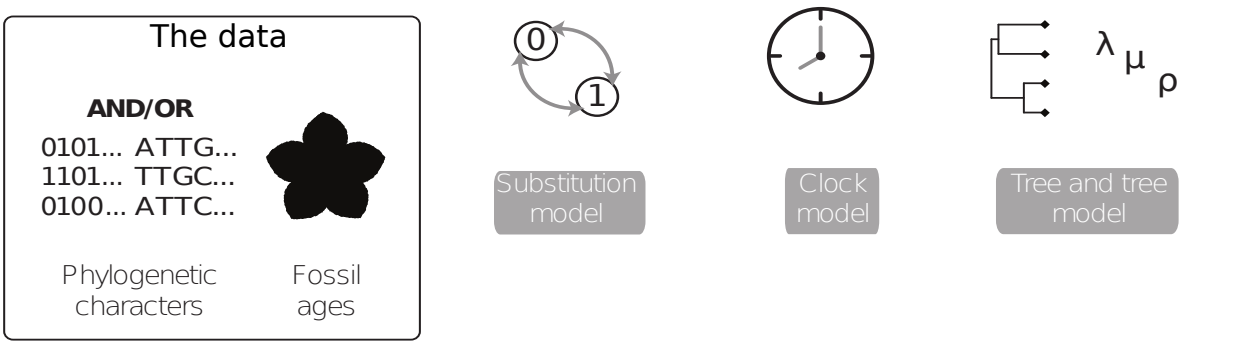
For this objective, I will use a hierarchical model-fitting approach that I have successfully applied in other taxa (Wright, Wagner and Wright 2020). Using stepping-stone model selection (Xie et al., 2011), I will optimize the character model for the molecular data and the morphological data. Stepping stone model selection allows researchers to quantify the goodness of fit of a Bayesian model to the data. Once each component data source has a model, I will combine the data to fit an FBD model describing the process of diversification that lead to the observed tree. As shown on Fig. 1, I have begun some preliminary tests have yielded support for a range of topologies. Stepping-stone model fitting will enable me to discriminate which of these results is best supported.

3.2 Research Objective 2: Using posterior predictive simulation to assess adequacy of timetree models

Stepping-stone model fitting is an important first step for this project. This type of model selection allows for a precise calculation of the marginal likelihood of the data (Xie et al., 2011). It cannot, however, tell us if any of the models in our candidate set of models are performing adequately. That is, are the models capturing crucial features of the generating process that lead to our data? Posterior predictive methods are a family of simulation-based tests that use our actual computed posterior sample to seed simulations with plausible values (Lewis et al., 2014; Hoehna et al., 2017). Summary statistics are then computed from the simulated datasets, and compared with the actual data. If the summary statistics are similar between the simulated and true data, this is evidence that the model is describing the data well. In this project, I will use the posterior sample from the three best models from **Objective 1** to simulate timetrees, distributions of fossils, and lineage through time plots to compare between the empirical and simulated samples. This will enable me to both determine if the models we are using for these analyses are adequate, and if there are features of topologies that may be diagnostic of model inadequacy.

3.3 Research Objective 3: Software solutions for complex model selection

Historically, model selection has been considered a crucial step for estimating phylogenetic trees from molecular sequence data. Partially constrained by the inavailability of multiple models of morphological evolution, this has not been the case with paleontolog-



9 genes, 666 taxa (Blanchard and Moreau 2017)	465 fossil dates	Mk model (Lewis 2001)	Strict Clock	Variable fossil sampling
		Site heterogeneous discrete character model (Wright, Pett, Heath 2017)	Uncorrelated lognormal (descendent evolution rates uncorrelated with parent rates)	Variable diversification rates
Morphology: 135 characters, 127 taxa (Barden and Grimaldi 2016, Keller 2011)			Early burst (decreasing evolutionary rates through time)	Variable diversification rates + fossil sampling
			Autocorrelated rates (descendent branch rates correlated with parent rate)	Fossilization correlated with angiosperm diversity

Figure 2: Table of character evolution, clock, and diversification models that will be tested.

ical data. Due to the current novelty of the FBD model and its derivatives, model selection has also not been common in these types of divergence time analyses. As a result, there is currently a paucity of usable software to do complex posterior predictive model adequacy testing, and little guidance on how to perform even stepping-stone analyses for these complex, hierarchical models. Many phylogenetic posterior predictive pipelines, for example, incorporate a piece of open-source software called RevBayes (Höhna, 2014; Höhna et al., 2016) and the statistical programming language, R (R Core Team, 2013). The final component of this project will be to develop new, robust and useful software tools, and expand already available ones to increase interoperability.

4 Proposed Research Approach

4.1 Research Objective 1: Identifying plausible hypotheses of diversification in Formicidae

4.1.1 Data and Taxa

I will use ants as a focal taxon for this work. The Formicidae is a large family, with around 16,000 extant species. They also have a relatively rich fossil record, compared to many

vertebrate groups. Because ants are of global interest due to their role in ecological interactions, including being common agricultural pests, there is an abundance of molecular and morphological data on this taxon. The ants are the nearly perfect group for the work proposed. As seen in Figure 1, different data sources do not produce the same topology. Different assumptions about evolution produce different topologies. Therefore, to tease apart which models are best supported in this group and use that model to estimate a robust phylogeny is still an open challenge.

The dataset I will be using to test methods and develop software has three components. The first is a set of molecular loci from Blandchard and Moreau (2017). This dataset has 666 species and 11 loci. The second component is a morphological data matrix. I recoded several characters from Barden and Grimaldi (2016) and Keller (2011) to synchronize characters between these two datasets and maximize both paleo- and neontological diversity in this matrix. These first two components are used to esimtate the phylogenetic relationships and evolutionary rates. The final component is a set of nearly 500 fossil occurrence times for ants that I have curated. These are used to estimate divergence times and macroevolutionary parameters under the FBD.

4.2 Hierarchical testing of character evolution models

Many researchers have estimated dated phylogenetic trees of ants, and have found a range of phylogenetic resolutions and ages on nodes (Moreau et al., 2006; Rabeling et al., 2008; Kück et al., 2011; Borowiec et al., 2019). Many of these estimations have been performed in a node calibration framework, which does not completely leverage the fossil information in this clade. Recent FBD analyses have also not used fossil morphology in estimation (Borowiec et al., 2019). I have implemented a mixture model for allowing character change asymmetry in morphological characters in the software RevBayes. The most commonly used character models assume that a character is as likely to be gained as lost. My model allows for characters to be gained and lost at different rates.

FBD analyses also require a clock model that describes how evolutionary rates are distributed across the tree. With morphological character information, it is common to use a strict clock (which implies a constant rate of evolution) or an uncorrelated clock (which implies the rate of evolution of one branch may be very different than its descendants). Very rarely is actual model fitting applied to understand which model actually is the best.

I have already performed model-fitting for the molecular sequence data. Therefore, I will use a hierarchical model fitting approach, following Wright, Wagner, and Wright (2020). First, I will use stepping stone model fitting in RevBayes to assess the best fit model for the character data in an undated phylogentic analysis. This step must come first in order to be able to generate a phylogeny. Then, I will test several clock models (Table 1) against one another. Stepping-stone model fitting calculates the precise marginal likelihood, allowing me to choose the best combination of character change and clock models.

4.2.1 Testing competing models for diversification in ants

Once the sequence and clock model are fit, the tree model can be fit. The tree model describes the distribution of speciation, extinction, and sampling events that generate the phylogeny. Each different tree model can be thought of as a hypothesis about how diversification happened in the ant clade. The FBD model is most commonly applied as a time-homogeneous process. However, most estimates put the clade at at least 120 million years old. Over that time, the rise of flowering plants, with which ants have mutualisms, occurred. There were ice ages, and sea level changes. The distribution of soil types and biomes has changed dramatically. It may be reasonable to expect that we could see differences in diversification rates over time associate with any one of these factors.

I have outlined the candidate diversifiction models on Fig. 2. These models describe a range of scenarios. In some, only specific parameters of the model vary. For example, the fossilization parameter, Psi, may vary more than others. This is because the rate of fossilization is heavily dependent on the presence of plants that generate sap, which crystallizes as amber. Therefore, there are periods in the ants' evolutionary history with no fossils, and some with many. Included in the models that will be competed are models that allow the rates of diversification and sampling to vary over time, as well as those which allow the rate of diversifcation and sampling to vary among clades of ants. I will use the best-fit character change models and clock models for all estimations. Using log Bayes Factors, I will choose the five best models for further testing in **Objective 2**.

4.3 Research Objective 2: Using posterior predictive simulation to assess the adequacy of timetree models

Model selection methods can tell us which of a candidate set of models fits our data best. But the true model may not be in our candidate set, or even proposed yet. Posterior predictive model adequacy testing, however, can tell researchers if the model is capturing key facets of the process that generated the data. In this approach, a Bayesian phylogenetic tree is estimated. Then, parameter values are chosen from the posterior sample and used to simulate new datasets. The degree of similarity between the empirical data and the simulated datasets is then quantified. A model that is capturing the generating process adequately will simulate datasets that are similar to the empirical dataset.

These methods have been applied to non-dated phylogenetic estimates (Hoehna et al., 2017). In these cases, they are typically used to generate nucelotide alignments, which are then compared to the empirical nucleotide alignment from which the tree was estimated. These comparisons normally focus on easy-to-calculate quantities, such as the proportion of different nucleotide bases, number of gaps, etc. In the case of the timetree model, however, we are interested in if the trees themselves are representative of the empirical patterns of diversification. Using the posterior samples of trees estimated in **Objective 1**, I will sample 1000 sets of speciation, extinction and sampling parameters for each of the five best diversification models. Using the R package 'FossilSim', I will then simulate 1000 timetrees for each of the sampled sets of parameters. 'FossilSim' is quite flexible, and implements most of the models listed on Fig. 2.

From each replicate, I will collect summary statistics. These summary statistics will

include the total age of the ant group (the origin time), the number of extant ants, the number of fossils and the distribution of fossils through time. I will then plot these summary statistics against the empirical data to assess how closely the simulated data are tracking the empirical data. Because all parameter values are logged in a Bayesian analysis, I can perform reciprocal posterior predictive simulation. As outlined above, the origin time can be a summary statistic calculated. But this is also a parameter that is fit in the Bayesian analysis. I will also simulate timetrees from the origin time and diversification parameters to see if the produced sampling rates on that tree are reasonable, as well. Likewise, I will simulate timetrees from the origin time and sampling parameters to see if the produced diversification rates match the empirical rates. Posterior predictive analyses allow researchers to detect if models are adequate. This reciprocal experimental design will allow me to detect if specific parameters or sets of parameters are the source of any inadequacy. This will ultimately be a very useful contribution not just to phylogenetics, but more broadly to other types of biological data, as discovery of incorrect of inappropriate assumptions in models is problematic across all disciplines.

4.4 Software solutions for complex model selection

The proposed work outlined above has steps in both RevBayes (Höhna, 2014; **?**) and R (R Core Team, 2013). RevBayes is used for estimating phylogenetic trees and associated model parameters. The R steps above involve using specific packages to simulate trees, calculate summary statistics and visualize results. R is a general-purpose computing language that is quite popular in biology, particularly comparative and population biology. RevBayes has a built-in R-like language, called Rev, that is used to construct phylogenetic models. I, and a co-author have done initial explorations into making RevBayes and R work better together. Called RevKnitR, this R package adapts the popular R document preparation package KnitR to produce interactive reports and tutorials that include both R and Rev code. However, the two languages (R and Rev) do not talk to each other directly. Developing statistical intuition and computational competence in biologists poses many challenges (see **Education Plan**), and phylogenetics is an inherently statistical and computational field. Being able to leverage the advanced statistical and graphical capabilities of R to understand models in RevBayes would be very powerful for many users.

Therefore, the third aim of this proposal is to better integrate R and RevBayes. I would like to build on my preliminary work to improve passing of objects between the two pieces of software. At the present, Rev can be used in the R interface RevKnitR, which works by passing Rev code to RevBayes using the R core command 'system'. The code is then run in RevBayes, and standard output to the command line is echoed back into the RStudio (a graphical user interface for R) console. At the present, it is not possible to pass an object, such as a model or dataset, between R and RevBayes. RevBayes outputs, such as posterior samples and trees, can be imported to R. When a researcher is setting up an analysis, they might want to visualize the uncertainty in a parameter. RevBayes doesn't have the ability to do this, as it has no graphics engine. Ideally, if a researcher was working in RStudio's interface, they could set their prior, then use R's core graphics to view that uncertainty. Other languages, such as Stan and Python, have R interfaces. I'd like to adapt the framework these two languages use for their interfaces by creating a cus-

tom RevBayesR class that can interact with the compiled RevBayes code. In my existing RevKnitR codebase, I will add functions to translate different types of RevBayes objects (distributions, delimited data matrices, trees, and posterior traces) into the R equivalents of those objects. This will allow researchers full control of manipulation over their created Rev objects in both languages.

In the case of the posterior predictive analysis (as decribed in **Objective 2**), this would allow me to run the posterior predictive simulations as the posterior sample is being generated, effectively checking model adequacy as the analysis is in progress. For a researcher performing an empirical analysis, this would provide the researcher with real-time feedback about the quality of their results, as opposed to only being able to obtain that data after the time- and compute-intensive MCMC analysis has completed. This would represent a major streamlining of the posterior predictive pipeline, and could lead adoption of these sorts of methods more broadly.

5 Education Plan

From genomes to ecology, large datasets are now the norm in biology. Many of these datasets are too large to work with tractably via graphical user interface software (such as Excel). Combining datasets, and multiple data layers is also becoming more normal. Scientific software is also most often not created by large corporations, but by other scientists, who may have little or no formal software training. As such, many cutting-edge analyses are performed at the command line, or in languages such as R or Python. As a consequence, researchers need to have biological knowledge of their datasets, the sources of uncertainty and biases in each component of their datasets, and an understanding of how to work computationally with bare-bones software. And yet, undergraduates often have little computational training (Barone et al., 2017; Wilson Sayres et al., 2018; Goldman and Fee, 2017). This has several negative impacts for students: visibility of careers in computational biology may be low (Bares et al., 2018), students may not be well-prepared for graduate or industry work (De Veaux et al., 2017), and faculty supervising research students may absorb more training burden than if computation was well-integrated in the curriculum. I aim to carry out several aims to improve the intercolation of computation into the biology curriculum at Southeastern, and to assess these activities to understand if they can be used to increase student participation in computation.

My goal in this proposal is to expand and evaluate offering early computational biology research training for undergraduate students. Prior research indicates that consistent exposure to computation reinforces learning (Mendez et al., 2016; Behringer and Engelhardt, 2017; De Veaux et al., 2017; Bares et al., 2018). We should integrate computation at all levels so students learn computation as a core tool for solving modern biological problems. Still, major gaps exist in the literature about how to perform this integration, and these gaps can cause faculty to be unsure about introducing computation (Brownell and Tanner, 2012; Williams et al., 2017). I would like to assess if including computation in course work requires biology subject matter to be lost. I will also evaluate if introducing computation into the lower levels of courses increases student awareness of computational careers, and increases the recruitment and representation of underrepresented students in computa-

tion. Lastly, I intend to expand retention programs for cross-disciplinary undergraduate scholars working with computation.

5.1 Preliminary Work

Since beginning as an assistant professor in 2017, I have added 5 classes to Southeastern's course catalog that have computational components. The most obvious of these is my Biological Data Analysis class, which is a mixed upper division undergraduate and Master's course on performing data analysis reproducibly with R. This course also covers other components of literate programming, such as use of a revision management system, documenting code, and use of high-performance cluster computing. This class has been a success on paper. Enrollments and student evaluations are strong. But the class is substantially less diverse than our undergraduate population, with only one-fifth of the students being of color, compared to about a third of the general student population. This lead me to be dissatisfied with my general approach.

In order to serve the broader student population, I began teaching a section of introductory biology with an integrated R component. Every week, students do an in-class activity aimed at analysing data in R to learn about a biological principle. For example, one week, we learned about how organisms' physical features constrain where they live. That week, we did a computational investigation into Bergmann's rule and animal body size. Additionally, I trained another faculty member in the use of R to be able to teach another section of this course. I have also added computational exercises to my genetics class, and in Fall 2020 have added a computational genetics lab to the course catalog. Other offerings, such as my applied systematics lab, also broaden computational training.

In preparing for these courses, I have done extensive research into how to teach these courses equitably. Being in a low-income region of a historically low-income state means that many students don't have home computers. Some work extensive hours outside school, or may have to work on a computer that doesn't belong to them while on a military deployment. Because of this, I have adopted the use of RStudio Server, which enables me to set up a central computational instance to which students connect to do their coursework. Students can connect to this via the internet, on any type of internet-enabled device. A student who does their homework in the library, where they don't have install permissions, can then do their homework as well as a student with a new laptop. Ultimately, these experiences lead to a collaboration with other instructors of computational biology to write a manuscript sharing information on how to use open technology to increase the equity of computing education (Wright et al., 2019).

5.2 Education Objective One: Addressing faculty concerns learning in the computational classroom

Wilson-Sayres et al. (2018) highlighted a number of concerns about teaching computational biology. Among them were concerns about fitting computation into stuffed curricula. Also expressed are concerns that teaching computation via active learning may lead to negative student evaluations of teaching (Potvin and Hazari, 2016), a core component

of promotion and tenure packets. A final is simply lacking the expertise. My first educational aim will be to address these concerns through faculty mentoring networks, and through assessment of student learning in coursework with computational components.

5.2.1 Surveying skill gains in students

A common concern when it comes to computation in biology is where to fit this type of instruction in the curriculum. One way in which this is addressed is through the 'code-to-learn' framework (Resnick, 2013), in which students learn to code through domain-specific exercises that emphasize biology knowledge alongside coding concepts. I have adopted this framework for my introductory biology class. Each week, I have the students work in pairs on an R exercise intended to emphasize a concept or concepts from lecture. However, devoting this time may still lead to less time for other concepts.

I will formally study if students learn less biology domain knowledge in courses with a computational component. I'd like to do this via a paired study design, in which I will pair with two other instructors. First, all the instructors will establish a common set of learning objectives for the course, and will design a pre- and post-test assessment to measure student skills on entrance into the course and after the course (Dugard and Todman, 1995; Lazarowitz and Lieb, 2006; Schiekirka et al., 2013). I will teach my version of the course with an R component. The other instructor will teach the course without the R component. The third instructor will teach one section of each type. Then we will compare post-test scores, and net gain of biological knowledge between the four sections.

5.2.2 Faculty mentorship

When I developed my biology courses, I developed them as R packages, so they can be easily installed and deployed (including a course website) by other instructors. However, having materials is different than having the confidence to teach from them. Many faculty don't have formal training in computation and may need more mentoring to feel comfortable teaching that topic. I have directly mentored other instructors at my home university. I will scale up these activities via a QUBES Faculty Mentoring Network.

Along with a collaborator, I have begun work on an FMN aimed at assisting faculty with adopting computation. At this mentoring network, faculty can sign up in cohorts, each beginning at the start of a semester, based on the type of class they're teaching. To get them started, faculty can adopt a few modules from other QUBES communities. By the end of the semester, though, they will write three code-to-learn computation activities. Each faculty member will host a discussion before teaching the materials with their fellow faculty to spot issues with the lesson and evaluation session. Based on student performance, and feedback of their fellow faculty, they will revise the lesson. If they have made all the revisions, their community mentor will write a letter for their tenure packet attesting to the educator's use of good educational practices and reviewing the literature showing that some students react negatively to this type of instruction. For faculty concerned that incorporating computation will lead to a drop in their student evaluations, this could alleviate those fears and drive adoption of active learning and computation.

5.3 Education Objective Two: Creation of a learning community centered on computation

I will create a learning community in computational science. The learning community model has been popularized at Iowa State University. This model uses research-based principles (Gregerman et al., 1998; Kosoko-Lasaki et al., 2006; Kobulnicky and Dale, 2016) to establish groups of students take similar coursework and meet together in structured discussion sections. These communities may be especially important for underrepresented minority students (URM) studying at primarily white institutions, who often report feeling alienated or discriminated against by majority students and faculty (Chang et al., 2009; Romero, 2018; Thompson et al., 2019). I work at a regional public institution in a poor state. Student retention is a serious problem, particularly for students from underrepresented backgrounds. At Southeastern, degree completion for URM students is 10% lower than for white students. I hypothesize that helping students establish community on campus will improve retention (**?**Newhall et al., 2014).

I am requesting funds for research stipends for 5 local URM students interested in the intersection of biology and computational science annually. Students will receive a $500 monthly stipend during the school year and $1000 a month in the summer to conduct research with a faculty mentor beginning the summer after their freshman year. Research supports financial reasons being a primary cause of dropout (Whittaker et al., 2015); most students at Southeastern work more than 20 hours a week outside school, and many report significant difficulty balancing work and school. Therefore, we might expect that students who develop expertise via active learning, join a mentoring group, and receive a stipend to engage in research might have better outcomes in terms of retention and degree completion (McCavit and Zellner, 2016; Gregerman et al., 1998). All non-transfer Southeastern students take a first year course called 'Southeastern 101', which covers study skills, communicating with professors, financial aid and other life skills. Sections have been piloted that are designed for local URM students, and taught by URM faculty. These five students will be assured entry into these sections to build community among students facing similar challenges.

I would like to use longitudinal surveys with these students, as well. In these surveys, I would like to assess degree completion, time to degree completion, retention in their major, and post-college plans. Students for the learning community will come from local parish high schools. Southeastern has a well-developed Upward Bound program. In collaboration with director Rob Abel, I will recruit students from this program to begin the learning community in the summers after their first year. Southeastern also has a STEM Cafe program, aimed at recruiting high schoolers in underserved Louisiana parishes (counties). In collaboration with director Wendy Conarro, I will develop a set of materials to recruit students. Each undergraduate paid on this project will be responsible for presenting the materials at one STEM Cafe annually. Undergraduates in their junior and senior years will also be allowed to use one hour of their weekly paid time to engage in mentoring of freshmen students via Project Pull, Southeastern's peer mentorship program for URM students.

6 Summary: Significance of proposed work

6.1 Intellectual Merit

Both the research and educational missions of this project will substantially improve their respective fields. The research objectives of this proposal will inform researchers of how to appropriately model complex and heterogeneous data in a hierarchical model. Currently, there has been much research into extensions of the FBD model, but little practical guidance for empiricists on how to actually use and apply these models. With a variety of different assumptions that can be made about the process of evolution that leads to the observed data, it is possible to use a model that does not capture the generating process that underlies the observed data. The proposed work will provide this type of guidance, and will implement software to make these types of analyses more tractable to end users.

The educational components also address key gaps in the literature on how to incorporate computation in undergraduate biology education. In particular, code-to-learn approaches are currently understudied in biology. The pervasive idea that material has to be lost from the curriculum in order to include computation may be limiting the adoption of these skills. Addressing this gap in the literature will have practical consequences for educators and students. The literature on diversity and inclusion in undergraduate education is very robust. But much of it comes from highly-resourced institutions in highly-resourced states. At a public regional institution in a poor state, it isn't necessarily possible to add staff to manage more programs for student retention. Investigating low-impact interventions, such as including computation in lower-division coursework or providing cost-of-living payments for students could be a very important contribution to the literature, as well as an important contribution for policy decisions.

6.2 Broader Impacts

6.2.1 Diversity and Inclusion

Retention of underrepresented students is a major facet of the educational aims of this proposal. Southeastern's student body make up is representative of the local area, approximately 33% African American. Southeastern has a number of recruitment efforts targeted for URM students, low-income students, and first generation students. These pathways will be leveraged to disseminate information about **Education Objective Two**. However, these students are disproportionately lower-income and higher dropout risk. Therefore, this project focuses less on recruitment of new diverse students, and more on enabling the students we have to complete their education and enter the STEM workforce. In particular, dropout is often associated not only with the cost of tuition, but the opportunity cost of a degree: needing money for housing, food, and to assist in supporting family members. The work on this project is aimed at understanding this burden, and attempting retention interventions to alleviate these costs.

The effect of underrepresentation does not end with students. The make-up of faculty is often disproportionately white and male relative to the workforce in a given domain, and to the make-up of the local community. Research consistently demonstrates

that women faculty and faculty of color are less likely to receive faculty positions and less likely to be awarded tenure once they have one. Student evaluation results are often lower for women and non-majority faculty, as well. Therefore, **Educational Aim One** is a faculty diversity retention aim. Because of the above factors, women and underrepresented faculty may be less likely to incorporate computation and active learning. The faculty mentoring network is intended to provide support to faculty to mitigate the effects of the above biases by helping faculty with the promotion and tenure process.

6.2.2 Disseminating results and software

Software associated with this work will be provided for free and hosted on GitHub. R packages will be sent for review at ROpenSci, a non-profit that provides software peer review for R packages. Tutorials associated with analyses in RevBayes will be made available via the RevBayes tutorial repository. Courses, including slides and code, will be made available via my GitHub site. Significant course development will also be sent for review at the Journal of Open Source Education (JOSE), which provides review of course materials and assigns a stable DOI for every reviewed course.

Travel support is requested to send myself, a postdoc and students to present at (1) the Evolution annual meetings, (2) the Geological Society annual meetings, and (3) the Society of Systematic Biology semi-annual meetings. These two meetings are the largest meetings of experts in relevant fields. I also am the head of the iEvoBio organization, which meets after Evolution. This organization hosts a meeting that consists of two sections: software development, and education. Education results will be disseminated here. I have also requested funds for open access publishing. Much morphological work is conducted in museums. Access issues between countries and types of institutions are often difficult, so making results freely available on the internet will be important.

7 Timeline

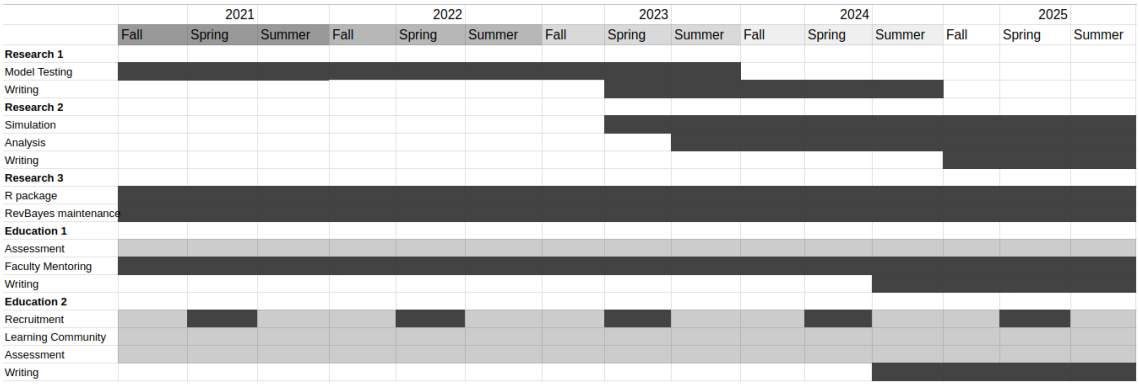


Figure 3: Expected project timeline. Dark gray bars indicate tasks involving substantial concentration and personnel effort. Light gray indicates ongoing tasks or periodic efforts.

References

Barden, P. and D. A. Grimaldi, 2016: Adaptive radiation in socially advanced stem-group ants from the cretaceous. *Current Biology*, **26(4)**, 515–521.

Bares, W., B. Manaris, and R. McCauley, 2018: Gender equity in computer science through computing in the arts – a six-year longitudinal study. *Computer Science Education*, **28(3)**, 191–210.

Barido-Sottani, J., G. Aguirre-Fernández, M. J. Hopkins, T. Stadler, and R. C. Warnock, 2019: Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process. *Proceedings of the Royal Society B*, **286(1902)**, 20190685.

Barone, L., J. Williams, and D. Micklos, 2017: Unmet needs for analyzing biological big data: A survey of 704 nsf principal investigators. *PLOS Computational Biology*, **13(10)**, 1–8.

Behringer, E. and L. Engelhardt, 2017: Guest editorial: Aapt recommendations for computational physics in the undergraduate physics curriculum, and the partnership for integrating computation into undergraduate physics.

Blanchard, B. D. and C. S. Moreau, 2017: Defensive traits exhibit an evolutionary trade-off and drive diversification in ants. *Evolution*, **71(2)**, 315–328.

Borowiec, M. L., C. Rabeling, S. G. Brady, B. L. Fisher, T. R. Schultz, and P. S. Ward, 2019: Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. *Molecular phylogenetics and evolution*, **134**, 111–121.

Brown, J. M., 2014: Predictive approaches to assessing the fit of evolutionary models. *Systematic Biology*, **63(3)**, 289–292.

Brownell, S. and K. Tanner, 2012: Barriers to faculty pedagogical change: Lack of training, time, incentives, and...tensions with professional identity? *CBE Life Sciences Education*, **11(4)**, 339–346.

Chang, M. J., M. K. Eagan, M. Lin, and S. Hurtado, 2009: Stereotype threat: Undermining the persistence of racial minority freshmen in the sciences. In *annual meeting of the American Educational Reserach Association, San Diego, CA*.

De Veaux, R. D., M. Agarwal, M. Averett, B. S. Baumer, A. Bray, T. C. Bressoud, L. Bryant, L. Z. Cheng, A. Francis, R. Gould, et al., 2017: Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, **4**, 15–30.

Donoghue, P. C. and M. J. Benton, 2007: Rocks and clocks: calibrating the tree of life using fossils and molecules. *Trends in Ecology & Evolution*, **22(8)**, 424–431.

Duchêne, D. A., S. Duchêne, E. C. Holmes, and S. Y. Ho, 2015: Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Molecular Biology and Evolution*, **32(11)**, 2986–2995.

Dugard, P. and J. Todman, 1995: Analysis of pre-test-post-test control group designs in educational research. *Educational Psychology*, **15(2)**, 181–198.

Eck, R. V. and M. O. Dayhoff, 1966: *Atlas of Protein Sequence and Structure, V. 3-5*. National Biomedical Research Foundation.

Felsenstein, J., 1981: Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17(6)**, 368–376.

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond, 2017: Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Systematic Biology*, **66(1)**, 57–73.

Goldman, M. S. and M. S. Fee, 2017: Computational training for the next generation of neuroscientists. *Current opinion in neurobiology*, **46**, 25–30.

Gregerman, S. R., J. S. Lerner, W. Von Hippel, J. Jonides, and B. A. Nagda, 1998: Undergraduate student-faculty research partnerships affect student retention. *The Review of Higher Education*, **22(1)**, 55–72.

Harvey, P. H., R. M. May, and S. Nee, 1994: Phylogenies without fossils. *Evolution*, **48(3)**, 523–529.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler, 2014: The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, **111(29)**, E2957–E2966.

Hillis, D., T. Heath, and K. John, 2005: Analysis and Visualization of Tree Space. *Systematic Biology*, **54(3)**, 471–482.

Hoehna, S., L. M. Coghill, G. G. Mount, R. C. Thomson, and J. M. Brown, 2017: P3: Phylogenetic Posterior Prediction in RevBayes. *Molecular Biology and Evolution*, **35(4)**, 1028–1034.

Höhna, S., 2014: Likelihood Inference of Non-Constant Diversification Rates with Incomplete Taxon Sampling. *PLoS One*, **9(1)**, e84184.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist, 2016: RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, **65(4)**, 726–736.

Jukes, T. and C. Cantor, 1969: Evolution of protein molecules. *Mammalian Protein Metabolism*, **3**, 21–132.

Keller, R. A., 2011: A phylogenetic analysis of ant morphology (hymenoptera: Formicidae) with special reference to the poneromorph subfamilies. *Bulletin of the American museum of natural history*, **2011(355)**, 1–90.

Kimura, M., 1980: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16(2)**, 111–120.

Kobulnicky, H. A. and D. A. Dale, 2016: A community mentoring model for stem undergraduate research experiences. *Journal of College Science Teaching*, **45(6)**, 17.

Kosoko-Lasaki, O., R. E. Sonnino, and M. L. Voytko, 2006: Mentoring for women and underrepresented minority faculty and students: experience at two institutions of higher education. *Journal of the national medical association*, **98(9)**, 1449.

Kück, P., F. H. Garcia, B. Misof, and K. Meusemann, 2011: Improved phylogenetic analyses corroborate a plausible position of martialis heureka in the ant tree of life. *PLoS One*, **6(6)**.

Lartillot, N., 2006: Conjugate Gibbs sampling for Bayesian phylogenetic models. *Journal of Computational Biology*, **13(10)**, 1701–1722.

Lazarowitz, R. and C. Lieb, 2006: Formative assessment pre-test to identify college students' prior knowledge, misconceptions and learning difficulties in biology. *International Journal of Science and Mathematics Education*, **4(4)**, 741–762.

Lemmon, A. R. and E. C. Moriarty, 2004: The Importance of Proper Model Assumption in Bayesian Phylogenetics. *Systematic Biology*, **53(2)**, 278–298.

Lewis, P. O., 2001: A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology*, **50(6)**, 913–925.

Lewis, P. O., W. Xie, M.-H. Chen, Y. Fan, and L. Kuo, 2014: Posterior predictive Bayesian phylogenetic model selection. *Systematic Biology*, **63(3)**, 309–321.

Maddison, W. P. and L. L. Knowles, 2006: Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55(1)**, 21–30.

Matzke, N. J. and A. Wright, 2016: Inferring node dates from tip dates in fossil canidae: the importance of tree priors. *Biology Letters*, **12(8)**, 20160328.

McCavit, K. and N. Zellner, 2016: Persistence of physics and engineering students via peer mentoring, active learning, and intentional advising. *European Journal of Physics*, **37(6)**, 065702.

Mendez, R. G., J. Torres, P. Ishwad, H. B. Nicholas, and A. Ropelewski, 2016: Assisting bioinformatics programs at minority institutions: Needs assessment, and lessons learned – a look at an internship program. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, Association for Computing Machinery, New York, NY, USA, XSEDE16, ISBN 9781450347556.

Mirarab, S., N. Nguyen, S. Guo, L.-S. Wang, J. Kim, and T. Warnow, 2015: Pasta: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, **22(5)**, 377–386.

Moreau, C. S., C. D. Bell, R. Vila, S. B. Archibald, and N. E. Pierce, 2006: Phylogeny of the ants: diversification in the age of angiosperms. *Science*, **312(5770)**, 101–104.

Newhall, T., L. Meeden, A. Danner, A. Soni, F. Ruiz, and R. Wicentowski, 2014: A support program for introductory cs courses that improves student performance and retains students from underrepresented groups. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pp. 433–438.

Potvin, G. and Z. Hazari, 2016: Student evaluations of physics teachers: On the stability and persistence of gender bias. *Phys. Rev. Phys. Educ. Res.*, **12**, 020107.

R Core Team, 2013: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabeling, C., J. M. Brown, and M. Verhaagh, 2008: Newly discovered sister lineage sheds light on early ant evolution. *Proceedings of the National Academy of Sciences*, **105(39)**, 14913–14917.

Resnick, M., 2013: Learn to code, code to learn. *EdSurge, May*, **54**.

Romero, D., 2018: Examining academic success for underrepresented minority science technology engineering and mathematics students within hispanic serving institutions and predominantly white institutions.

Schiekirka, S., D. Reinhardt, T. Beibarth, S. Anders, T. Pukrop, and T. Raupach, 2013: Estimating learning outcomes from pre-and posttest student self-assessments: a longitudinal study. *Academic Medicine*, **88(3)**, 369–375.

Slater, G. J. and M. W. Pennell, 2014: Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Systematic Biology*, **63(3)**, 293–308.

Stadler, T., 2010: Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, **267(3)**, 396–404.

Tavaré, S., 1986: Some probabilistic and statistical problems in the analysis of DNA sequences. *Some Mathematical Questions in Biology: DNA Sequence Analysis*, **17**, 57–86.

Thompson, G., J. Ponterotto, and C. Dyar, 2019: Social identity pathways to college choice and attitudes toward help-seeking among black students at a minority serving institution. *College Student Journal*, **53(1)**, 113–129.

Warnock, R. and A. Wright, 2020: Understanding the tripartite approach to bayesian divergence time estimation.

Warnock, R. C., T. A. Heath, and T. Stadler, 2020: Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. *Paleobiology*, 1–21.

Warren, D., A. Geneva, D. Swofford, and R. Lanfear, 2016: rwty: R we there yet. *A package for visualizing MCMC convergence in phylogenetics*.

Whittaker, J. A., B. L. Montgomery, and V. G. M. Acosta, 2015: Retention of underrepresented minority faculty: Strategic initiatives for institutional value proposition based on perspectives from a range of academic institutions. *Journal of Undergraduate Neuroscience Education*, **13(3)**, A136.

Williams, J., J. Drew, S. Galindo-Gonzalez, S. Robic, E. Dinsdale, W. Morgan, E. Triplett, J. Burnette, S. Donovan, S. Elgin, et al., 2017: Barriers to integration of bioinformatics into undergraduate life sciences education. *BioRxiv*, 204420.

Wilson Sayres, M. A., C. Hauser, M. Sierk, S. Robic, A. G. Rosenwald, T. M. Smith, E. W. Triplett, J. J. Williams, E. Dinsdale, W. R. Morgan, J. M. Burnette, III, S. S. Donovan, J. C. Drew, S. C. R. Elgin, E. R. Fowlks, S. Galindo-Gonzalez, A. L. Goodman, N. F. Grandgenett, C. C. Goller, J. R. Jungck, J. D. Newman, W. Pearson, E. F. Ryder, R. Tosado-Acevedo, W. Tapprich, T. C. Tobin, A. Toro-Martínez, L. R. Welch, R. Wright, L. Barone, D. Ebenbach, M. McWilliams, K. C. Olney, and M. A. Pauley, 2018: Bioinformatics core competencies for undergraduate life sciences education. *PLOS ONE*, **13(6)**, 1–20.

Wright, A. M., 2019: A Systematist's Guide to Estimating Bayesian Phylogenies From Morphological Data. *Insect Systematics and Diversity*, **3(3)**.

Wright, A. M. and G. T. Lloyd, 2020: Bayesian analyses in phylogenetic paleontology: Interpreting the posterior sample. *Palaeontology*.

Wright, A. M., G. T. Lloyd, and D. M. Hillis, 2016: Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, **65(4)**, 602–611.

Wright, A. M., R. S. Schwartz, J. R. Oaks, C. E. Newman, and S. P. Flanagan, 2019: The why, when, and how of computing in biology classrooms. *F1000Research*, **8(1854)**, 1854.

Xie, W., P. Lewis, Y. Fan, L. Kuo, and M. Chen, 2011: Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, **60(2)**, 150–160.

Yang, Z. and B. Rannala, 2006: Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Molecular Biology and Evolution*, **23(1)**, 212–226.

Zuckerkandl, E. and L. Pauling, 1962: Molecular disease, evolution, and genetic heterogeneity. In Kasha, M. and B. Pullman, eds., *Horizons in Biochemistry*, Academic Press, New York, pp. 189–225.

—, 1965: Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, **97**, 97–166.

BUDGET JUSTIFICATION

Funding for a five-year project (start date approximately 08/01/2021) is requested.

A. Senior Personnel

Funds are requested to support 2 months of summer salary per year for 5 years for the PI. The PI will be responsible for leading the project, including (1) directing, participating in, and disseminating research, (2) leading computational training, developing coursework, and disseminating tutorials, and (3) mentoring the postdoc and students participating in the research and educational components of the grant. This calculation includes 4% cost of living allowance In the event that Southeastern is able to provide such in accordance with an approved raise plan. (11,733 year one, 12,202 y2, 12,672 y3, 13,141 y4, 13,610 y5, 63,358 for all five years)

B. Other Personnel

Funds are requested to support one postdoctoral researcher for each year of the grant starting at $42,000 per year, increasing by 4% per year. Note that the individual occupying that position is expected to change at least once over the project, in anticipation of the first post-doc moving onto a job elsewhere before the end of the grant. This calculation includes 4% cost of living allowance In the event that Southeastern is able to provide such in accordance with an approved raise plan. (42,000 y1, 43,680 y2, 45428 y3, 47245 y4, 49134 y5, $227,488 total) Funds are requested for graduate student stipend per academic year for years 2-5 years ($12,000 per year $48,000 total), and in-state tuition (11915 y2, 12391 y3, 12887 y4, 13402 y5, $62051 total). Funding will support graduate students assisting in the development of coursework, conducting research, collecting and managing assessment data, and mentoring laboratory undergraduates. Funds ($13860 annually for five years, 69300 total) are also requested to support two undergraduate students conducting research in the lab for 10 hours per week during the school year, full time during the summer. These students will work closely with the postdoc and graduate students on data collection and analysis.

C. Fringe Benefits

Fringe benefits are calculated and requested for the postdoc ($ 86,446 over five years) and PI summer salary ($4013*5 = $26,429) at the current Southeastern percentage rate of 38% of the salary (full family benefits, as per Southeastern policy).

D. Equipment

Funds are requested to purchase a small, networked data storage server (500). Funds are also requested to purchase desktop computers in years 4 and 5 to replace aging desktops in the lab ($10,000).

E. Travel

Domestic: Funds are requested for the postdoc, PI, and students to attend the annual meeting of the Society for the Study of Evolution or the semi-annual meeting of the Society of Systematic Biology (or similar meetings) to present on results in Years 1-5 (37,500).

F. Participant Support Costs

To support educational aim three, funds are requested to pay students a research stipend. The stipend will be $500 monthly during the academic year and $ 1000 monthly in the summer. ($7000 per participant per year, 15 participants over 5 years, $337500 total).

G. Other Direct Costs

Computers: Because southeastern Louisiana is a historically impoverished area, funds are requested to buy a laptop for each postdoc, Masters' student, and undergraduate student. These laptops will be returned on graduation from lab, and so fewer laptops are requested than total personnel (18,000).

Publication Costs: $1,800 per year in Years 2-5 is requested to cover the cost of one open-access publication per year in the journal Systematic Biology (or similar).

I. Indirect Costs

The indirect cost rate for Southeastern is 38.6% of the Modified Total Direct Costs ($647,505), which include all direct costs with the exception of large equipment, any additional cost of subawards above the first $25,000 (none in this proposal), participant costs, and tuition. ($ 249,937)

J. Total Direct and Indirect Costs

The total direct and indirect costs requested for the project period is $1,258,398.

Facilities, Equipment and Other Resources

Facilities Physical spaces are available for all aspects of this project at Southeastern. The Wright lab space is a large, open plan space with room for six researchers. This space is located in the same building, Thelma Ryan Hall, as Dr. Wright's personal office. Graduate students also have cubicles available to them. Teaching computer lab space will be used for the sections of introductory biology with integrated R components. The postdoc will both have a private office at Southeastern, and desk space in the laboratory of Dr. Jeremy Brown at Louisiana State University (see letter of collaboration from Dr. Brown.

Equipment

The state of Louisiana operates the Louisiana Optical Network Infrastructure (LONI) supercomputer. LONI is a 1.5 Petaflop high-performance compute cluster containing over 10,000 Intel Xeon processing cores. The nodes the Wright lab has typically been using for computation have 64 GB RAM, and flexible walltime limits. There is an array of computational software pre-installed on the cluster, as well as a standard set of compilers for building custom software. The Wright Lab additionally has two lab servers, each with 64 GB of RAM, 254 GB solid state drives, and automated backups.

For use in the educational aims, the free educational service RStudio Cloud, will be used.

Other Resources

Additional assistance recruiting students from federal student assistance pipelines (TRiO, Upward Bound), will be provided by Mr. Ron Abels and Ms. Wendy Conarro (see letters of commitment). Abels and Conarro each run a TRiO program for different parts of Louisiana.

Data Management Plan

Research data

New raw data will not be generated by this research. Biological data will be obtained from public databases. Simulated data and estimated phylogenies will be deposited in appropriate data repositories such as Dryad (at the time of peer review) and TreeBase. At the time of publication, phylogenies will be added to the Open Tree of Life.

Research code and software

Code to replicate research findings will be stored on GitHub as part of the version control pipeline. Upon review of any manuscripts from the project, the code will be deposited on data Dryad. Code will be annotated appropriately, and a small written manual will be provided to facilitate reproducibility of results and appropriate use of the code. An example of this structure from a recent paper from the Wright Lab can be seen here.

There is a moderate R package development component to this project. Upon completion, R packages will be sent to ROpenSci, a non-profit organization that provides code and documentation review for research software written in R. When a package is reviewed by ROpenSci, it is subsequently given a stable Digital Object Identifier (DOI) and archived via the ROpenSci website. For an example of an R package from the PI's lab that is archived in ROpenSci, see here.

Educational materials

Tutorials for use in courses will be maintained on Github in the easy-to-use R Markdown format, and made available on the PI's website. After several iterations of the introductory biology course will be sent to the Journal of Open Source Education, a scientific journal that does code review for educational materials that involve heavy computation components. These courses are then backed up in the journal's Zeonodo repository, providing a stable DOI for the materials. An example of a course the PI has had peer reviewed at JOSE can be seen here.

Surveys are conducted to assess the impact of integrating computational skills in coursework, the data will be collected anonymously and/or de-identified in accordance with IRB protocols (see letter of support from IRB director Dr. Michelle Hall). Dr. Hall will also advise on data archival at the time of publication.

Postdoctoral Mentoring Plan

This project will fund postdoctoral training for 1-3 individuals. I anticipate hiring one postdoc right away, and that this individual will likely leave for a permanent position. Each postdoc will complete the American Societies for Experimental Biology mentoring plan (http://myidp.sciencecareers.org/). This will be a springboard for us to evaluate the skills the postdoc has, the skills they would like to gain, and the timeline for doing so. We will need to re-evaluate skills gains according to this tool annually during the award period.

Southeastern does not have the research activity of an R1 institution. This means that the postdoc would have less access to seminars, to a community of other postdoctoral researchers, and to other training. Therefore, I am proposing to have the postdoc housed with Dr. Jeremy Brown (senior personnel on this grant) three days a week. This will permit them to interact with other early career researchers more readily. The postdoc will be responsible for teaching one section of introductory biology with R each semester, and one other course of their choosing. This could be another section of introductory biology without R (pursuant to **Educational Aim One**), or another course they would like to develop. The postdoc will be expected to be the lead author on the manuscript on educational gains that results from **Educational Aim One**. Therefore, they will be expected to have an active teaching role, but also to analyse the data associated with the assessments described in this aim.

Postdocs hired on this proposal will work with me to mentor Master's and undergraduate students to pursue **Research Aims One and Two**, and to help develop software solutions for this work. Funds are requested for the postdoc to present their research at the Evolution or Systematic Biology meetings. While the postdoc will not be the primary mentor of the students, they will develop mentorship as a skill by working with me to supervise the students. I have also requested one month of summer salary for Dr. Jeremy Brown for his role in housing the postdoc. The postdoc will be encouraged to develop an independent project with Dr. Brown and any students in his lab.

To assess the success of the postdoc's progress, I will do the following:

- Keep track of the postdoc's progress toward goals as laid out in the mentoring plan
- Research outputs, including papers, software, and talks
- Teaching evaluation scores

Together, these three items should provide a wholistic look at how the postdoc is progressing. I will compile these into a report using Southeastern's annual faculty evaluation template, and discuss this as a performance review with the postdoc annually.