2013

Model-based approaches to discovering diversity: new implementations, tests of adequacy and an empirical application to central American Diptera

Noah M. Reid
Louisiana State University and Agricultural and Mechanical College, nreid1@tigers.lsu.edu

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/18

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
MODEL-BASED APPROACHES TO DISCOVERING DIVERSITY: NEW IMPLEMENTATIONS, TESTS OF ADEQUACY AND AN EMPIRICAL APPLICATION TO CENTRAL AMERICAN DIPTERA

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

The Department of Biological Sciences

by
Noah Mattoon Reid
B.S. Boston University 2003
M.S. University of Idaho 2008
August 2013
Acknowledgments

I want to first thank my primary mentor in this work, Bryan Carstens. He was always full of ideas, available for meetings, and turned around writing within 48 hours. He encouraged me to develop my own research system and explore new ideas while giving me the space and support to do so. That kind of independence was one of the things I sought in doctoral training, and though it turned out to be much more challenging than I had imagined, I’m grateful that Bryan gave me the opportunity to have that experience. I also want to thank my committee members. Discussions about phylogenetics and Bayesian statistics with Jeremy Brown were critical in motivating Chapters 2 and 3. Richard Stevens’ macroecology course planted the seeds that grew into Chapter 4. When I reached what I thought were impasses in my work, discussions with Michael Hellberg inevitably introduced me to other approaches and perspectives.

The Carstens Lab at LSU was also critical in my personal and professional development. John McVay and I embarked on this journey together and are now finishing it simultaneously. My time as lab mates was marked by lively discussions of developments in the literature, critical discussion of my respective projects and invaluable blowing off of steam in the form of beers and YouTube videos. Tara Pelletier and Jordan Satler were great lab mates and friends. Postdoctoral researchers Maggie Hanes, Eric Tsai and Amanda Zellmer, all provided valuable perspective on both the design of my project and my ultimate career plan.

I also want to thank the community of graduate students at LSU with whom I interacted, which is too large to list here. The constant churn of ideas in the Department, in both the main building and the Museum of Natural Science is inspiring and I have benefited greatly from it.

Various other individuals have contributed to the completion of this work, including the departmental administrative staff, especially Priscilla Milligan, who helped me out of many binds, and Dr. James Moroney, who made sure I had access to an equipment-filled lab and office space after the Carstens Lab moved to Ohio State. Francisco Campos at the Organization for Tropical Studies was instrumental in helping me secure permits to conduct fieldwork in Costa Rica, and a number of individuals at MINAE and SINAC were also helpful in the execution of that work.

I want to thank my family. They have been incredibly supportive over the last five years. I particularly want to acknowledge my wife, Sarah Hird, who, simply put, was the single essential ingredient in every good thing I have accomplished in the last five years.

Funding for this work was provided by the Louisiana Board of Regents, the National Science Foundation [DEB – 0918212], the Society of Systematic Biologists and the American Alpine Club. Some of the work was conducted using the resources of the LSU High Performance Computing facility.
# Table Of Contents

Acknowledgments ................................................................................................................................. ii

Abstract ......................................................................................................................................................... iv

Chapter 1. Introduction ................................................................................................................................... 1

Chapter 2. Phylogenetic Estimation Error Can Decrease The Accuracy of Species Delimitation: A Bayesian Implementation Of The General Mixed Yule-Coalescent Model ................................................................................................................................. 4

Chapter 3. Poor Fit To The Multispecies Coalescent Is Widely Detectable In Empirical Data ................................................................................................................................................................. 19

Chapter 4. Integrating Ecological And Evolutionary Methods For Inference Of Species Diversity And Community Structure: An Example Using A Tropical Alpine Fly Community ............................................................................................................................................... 35

Chapter 5. Conclusions ................................................................................................................................... 54

References ...................................................................................................................................................... 57

Appendix A. bGMYC Instructions For Use .............................................................................................................. 74

Appendix B. starbeastPPS Instructions For Use ..................................................................................................... 77

Appendix C. License For Reproducing Chapter 2 ................................................................................................. 82

Vita ................................................................................................................................................................. 86
Abstract

Most of the earth’s biodiversity is unknown to science. With global climate change set to drastically alter its distribution, it is imperative to catalogue it and understand its function in order to preserve it and better understand how this change will impact humanity. Recent technological and statistical advances have in theory made possible increasingly rapid discovery and description of diversity. The statistical properties and performance of these new approaches are still poorly known, however, their integration with complementary methods from disparate disciplines has not been achieved. In this dissertation we present three chapters of original research that advance these areas of biodiversity science. The first introduces a new implementation of the GMYC, a statistical model used for species delimitation. This implementation fully accounts for uncertainty in model parameters, and we test its performance under various historical scenarios. We find that the model generally performs well, but that failing to account for uncertainty in nuisance parameters inflates confidence in species limits. The second introduces a method to examine the fit of empirical data to a multispecies coalescent model commonly used in phylogenetic inference. Systematic and phylogeographic studies are generating ever-larger datasets which often range up to the genome scale and often wish to use coalescent models to infer parameters such as phylogenies, divergence times and effective population sizes. Though the multispecies coalescent can infer these parameters, it is unclear the extent to which it is a good fit for these new empirical datasets. We employ our new approach to 25 published datasets and find that a majority of them show poor fit to the data and that for some of them, that poor fit affects inference. In the last chapter we integrate statistical approaches from both ecology and systematics to infer species limits, phylogeny, population genetic structure and ecological community structure in a study of a poorly known tropical alpine fly fauna. We find that we can effectively describe patterns of diversity in the absence of a low-level taxonomic framework, but that inference of the processes structuring that diversity remains difficult. We also find that some of our inferences of community structure are sensitive to uncertainty in species limits and phylogeny.
Chapter 1.
Introduction

We have entered a period of rapid anthropogenic global change that will permanently alter the distribution of earth’s biodiversity (Parmesan, 2006, Pounds et al., 2006, Thomas et al., 2004). The consequences, both to humanity and to that biodiversity are poorly understood. In part, this is because biodiversity itself is poorly understood: in spite of over 250 years of taxonomic work, less than 13% of an estimated 10.9 million species have been given formal scientific descriptions (Mora et al., 2011). Of these species, distributions, historical relationships, and life histories are known in detail for only a small portion and general knowledge of the types and magnitudes of climatic and biotic change to which organisms can adapt is meager. Mitigating the effects of global change, both on biodiversity and on humanity’s interaction with it will require a greater understanding of the patterns and processes of biodiversity.

Methods for describing species and their distributions historically relied heavily on the judgments of relatively few expert taxonomists. Their skills, while extraordinary in their depth, are typically narrow in scope, encompassing from a few hundreds to a thousand species. In fields such as entomology, these skills may take up to a decade of training to master and have limited transferability among taxa. The application of these skills is also extremely labor intensive. The current average time from the collection of a specimen of a new species until its description is 21 years (Fontaine et al., 2012). A recent study estimates that the average taxonomist describes 25 species in their career. At the current rate of ~16,000 new species descriptions per year, it would take 360 years to catalogue global animal diversity (Carbayo & Marques, 2011). Furthermore, under the classical taxonomy model, the primary means of making taxonomic information available to evolutionary biologists and ecologists downstream is through monographs and keys. These works can be extremely technical and rely on obscure, sometimes subjective interpretations of organismal characteristics for identification, with the end result being that input from expert taxonomists is required to use them.

During the last decade, statistical advances have increased the speed, accuracy, and efficiency of efforts to describe biodiversity. Evolutionary models based on genetic data allow the inference of phylogenies, species boundaries, and assignment of individuals to species (Heled & Drummond, 2010, Yang & Rannala, 2010, Pritchard et al., 2000, Knowles & Carstens, 2007a). Distributional models enable inference of species ranges (Anderson et al., 2006), and combinations of these models and paleoclimatic data can give information on species historical distributions and how they may influence contemporary patterns (Carstens & Richards, 2007, Knowles & Alvarado-Serrano, 2010). Ecological statistics that quantify and estimate patterns of species diversity have become more rigorous (Chao et al., 2005, Colwell et al., 2012), and models of community assembly have been developed to infer the determinants of community structure, although their use for this purpose has become controversial (Kembel et al., 2010, Webb et al., 2002, Helmus et al., 2007).
Historically, researchers interested in species distributions or the structure of ecological communities behaved as consumers of taxonomic information and were restricted from studying organisms that had not received sufficient attention by taxonomists. These advances suggest the possibility that application of more general modeling skills could enable researchers to generate sufficient estimates species identifications, distributions, and ecologies without requiring either assistance from taxonomists or prior experience in the group of interest. Along with effective data archiving and maintenance of voucher specimens, this would also free trained taxonomists to work on difficult problems and mine data generated by researchers taking this approach (Maddison et al., 2012).

A case that highlights the extent to which this development of statistical approaches to describing biodiversity will improve over the status quo is that of the Raspberry crazy ant (RCA). In Houston, Texas in 2002, this exotic ant species appeared and began an explosive invasion. It forms very large, dense colonies and is successful at displacing the invasive fire ant, Solenopsis invicta, which is considered one of the most aggressive and destructive invasive insects in the world (Morrison et al., 2004). For ten years, taxonomists failed to determine the specific identify of the RCA. This means that information regarding its geographic origins, native climate, natural enemies and past history of invasiveness that is fundamental to control efforts was inaccessible. Finally, in 2012, a group of taxonomists determined that the RCA was conspecific with an invasive ant that appeared in Florida decades ago, Nylanderia fulva, itself misidentified as a congener, N. pubens, that has widely invaded the Caribbean (Gotzek et al., 2012).

The further development of novel approaches to species delimitation is warranted because the time requirement for classical taxonomic work is prohibitive in situations such as the RCA. It will also allow community-scale research to be conducted in the absence of reliable or accessible traditional taxonomic frameworks. While substantial work remains to achieve this, including the development of new models, refinement of existing ones, and simulation testing under a variety of conditions, the work presented in the following dissertation makes fundamental advances toward this goal.

1.1. OVERVIEW OF CHAPTERS

In chapter two I develop a new implementation of a single-locus species delimitation model, the general mixed Yule-coalescent model (Pons et al., 2006). This model takes single-locus DNA sequence data, such as DNA barcodes, from a collection of unknown species and uses a tree estimated from that data to estimate species limits. This model has been widely used in the literature. The original publication has over 200 citations, but its performance under various demographic scenarios and sensitivity to uncertainty in the tree used to infer species limits were unknown. I developed a Bayesian implementation of the model that integrates over phylogenetic uncertainty, and uncertainty in the model parameters using Markov chain Monte Carlo and tested performance of the model under varying species divergences and with varying phylogenetic information. Results from this
chapter are published in *BMC Evolutionary Biology* (Reid & Carstens, 2012) and the model implementation is released in an R package, bGMYC.

In chapter three I develop a method of posterior predictive model check for a multispecies coalescent model used in phylogenetic estimation. Multispecies coalescent models use multilocus genetic data to infer phylogenies, species limits and demographic histories. In simulation they have proven to be effective, useful tools, but their fit to empirical data has rarely been evaluated. As genome-scale datasets become available, it is to be expected that some loci will have complex evolutionary histories that are not well described by these models. The pervasiveness of such loci and the sensitivity of the models are poorly known. I developed a posterior predictive model check that evaluates model fit and can identify outlier genes with inconsistent evolutionary histories. I applied this model check to 25 published datasets and discuss the sources of poor fit I found. Results from the chapter are accepted for publication in the journal *Systematic Biology* (Reid et al., 2013), and the method is released in an R package, MCPoPS.

In the fourth chapter I integrate species delimitation models, phylogenetic estimation and ecological diversity analyses to elucidate the processes that have structured the Diptera fauna of a poorly known tropical alpine ecosystem. I examine the prospects and limitations of this approach for identifying community assembly processes at the regional scale in the absence of any prior taxonomic framework below the Family level. I explore the sensitivity of my conclusions to uncertainty in phylogenetic relationships and to the approach I take to defining species.
Chapter 2.
Phylogenetic Estimation Error Can Decrease The Accuracy Of Species Delimitation: A Bayesian Implementation Of The General Mixed Yule-Coalescent Model*

2.1. BACKGROUND

A common challenge faced by empirical researchers in studies of ecological communities is to identify individuals at the species level from limited information collected from a broad taxonomic range of organisms. In many cases, useful taxonomic keys for particular groups or regions are not available. This is because many diverse groups are morphologically cryptic, contain many undescribed taxa, or existing taxonomic literature is conflicting, an issue referred to as the "taxonomic impediment" (Hoagland, 1996). In these cases, short DNA sequence tags (the DNA barcode region of the mitochondrial gene COI, or a hypervariable region of the microbial 16S rRNA gene) are frequently surveyed because they can be rapidly and inexpensively collected (Hebert et al., 2003, Hughes et al., 2001). DNA barcoding initiatives aim to connect these sequence tags to taxa validated by expert taxonomists (Cole et al., 2009, Ratnasingham & Hebert, 2007), but at present this is not possible for most groups. As a result, diversity must frequently be quantified in the absence of a low-level taxonomic framework. In order to accomplish this, observed DNA sequences must be clustered into putative species. While the delimitation of species is a complex philosophical and biological problem (Hey et al., 2003), species concepts widely share the idea that species are independently evolving metapopulation lineages (De Queiroz, 2007). This provides a justification for using genetic data (such as DNA barcodes) as the primary data for the diagnosis of these lineages, as they contain the signal of historical processes involved in lineage divergence (Nielsen & Wakeley, 2001). As a caveat, lineages identified in this way will not necessarily meet the criteria for species status under any given species concept, such as reproductive isolation from other such lineages, or exhibit morphological, ecological or behavior divergence.

Methods used for delimitation of species from barcode data are a subset of those developed for the larger problem of species delimitation. They can be considered species discovery methods because they must be functional in the absence of good a priori taxonomic information (Ence & Carstens, 2011, Hebert et al., 2004, Smith et al., 2007). This contrasts with validation methods (e.g. (Ence & Carstens, 2011, Yang & Rannala, 2010)), which test specific hypotheses of species status, and assignment methods, which assign unknown individuals to existing species (e.g. (Abdo & Golding, 2007, Munch et al., 2008, Nielsen &

Matz, 2006, Kelly et al., 2007). Of the handful of approaches typically used to discover species limits using genetic data, thresholds based on pairwise sequence distances among individuals are perhaps most commonly applied to cluster sequences into putative species (Ratnasingham & Hebert, 2007, Schloss et al., 2009). These methods identify some level of sequence divergence beyond which two individuals cannot be considered the same species. Distance threshold methods have been criticized because they do not account for evolutionary processes (Hickerson et al., 2006), and the uncertainty in selecting an appropriate threshold (Nielsen & Matz, 2006), which relies on an observable “barcode gap” between pairwise intraspecific and interspecific DNA sequence distances((Meyer & Paulay, 2005, Meier et al., 2008, Meier et al., 2006, Wiemers & Fiedler, 2007); but see (Puillandre et al., 2011)).

Pons et al. (Pons et al., 2006) introduced a model-based alternative to distance threshold methods. The model, the general mixed Yule-coalescent (GMYC), takes a phylogenetic tree estimated from DNA sequence data and assumes that the branching points in the tree correspond to one of two events: divergence events between species level taxa (modeled by a Yule process (Yule, 1925)), or coalescent events between lineages sampled from within species (modeled by the coalescent (Kingman, 1982b)). Because the rate of coalescence within species is expected to be dramatically greater than the rate of cladogenesis, the GMYC aims to find the demarcation between these types of branching. This model has been shown to be useful in several empirical studies (Barracough et al., 2009, Monaghan et al., 2009a, Papadopoulou et al., 2011, Pons et al., 2006, Pons et al., 2011, Powell et al., 2011). Because it is based on a Likelihood function that directly models evolutionary processes of interest, it provides a means to ameliorate some of the criticisms leveled at threshold methods. Notably, it has allowed for quantification of uncertainty in delimitation of species (Powell, 2012) and avoids the use of non-independent pairwise sequence distances (e.g. in (Puillandre et al., 2011)) as data.

The GMYC model as presently implemented, however, does not account for three potentially large sources of error. First, it is widely recognized that a variety of factors can cause the genealogy from a particular locus to be discordant with the true history of speciation (Maddison, 1997), and the GMYC, like all methods based on a single locus, can be mislead by this discordance. Second, there may be error in the model estimates. Under certain circumstances, the transition from speciation events to coalescent events may be indistinct (e.g., a combination of rapid speciation events and large effective population sizes) causing the model to have a wide confidence interval. A recent implementation by Powell (Powell, 2012) accounts for uncertainty in the threshold parameter and produces model-averaged species limits, but uses point estimates for the other parameters. Finally, phylogenetic error can diminish the accuracy of delimitation results. The GMYC model requires the user to input a point estimate of the phylogenetic tree and inference is premised on the accuracy of this point estimate. Diversity studies using sequence tags, however, typically use relatively short loci that yield estimates of topology and branch lengths that may have high levels of uncertainty. This uncertainty could influence the accuracy of the model.
In order to address the second and third potential sources of error, I introduce a Bayesian implementation of this model with flexible prior distributions in the statistical scripting language R (R Development Core Team, 2011). It accounts for the error in phylogenetic estimation and uncertainty in model parameters by integrating over uncertainty in tree topology and branch lengths and in the parameters of the model via Markov Chain Monte Carlo simulation (MCMC) (Gelman et al., 2009). It produces marginal posterior probabilities for species that are independent of these factors along with output characterizing the posterior distribution that is suitable for downstream analyses of community structure accounting for uncertainty in species limits and phylogeny using R packages such as Picante (Kembel et al., 2010), Vegan (Dixon, 2003), and APE (Paradis et al., 2004). I also conduct simulation tests to evaluate the performance of the model and re-analyze a dataset previously analyzed with the Likelihood version of the model.

2.2. METHODS

2.2.1. Model

Given an ultrametric phylogenetic tree estimated from a set of sequences consisting of multiple species and multiple individuals within species, the GMYC model decomposes the tree into its component waiting times between branching events. These waiting times are the data to be modeled. The parameter of primary interest is a threshold parameter before which the waiting times are modeled according to a Yule process and after which, according to $k$ intraspecific coalescent processes, where $k$ is the number of species with more than one unique sequence sampled. The Likelihood of a waiting time under a Yule model is:

$$L_{(x_i)} = \lambda n_i^p e^{-\lambda n_i^p x_i}$$

where the waiting times ($x_i$) are assumed to be exponentially distributed and a function of: the branching rate ($\lambda$), the number of lineages in the interval ($n_i$), and a rate change parameter that accounts for the possibility of increasing or decreasing diversification rate with time ($p$; Pons et al. (Pons et al., 2006)). The Likelihood ($L$) of a waiting time under the coalescent model is:

$$L_{(x_i)} = \lambda n_i (n_i - 1)^p e^{-\lambda n_i (n_i - 1)^p x_i}$$

where the branching rate ($\lambda$) can be interpreted as $1/N_e \mu$ (where $\mu$ is the per generation mutation rate, or the number of generations per year, depending on the branch length units of the tree) and the rate change parameter ($p$) can be interpreted as population size change over time (Pons et al., 2006).
The GMYC model combines the above models, and the Likelihood of the full model is calculated by assigning lineages in each waiting interval to either the Yule process or one of the coalescent processes such that:

\[
b = \lambda_{k+1} n_{i,k+1}^{p_{k+1}} + \sum_{j=l,k} \lambda_j (n_{ij} (n_{ij} - 1))^p_j
\]

making the full Likelihood of a waiting time:

\[
L_{x_i} = b e^{-bx_i}
\]

where \( k \) indexes the branching processes (1: \( k \) are intraspecific coalescent processes, \( k+1 \) is the Yule process), \( \lambda_{k+1} \) and \( p_{k+1} \) are the branching rate and rate change parameters for the Yule process, and \( \lambda_j \) and \( p_j \) are the branching rate and population size change parameters for the coalescent process. Following Pons et al. (Pons et al., 2006), I constrain \( \lambda_j \) and \( p_j \) to be identical across coalescent processes. The number of lineages assigned to the Yule and coalescent processes in each waiting interval are \( n_{ik+1} \) and \( n_{ij} \), respectively. Assignment of lineages in this case is determined by the selection of a threshold.

Because the sequence data employed in these analyses are typically from short fragments, and thus likely to yield trees with high levels of uncertainty in topology and branch lengths I implemented this model in a Bayesian statistical framework. It eliminates the reliance on point estimates of the phylogeny and model parameters and by estimating the marginal probabilities of the identity of species, allows one to incorporate that uncertainty in downstream analyses. My implementation operates as follows. First, the posterior distribution of trees and branch lengths are characterized using BEAST (Drummond & Rambaut, 2007). Second, a post-burn-in sample of the trees from that posterior distribution is taken, and for each tree, an MCMC simulation is conducted to characterize the posterior distribution of the GMYC model. Following Pons et al., I did not allow the \( \lambda \) parameters to be freely estimated, but used a Maximum Likelihood estimator (Nee, 2001). The non-normalized posterior density function is as follows:

\[
P(T, \lambda_j, \lambda_{k+1}, p_j p_{k+1} | \tau) \propto P(T, \lambda_j, \lambda_{k+1}, p_j p_{k+1}) P(\tau | T, \lambda_j, \lambda_{k+1}, p_j p_{k+1})
\]

where \( \tau \) is the threshold. Because each MCMC evaluates the posterior of the GMYC conditioned on the tree, pooled samples from analyses of many trees sampled from the tree posterior result in a posterior probability distribution of species delimitations conditioned only on the sequence data, the phylogenetic model and the GMYC model.

### 2.2.2. Simulation Testing

I evaluated the utility of this implementation of the GMYC using three simulation experiments. In each, I simulated gene trees from species trees using ms (Hudson, 2002). All species trees contained 50 species and were generated via a Yule process in Mesquite (Maddison & Maddison, 2011), randomly sampling 50 species from a tree of 150 species.
This design was intended to mimic environmental samples of a given taxon, which would not be expected to contain all species in a clade.

In the first experiment I examined the effect of tree depth on model accuracy. I simulated 50 species trees as above and scaled them to four different depths (20N, 40N, 80N, 160N generations, where N is the effective number of diploid individuals in the species). When considering how the results translate to haploid, maternally inherited organellar DNA, the equivalent tree depths are halved (e.g. 10N, 20N…) and N becomes the effective number of females in the population. I then simulated a single gene tree from each species tree at each depth, sampling five alleles per species. For each of these trees I sampled from the posterior for 100,000 generations, discarding the first 10,000 generations as burn-in and thinning every 100 generations, assessed stationarity by examining plots of the parameters by eye, and characterized the posterior distribution of the threshold parameter, which determines the species limits given a tree. Priors on all parameters were uniform distributions; in the case of the threshold parameter, from $\text{U}(2,250)$ and for the $p$ parameters $\text{U}(0,2)$.

In the second experiment I looked at the influence of sampling. The species trees with a depth of 80N from the first experiment were used with four different sampling schemes: 2 alleles per species, 5 alleles per species, 10 alleles per species, and a random number of alleles per species, drawn from a lognormal distribution, with a mean and standard deviation of 1 (an average of 5 alleles per species; approximately 17% of species were represented by singletons). I used the lognormal distribution because it approximates some real species-abundance distributions (Hubbell, 2001) that might be observed in actual species delimitation datasets. I conducted the analyses as in the first experiment.

In the third experiment, I tested the effect of nucleotide sampling and tree estimation on the accuracy of the model (in My simulations, sequence length is directly correlated with the number of variable sites). I selected 10 of the simulated gene trees from 10 species trees scaled to 160N generations for which the confidence intervals in the analysis overlapped the true value of 50 species. I then simulated DNA sequences on those gene trees of 300bp, 600bp, 1200bp and 2400bp using Seq-Gen (Rambaut & Grass, 1997). I assumed $\theta = 0.015$ (corresponding to an $N_e$ of 250,000 and a mutation rate of 1.5% per million generations) and an HKY+G model. I characterized the posterior distribution of trees using the true model of sequence evolution and a strict clock model in BEAST. I pruned all identical sequences and ran BEAST for 10 million generations, discarding the first million as burn-in, at which point all parameters for all replicates had effective sample sizes above 150 and most above 200. I then ran independent GMYC MCMC analyses on 100 trees sampled every 50,000 generations from the BEAST posterior distribution of trees, pooled the results and characterized the marginal posterior distribution of the threshold parameter compared to the distribution produced using the true tree.

2.2.3. Empirical Data Analyses

To illustrate how this implementation of the GMYC could be applied; I downloaded from GenBank and reanalyzed the dataset from Pons et al., the original publication of the GMYC
(Coleoptera: Carabidae: Rivacindela; AJ617921–AJ618351, AJ618352–AJ618766, AJ619087–AJ619548; (Pons et al., 2006). I first pruned each alignment to consist only of unique sequences. Since I am not using a true genealogy sampler (sensu (Kuhner, 2009)), identical sequences will result in many zero-length branches at the tip of the tree, and will cause the model to over-partition the dataset. I then applied a phylogenetic clock model to estimate the posterior distribution of ultrametric trees using BEAST. I partitioned models of sequence evolution by codon, and also by gene when multiple genes were used, applying models of sequence evolution selected by DT ModSel (Minin et al., 2003). I accepted that runs converged when all parameters reached ESS values greater than 200 and checked that posterior distributions differed from priors. I explored the use of different tree priors, as it is conceivable that in cases where branch-length information is lacking, the prior could strongly influence the posterior. For Bayesian GMYC MCMC analyses, I ran each tree for 10,000 generations, discarding the first 1000 as burn-in and sampling every 100 generations. Using 100 trees sampled from the BEAST posterior distribution of trees, this resulted in 9000 samples. I selected this length of Markov-chain because preliminary analyses suggested that stationarity was usually achieved by 1000 generations. I compared the posterior distribution from sampling multiple trees to that from the maximum clade credibility tree and examined the effect of changing the prior on the Yule rate change parameter ($p_{k+1}$). I compared the posterior distribution to the point estimate produced by the Likelihood version of the model, and to the Akaike weights (Anderson, 2008) of each threshold point.

2.3. RESULTS AND DISCUSSION

2.3.1. Simulation Tests

I first tested the influence of tree depth on model performance. When deeper trees are simulated, coalescent and Yule branching processes are expected to occur on more distinct time scales, and thus in general the model should perform better. The influence of tree depth is actually confounded by two issues, however. First, as the tree depth becomes shallower the implied rate of speciation increases because all trees contain 50 species. If the rate of speciation approaches the rate of coalescence within species, then a sharp transition between processes should not be detectable. Second, as the implied rate of speciation increases, more species originate relatively recently. The expected time to coalescence for a diploid, panmictic population is 4N generations. Cladogenic events occurring more recently than this are expected to be increasingly difficult to delimit for two reasons: they are more likely to yield species that are not monophyletic and thus impossible to accurately identify under this model, and the most recent common ancestor (MRCA) of the daughter species is more likely to occur more recently than the threshold point. Assuming species monophyly, the expected time to the MRCA for two species that diverged 4N generations ago is 6N generations. Therefore all probability should be on thresholds older than 4N generations, and most on thresholds older than 6N generations. Again, when considering maternally inherited, haploid, organellar DNA, equivalent times in N generations are halved, and N becomes the effective number of females in the population. This would give an expected time to MRCA of 3N generations.
The results of this test are dramatic (Figures 2.1, 2.2A). There is a clear increase in accuracy as well as a decrease in the range of the 95% highest posterior density interval (HPD) with increasing tree depth. At shorter tree depths, the model’s performance diverged from expectations. When trees are short, larger numbers of species have divergence times younger than 4N generations and thus should not be detectable under the model. Therefore I expected the number of species delimited to be smaller than for deeper trees. I did not observe this. Instead, replicate posterior means became widely scattered around the true value of 50 species. 95% HPDs also did not uniformly increase with decreasing tree depth, instead, a wide distribution of HPDs were observed. I also expected threshold times to have a minimum value of approximately 4N generations. At deeper tree depths this is observed, with 0.13% and 1.6% of MCMC steps sampling thresholds younger than 4N generations for tree depths of 160N and 80N, respectively. However, at 40N and 20N tree depths, 20% and 40% of MCMC steps sampled thresholds younger than 4N generations.

These results indicate that the model performs well under demographic or sampling conditions that result in coalescent and Yule processes occurring on very different time scales. It does not, however, perform optimally when those conditions are not met.

Ideally one would hope that as inference of the threshold point became more difficult, that the 95% HPDs would increase, but still encompass the true value 95% of the time. This is not the case at the 20N and 40N tree depths. HPDs generally become broader, but for
Figure 2.2. The distribution of MCMC samples for each treatment and each replicate within treatments for simulated data. (A) is results from the tree depth simulation, (B) is the results from the allele sampling simulation and (C) is the results from the nucleotide sampling simulation.
increasing numbers of simulation replicates, they fail to encompass the true value. 50 species arising in 40N generations constitutes a very rapid radiation, with an average of 89% of branches in the species tree shorter than the expected population coalescence time of 4N generations. Failure to accurately assess credibility intervals in this case is likely because in this area of parameter space, the GMYC is no longer an accurate approximation of the real branching process in the gene tree. Rather than there being a threshold between coalescent and speciation branching processes, the two processes are intermixed because there is little time for the independent evolution of lineages prior to speciation. Note that these conditions will cause any DNA barcode-based method of species discovery to fail and will also challenge more realistic models utilizing multilocus data and prior information on population assignment.

![Image](image_url)

**Figure 2.3.** Results of a simulation testing the effect of the number of sampled sequences per species on estimates of the threshold parameter and highest posterior density interval (HPD). Both panes show replicates for four sampling schemes: 2, 5, and 10 sequences per species, and a lognormally distributed number of samples per species with a mean of 5. The left pane shows the mean of the posterior for each analysis. The blue line shows the number of species simulated; the red lines show the expected number of detectable species. The right pane shows the 95% HPD. Gray points indicate that the HPD overlapped the true number of species; black ones that it did not.

Next I examined the effect of intraspecific sampling. Because the data points used by the model are waiting times between branching events, I expected that with 50 species, I would not need extremely high sampling to accurately characterize the model, and that the distribution of samples among species would not be particularly important. My expectations were met. I found that sampling of 2 individuals per species yielded poor results (Figures 2.3, 2.2B). Replicate posterior means showed a strong bias towards inference of a large number of species. Sampling of greater than 2 individuals per species provided an improvement in the accuracy of the posterior means, but no change in the 95% HPD range. All sampling schemes greater than 2 individuals per species appeared to yield
similar results, including the more realistic condition of a lognormal distribution of alleles among species in which a large number (~17%) of species are represented by singletons. While delimiting rare species, particularly from single specimens, is a challenge faced by taxonomists (Lim et al., 2011), My results suggest that the GMYC method may efficiently delimit species represented by singletons by calibrating the divergence threshold using data from better represented species.

Finally, I tested the effects of nucleotide sampling and the incorporation of phylogenetic uncertainty. I expected to find wider HPDs with less sequence data, as uncertainty, particularly in branch lengths should be greater. I found a mild reduction in accuracy of the posterior means with up to 600bp of sequence, but after that, posterior means converged on those of the true tree. The 95% HPDs improved with the addition of more sequence, but had not quite converged on those estimated from the true tree, even at 2400bp (Figures 2.4, 2.2C). This suggests that uncertainty in phylogenetic estimation, particularly in typical DNA barcode datasets of ~650bp will contribute substantially to uncertainty in species delimitation.

![Figure 2.4. Results of a simulation testing the effect of the number of nucleotide positions per individual on estimates of the threshold parameter and highest posterior density interval (HPD). Both panes show replicates for four sampling schemes and results from the true tree. The left pane shows the mean of the posterior for each analysis. The blue line shows the number of species simulated; the red lines show the expected number of detectable species. The right pane shows the 95% HPD. Gray points indicate that the HPD overlapped the true number of species; black ones that it did not.](chart)

Three factors that could influence the accuracy of the model that were not explored here: migration, population substructure and selection. Papadopoulou et al. (2008) examined the effects of migration on the formation of detectable GMYC clusters. They simulated datasets under an island model and found that even very low levels of migration (far less than the
Bayesian GMYC was 44 a from the BEAST posterior distribution, the mean number of species estimated by the GMYC so as to provide a direct comparison of the implementations using representative data. The BEAST run converged after 27 million generations and I discarded 2.7 million as burn-in. The estimate of the standard deviation of the lognormal distribution of rates did not overlap 0, so I could not use a strict clock with these data. When using samples of trees from the BEAST posterior distribution, the mean number of species estimated by the Bayesian GMYC was 44 and the 95% HPD ranged from 34 to 57. The rate change parameter

2.3.2 Empirical Analyses

I reanalyzed the empirical data used by Pons et al. to illustrate the original formulation of the GMYC so as to provide a direct comparison of the implementations using representative data. The BEAST run converged after 27 million generations and I discarded 2.7 million as burn-in. The estimate of the standard deviation of the lognormal distribution of rates did not overlap 0, so I could not use a strict clock with these data. When using samples of trees from the BEAST posterior distribution, the mean number of species estimated by the Bayesian GMYC was 44 and the 95% HPD ranged from 34 to 57. The rate change parameter
for the Yule process ranged as high as 1.9. In this model, the fold change in speciation rate from the root to the last speciation event is equal to $n^p/n$ where $n$ is the estimated number of species and $p$ is the Yule rate change parameter. In this case, given 44 species, $p=1.9$ allows for a 30-fold speciation rate increase. I thought this might be unrealistically high (A sampling of three recent papers examining diversification rate shifts yielded a maximum increase from the background rate of approximately 8-fold (Alfaro et al., 2009, Roelants et al., 2007, Stadler, 2011)), and thus re-ran the analysis with the prior distribution set to $U(0,1.2)$, or a maximum 2-fold increase. This minimally influenced the results. I also analyzed the maximum clade credibility tree under both priors, and using the likelihood implementation of the GMYC. I compared the results of Likelihood and Bayesian analyses by calculating Akaike weights for each possible threshold in the Likelihood analysis. Akaike weights are the relative likelihoods of a set of models and thus suited for qualitatively comparing results among these analyses. The Bayesian GMYC analysis of the maximum clade credibility tree yielded a mean of 44 species and a narrower 95% HPD of 38-55 species. When applied to the maximum clade credibility tree, the $U(0,1.2)$ prior did change the results, yielding a posterior mean of 43 species and a 95% HPD of 38 to 51. The Maximum Likelihood analysis resulted in 44 species with a 95% confidence interval of 39 to 51 species. The results of the Likelihood and Bayesian analyses of the maximum clade credibility tree are very similar (Figure 2.4A), particularly when the prior $U(0,2)$ is placed on the Yule parameter, but when sampling trees, the posterior distribution is substantially broader (Figure 2.4B). At least some of this uncertainty stems from variation in topology. Plots of pairwise probabilities of conspecificity demonstrate this (via off-diagonal variation in probability when ordered by a single tree; Figure 2.5). The probability distribution of the number of species in the sample is also wider for the posterior distribution than for the Akaike weights ((Anderson, 2008), Figure 2.4). My Bayesian approach is similar in spirit to the model-averaging approach of Powell (Powell, 2012) in that its goal is to make the inference of species limits and analyses based on them more robust to uncertainty. There are, however, three major differences. First, I take into account phylogenetic uncertainty. As indicated by my results, this is perhaps the most influential difference, although in theory it could also be accounted for using model averaging. Second, my Bayesian method requires the specification of prior knowledge. Depending on a researcher’s comfort with assigning prior probability distributions to the rate and threshold parameters, this is either an advantage or a disadvantage. Finally, the treatment of nuisance parameters (here, the rate change parameters) is fundamentally different. While in the model-averaging approach, inferences are conditioned on Maximum-Likelihood point estimates of nuisance parameters, the Bayesian approach integrates out nuisance parameters, giving marginal probabilities of species limits. These final two differences are intrinsic to Bayesian inference, and researchers choosing among methods will need to consider their choice of statistical paradigm. I note that my confidence intervals in all analyses (including from the Likelihood method) are far wider than those of Pons et al., which were 46-51 species. This is most likely because of the difference in obtaining ultrametric trees. Pons et al. used non-parametric rate smoothing (Sanderson, 2003) on a Maximum Likelihood estimate of the tree. While they achieved sensible results with narrower confidence limits than ours, I nevertheless advocate an approach that samples trees and fits them to a clock model using a parametric method. This allows for a full accounting of uncertainty associated with phylogenetic estimation, albeit at the cost of some precision.
2.4. CONCLUSIONS

My results demonstrate that the Bayesian implementation of the GMYC model is reasonably reliable given two caveats. First, the length of the DNA sequence is important. I found that when I sampled only 300bp, or only 2 alleles per species, that the performance of the model declined strongly. Second, the model is only useful when the underlying history of the species under consideration lies in particular regions of parameter space. Species that have recently diverged, or clades undergoing rapid radiation are unlikely to be identifiable under the model. In the latter case, the model may provide misleading estimates and confidence. Cases such as these, however, may be recognizable because the results may be highly unexpected in the context of other sources of data such as morphology or geography.

My implementation of the model provides two main improvements over the original. First, it allows the specification of prior probabilities on model parameters. It is my experience that very high values of the Yule process rate change parameter sometimes have high likelihood and result in high uncertainty in the threshold parameter (unpublished empirical data). These high values may be biologically unrealistic, and the specification of an informative prior can reduce the posterior probability of those areas and produce a more accurate estimate of diversity. Second, it allows for the characterization of species limits without use of a point estimate of the phylogeny. I know that many datasets are associated with substantial uncertainty owing to limited sequence data collection. The Bayesian GMYC method provides marginal probabilities of species identities and will allow downstream estimates of species diversity and community structure (which are often the goal of environmental sequencing studies; (Powell, 2012)) to account for uncertainty underlying species designations.

An important future direction for this work is to implement the multiple-threshold version of the model proposed by Monaghan et al. (Monaghan et al., 2009b), which can account for greater variation in divergence times and effective population sizes than the model implemented here. It has been shown to provide a better fit to some datasets (Powell, 2012), but will require implementation of a more complex reversible-jump MCMC that allows proposals that change the number of parameters in the model.

It is widely acknowledged that single-locus data are not optimal for the inference of phylogeny, historical demography, or species limits (Galtier et al., 2009, Hey & Nielsen, 2004, Funk & Omland, 2003, Knowles & Carstens, 2007b). Nevertheless, vast amounts of biological diversity remain undescribed at the level of species, and this limits my ability to understand the evolutionary history of my planet and its current ecological functioning. Available alternative means of describing species diversity, either from molecular or morphological data have major drawbacks in that they are time consuming, expensive, or subject to their own biases and inaccuracies. Single-locus data for many groups are currently being generated on a large scale, and I advocate making the best of this data. I believe that under certain conditions, the GMYC model can be useful, and that a Bayesian framework accounting for uncertainty is most appropriate for these data.
Figure 2.5. We compared the new Bayesian implementation with the Likelihood method, with the effect of varying priors, and the inclusion of phylogenetic uncertainty. (A) shows Akaike Weights from the Likelihood method at each threshold point (gray circles), posterior probabilities given a $\text{U}(0,2)$ prior (black triangles) and a $\text{U}(0,1,2)$ prior (black crosses) on the Yule rate-change parameter. All three results were calculated from the maximum clade credibility tree. (B) shows the same results, except the posterior probabilities were calculated by running the analysis on 100 trees sampled from the posterior distribution of trees generated in BEAST.
Figure 2.6. We compare the results of Likelihood and Bayesian analyses of the Pons et al. *Rivacindela* dataset. The phylogenetic tree is the maximum clade credibility tree from BEAST and the clades highlighted in red represent the Maximum Likelihood species limits. The colored table is a sequence-by-sequence matrix. Cells are colored by the posterior probability that the corresponding sequences are conspecific, and allows for the visualization of uncertainty in species limits. Off-diagonal color patterns indicate that some uncertainty in species limits owes to uncertainty in the topology of the phylogeny.
Chapter 3.
Poor Fit To The Multispecies Coalescent Is Widely Detectable In Empirical Data

3.1. BACKGROUND

The introduction of multispecies coalescent models to phylogenetic inference marked a fundamental advance in systematic biology (Degnan & Rosenberg, 2009, Edwards, 2009). These models treat populations, rather than alleles sampled from a single individual, as the focal units in phylogenetic trees. The multispecies coalescent model connects traditional phylogenetic inference, which seeks primarily to infer patterns of divergence between species, and population genetic inference, which has typically focused on intraspecific evolutionary processes. The development of these models was motivated by the common empirical observation that genealogies estimated from different genes are often discordant (e.g. Rokas et al. (2003), Jennings and Edwards (2005)) and the discovery that, if ignored, this discordance can bias parameters of direct interest to systematists, such as the relationships and divergence times among species (Degnan & Rosenberg, 2006, Kubatko & Degnan, 2007, McCormack et al., 2011).

In order to reconcile discordance among gene trees and uncover true species relationships, the first gene tree/species tree models assumed that discordance is solely the result of stochastic coalescence of gene lineages within a species phylogeny (Kubatko et al., 2009, Edwards et al., 2007, Heled & Drummond, 2010, Rannala & Yang, 2003). These approaches estimate topology, divergence times and effective population sizes (except Kubatko et al. (2009)) of the species tree using a model where the probability of a gene tree being discordant with a species tree increases with the ratio of effective population size along a branch to the length of the branch (Rosenberg, 2002, Takahata, 1989). When their assumptions are met, these models are consistent (Liu & Edwards, 2009) across a wide range of divergence histories (Leache & Rannala, 2011, Hird et al., 2010). However, the number of independently segregating loci needed to accurately infer the species tree increases with the above ratio (Hird et al., 2010, Huang et al., 2010).

Coalescent stochasticity, however, is not the only source of gene tree discordance (Maddison, 1997). Selection, hybridization, horizontal gene transfer, gene duplication/extinction, recombination and phylogenetic estimation error can also result in discordance. Maddison (1997) described the product of these disparate genealogical processes as a “cloud” of gene trees. Given that these processes are common (Mallet, 2005, Charlesworth, 2006, Zhang, 2003), I can expect that they will be ubiquitous in new phylogenomic datasets. Unfortunately, beyond a few studies on recombination (Lanier & Knowles, 2012), migration (Eckert & Carstens, 2008) and horizontal gene transfer (Chung & Ane, 2011), I have little idea of the extent to which these factors, unaccounted for, may bias the inference of topology and divergence times in species tree inference. Currently, no method can account for all of these factors. For example, some methods estimate species trees while accounting for gene duplication and extinction (Rasmussen & Kellis, 2012),
some incorporate gene flow (Gerard et al., 2011, Pickrell & Pritchard, 2012), others conduct species delimitation (O’Meara, 2010, Yang & Rannala, 2010), and at least one models discordance without reference to a specific biological process (Ané et al., 2007). Other than a recent model restricted to a three-taxon case (Choi & Hey, 2011), no model accounts for more than two factors, and none accounts for natural selection. Discordance can also be caused by methodological problems, such as errors in species delimitation and mis-specified models of DNA sequence evolution. While the potential for methodological error is well-established, very little work has been to quantitatively estimate its prevalence in empirical studies.

Given that all models must make some simplifying assumptions, and available models make assumptions that are known to be frequently violated, it is imperative to assess the statistical fit of the models to the data. Evidence of poor model fit should encourage researchers to treat phylogenetic estimates with caution and to explore important biological processes that they might not have previously considered. Model checking in this sense seeks to evaluate the absolute fit of models to the data, in order to determine whether any of the models under consideration sufficiently describe the data. This approach complements more widely applied methods of model selection that choose among a set of available models (Goldman, 1993).

Posterior predictive simulation (PPS) is a commonly used method for model checking in a Bayesian framework (Gelman et al., 2009). Although the use of PPS has been advocated for phylogenetic inference (Bollback, 2002, Brown & ElDabaje, 2009, Huelsenbeck et al., 2001, Nielsen & Bollback, 2005, Nielsen, 2002), it has yet to be widely adopted. PPS has been used to show that some common macroevolutionary models are a poor fit to the true process of diversification (Rabosky et al., 2012); that two common population genetic models perform poorly in describing the history of the duck, *Anas strepsera* (Peters et al., 2012) and recommended for use in evaluating models of DNA sequence evolution in the inference of gene trees (Bollback, 2002). However, aside from a recent paper suggesting its use in identifying instances of introgressive hybridization (Joly, 2012) it has not been used to check the fit of multispecies coalescent models.

Here I develop a model checking method in the PPS framework to test the fit of a commonly used Bayesian multispecies coalescent model implemented in *BEAST* (Heled & Drummond, 2010, Drummond & Rambaut, 2007) to 25 published datasets. I then hypothesize about the sources of identified model misspecification and discuss the consequences for inferences based on the multispecies coalescent.

### 3.2. MATERIALS AND METHODS

#### 3.2.1. The Multispecies Coalescent Model in *BEAST*

*BEAST* implements a Bayesian hierarchical model to estimate a species tree with divergence times and effective population sizes from multilocus DNA sequence data. The model hierarchy has three levels (Fig. 1). The bottom level connects the data, a series of DNA sequence alignments for *n* independently segregating loci (*D = d₁, d₂, ..., dₙ*) to their
respective gene trees \((G = g_1, g_2, ..., g_n)\) through the standard phylogenetic likelihood (Felsenstein, 1981):

\[
L(g_i) = P(d_i \mid g_i)
\]

Gene tree likelihoods are conditioned on a chosen model of sequence evolution. Gene tree branch lengths are measured in substitutions per site, the product of mutation rate and time. The second level connects the gene trees \((G)\) to ultrametric coalescent genealogies \((U = u_1, u_2, ..., u_3)\), whose branch lengths are proportional to time, through a molecular clock model:

\[
L(u_i) = P(g_i \mid u_i)
\]

Several molecular clock models, including relaxed clocks (Drummond et al., 2006) and a random local clock (Drummond & Suchard, 2010) are available in *BEAST, each of which make differing assumptions about the distribution of mutation rates among branches in the gene trees (*BEAST incorporates the molecular clock model into the gene tree likelihood, but I depict it separately here for clarity). The third level connects the ultrametric coalescent genealogies \((U)\) with the species tree, including divergence times and effective population sizes \((S)\), through the multispecies coalescent model:

\[
L(S) = P(u_i \mid S)
\]

The likelihood of the multispecies coalescent \((P(u_i \mid S))\) is calculated as the product, across all branches in the species tree, of the probabilities of the coalescent processes within each branch (Rannala & Yang, 2003). The most general form of the model in *BEAST allows the population size on each branch to change linearly, with the constraint that the sum of the population sizes of daughter branches must always equal the population size of their parent (piecewise linear). The marginal posterior probability of the species tree given the data for the full model is then

\[
P(S \mid D) \propto \prod_{i=1}^{n} \int \int P(d_i \mid g_i)P(g_i \mid u_i)P(u_i \mid S)P(S)du_i dg_i
\]

\(P(S)\) is the joint prior probability distribution on the species tree topology, branch lengths and effective population sizes. *BEAST estimates the posterior distribution of the model using a Markov Chain Monte Carlo (MCMC) algorithm.

### 3.2.2. PPS Approach to Checking the Fit of the Multispecies Coalescent

To conduct PPS, one first obtains the joint posterior distribution of parameters for a model (in this case, the *BEAST model, sampled via MCMC), draws sets of parameters from that joint distribution and uses them to simulate data (Fig. 3.1). The simulated datasets form a
The posterior predictive distribution representing reasonable outcomes of the model conditioned on the observed data. One can then compare the empirical data with the posterior predictive distribution using well-chosen test statistics. The ways in which the empirical data do not match the predictive distribution can identify failures of a model to capture important biological processes.

Figure 3.1. A schematic of the *BEAST analysis model and how our PPS model checks relate to it. We check the fit of two parts of the model independently: (a) the fit of the coalescent genealogies to the species tree and (b) the fit of DNA sequence data to the gene trees. Information from the data filters up through the model (black lines), while prior information filters down (dashed lines). Simulated data (gray) are influenced by the empirical data, the model structure, and the prior.

Because *BEAST implements a hierarchical model, and I am most interested in whether the multispecies coalescent component of the model is an appropriate fit the data, my approach to using PPS isolates and checks two levels of the model independently: the multispecies coalescent and the phylogenetic likelihood. I do not try to isolate and check the fit of molecular clock models here.

I check the multispecies coalescent by comparing coalescent genealogies simulated from the posterior distribution of species trees with those estimated in the empirical analysis. I used two test quantities for the comparison, both of which directly assess the fit of the simulated and estimated coalescent genealogies to the estimated species tree: the multispecies coalescent likelihood (i.e., the probability of a coalescent genealogy given the species tree \( P(u|\mathcal{S}) \)), and the number of deep coalescences (Maddison, 1997, Rannala & Yang, 2003). I predicted that for the coalescent likelihood, poor fit would be reflected by individual loci with extremely low probabilities, a low product of probabilities across loci, or an unexpectedly high coefficient of variation of probabilities across loci. For the number of deep coalescences, I expected that poor fit would manifest itself either in individual loci with unexpectedly high or low numbers of deep coalescences, excessively high or low sums of deep coalescences across loci, or a high coefficient of variation across loci. Each of these values measures the degree of discrepancy between gene trees and species trees or across gene trees. In order to generate posterior predictive distributions with expectations of 0, I use test quantities (sensu Gelman et al. (2009)) that are conditioned on particular parameter values sampled from the posterior distribution. To do so, I simulate one set of coalescent genealogies for each draw from the posterior (sampled from the MCMC), calculate test statistics for the coalescent genealogies from that draw as well as the coalescent genealogies simulated from the species tree in that draw, and take their
difference. A 95% highest posterior predictive density interval that does not contain 0 indicates poor fit of the model to the data with respect to that test quantity. For clarity, I refer to all ultrametric gene genealogies estimated or simulated under the model as coalescent genealogies, even if there is evidence that non-coalescent processes influenced them.

Although my primary interest in this study is assessing the fit of the multispecies coalescent, there are two reasons to simultaneously assess the fit of the phylogenetic likelihood. First, poor fit of the multispecies coalescent could be due to poorly fitting models of sequence evolution that result in inaccurate estimates of coalescent genealogies. Second, I speculated that the prior distribution on gene trees induced by the coalescent model may put very low probability on gene tree topologies generated by processes other than stochastic coalescence. For such gene trees, this informative prior could result in estimates that fit the multispecies coalescent model well but strongly disagree with the underlying sequence data.

I check the fit of the phylogenetic likelihood by comparing DNA sequence data simulated from the estimated gene trees \((G)\) with the empirical data. I use four test quantities: (1) the number of variable sites, the (2) multinomial and (3) phylogenetic likelihoods, and (4) the Goldman-Cox (GC) statistic, which is the difference between the multinomial and phylogenetic likelihoods. I expect the number of variable sites to be roughly related to total tree length. I predicted that in some cases of poor fit, the empirical phylogenetic likelihood would be lower than expected based on the posterior predictive distribution of phylogenetic likelihoods since posterior predictive sequence datasets would not conflict with their corresponding gene trees. The multinomial likelihood is the product, across all sites in an alignment, of the frequency of their respective site patterns, and has been frequently used as a test statistic in phylogenetic PPS (Brown & ElDabaje, 2009, Bollback, 2002). The GC statistic has been used in assessing the fit of models of DNA sequence evolution in a likelihood framework (Goldman, 1993, Ripplinger & Sullivan, 2010). The multinomial and phylogenetic likelihoods are expected to converge with very large amounts of data, so when the difference between them is larger than expected, it is a sign that some part (sequence model or tree) of the evolutionary model is a poor fit to the data. It is also worth noting that the phylogenetic and multinomial likelihoods and the GC statistic are all strongly correlated with the number of variable sites in an alignment. Therefore, inaccurate estimates of tree length, even absent other reasons for poor fit, can lead to deviation in these statistics.

### 3.2.3. Empirical Data

I obtained 25 datasets from Genbank or directly from the authors (Table 3.1). With a few exceptions, I avoided publications that dealt with hybridization directly, or those that
Table 3.1. Summary of empirical datasets used in this analysis, including number of OTUs (O), number of alleles (A), number of loci (L) and inclusion of organellar DNA (OD).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reference</th>
<th>Source</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Level</th>
<th>O</th>
<th>A</th>
<th>L</th>
<th>OD</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Thomomys</em></td>
<td>Belfiore et al. 2008</td>
<td>BEAST help</td>
<td>Rodentia</td>
<td>Geomyidae</td>
<td><em>Thomomys</em></td>
<td>population</td>
<td>9</td>
<td>26</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td><em>Manacus</em></td>
<td>Brumfield et al. 2008</td>
<td>web</td>
<td>Passeriformes</td>
<td>Pipridae</td>
<td><em>Manacus</em></td>
<td>species</td>
<td>5</td>
<td>47</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td><em>Sceloporus</em></td>
<td>Leaché 2009</td>
<td>TreeBASE</td>
<td>Squamata</td>
<td>Iguanidae</td>
<td><em>Sceloporus</em></td>
<td>population</td>
<td>9</td>
<td>21</td>
<td>8</td>
<td>-</td>
</tr>
<tr>
<td><em>Phylomedusa</em></td>
<td>Brunes et al. 2010</td>
<td>author</td>
<td>Anura</td>
<td>Hylidae</td>
<td><em>Phylomedusa</em></td>
<td>species</td>
<td>5</td>
<td>27</td>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>Phocidae</td>
<td>Fulton and Strobeck 2010</td>
<td>author</td>
<td>Pinnipedia</td>
<td>Phocidae</td>
<td>genus</td>
<td>24</td>
<td>39</td>
<td>16</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td><em>Ctenosaura</em></td>
<td>Paschkink et al. 2010</td>
<td>author</td>
<td>Squamata</td>
<td>Iguanidae</td>
<td><em>Ctenosaura</em></td>
<td>species</td>
<td>5</td>
<td>27</td>
<td>4</td>
<td>+</td>
</tr>
<tr>
<td>Cettiidae</td>
<td>Alstrom et al. 2011</td>
<td>author</td>
<td>Passeriformes</td>
<td>Cettiidae</td>
<td>genus</td>
<td>29</td>
<td>97</td>
<td>4</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Locustellidae</td>
<td>Alstrom et al. 2011</td>
<td>author</td>
<td>Passeriformes</td>
<td>Locustellidae</td>
<td>genus</td>
<td>41</td>
<td>41</td>
<td>5</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td><em>Brachycephalus</em></td>
<td>Clement- Carvalho et al. 2011</td>
<td>genbank</td>
<td>Anura</td>
<td>Brachycephalidae</td>
<td><em>Brachycephalus</em></td>
<td>species</td>
<td>15</td>
<td>15</td>
<td>4</td>
<td>+</td>
</tr>
<tr>
<td><em>Buarremon</em></td>
<td>Florez-Rodriguez et al. 2011</td>
<td>author</td>
<td>Passeriformes</td>
<td>Emberizidae</td>
<td><em>Buarremon</em></td>
<td>species</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>+</td>
</tr>
<tr>
<td>Psittacidae</td>
<td>Joseph et al. 2011</td>
<td>author</td>
<td>Psittaciformes</td>
<td>Psittacidae</td>
<td>genus</td>
<td>27</td>
<td>27</td>
<td>8</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td><em>Sistrurus</em></td>
<td>Kubatko et al. 2011</td>
<td>TreeBASE</td>
<td>Squamata</td>
<td>Crotalidae</td>
<td><em>Sistrurus</em></td>
<td>population</td>
<td>8</td>
<td>52</td>
<td>19</td>
<td>+</td>
</tr>
<tr>
<td><em>Certhia</em></td>
<td>Manthey et al. 2011</td>
<td>author</td>
<td>Passeriformes</td>
<td>Certhiidae</td>
<td><em>Certhia</em></td>
<td>population</td>
<td>11</td>
<td>141</td>
<td>20</td>
<td>-</td>
</tr>
<tr>
<td><em>Lepus</em></td>
<td>Melo-Ferreira et al. 2011</td>
<td>TreeBASE</td>
<td>Lagomorpha</td>
<td>Leporidae</td>
<td><em>Lepus</em></td>
<td>species</td>
<td>13</td>
<td>55</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td><em>Myotis</em></td>
<td>Salicini et al. 2011</td>
<td>author</td>
<td>Chiroptera</td>
<td>Vespertilionidae</td>
<td><em>Myotis</em></td>
<td>species</td>
<td>6</td>
<td>49</td>
<td>7</td>
<td>+</td>
</tr>
<tr>
<td><em>Aliatypus</em></td>
<td>Satler et al. 2011</td>
<td>author</td>
<td>Araneae</td>
<td>Antrodiaetidae</td>
<td><em>Aliatypus</em></td>
<td>population</td>
<td>13</td>
<td>102</td>
<td>5</td>
<td>+</td>
</tr>
<tr>
<td><em>Herpystichum</em></td>
<td>Tepe et al. 2011</td>
<td>author</td>
<td>Solanales</td>
<td>Solanaceae</td>
<td><em>Herpystichum</em></td>
<td>species</td>
<td>12</td>
<td>18</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td><em>Liolaemus</em></td>
<td>Carmago et al. 2012</td>
<td>author</td>
<td>Squamata</td>
<td>Liolaemidae</td>
<td><em>Liolaemus</em></td>
<td>population</td>
<td>16</td>
<td>48</td>
<td>20</td>
<td>+</td>
</tr>
<tr>
<td><em>Ursus</em></td>
<td>Hailer et al. 2012</td>
<td>supplemental</td>
<td>Carnivora</td>
<td>Ursidae</td>
<td><em>Ursus</em></td>
<td>population</td>
<td>7</td>
<td>90</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td><em>Ettheostoma</em></td>
<td>Harrington and Near 2012</td>
<td>TreeBASE</td>
<td>Perciformes</td>
<td>Percidae</td>
<td><em>Ettheostoma</em></td>
<td>population</td>
<td>5</td>
<td>28</td>
<td>4</td>
<td>+</td>
</tr>
<tr>
<td><em>Malurus</em></td>
<td>Lee et al. 2012</td>
<td>TreeBASE</td>
<td>Passeriformes</td>
<td>Maluridae</td>
<td><em>Malurus</em></td>
<td>population</td>
<td>25</td>
<td>84</td>
<td>17</td>
<td>+</td>
</tr>
<tr>
<td><em>Bufo</em></td>
<td>Recuero et al. 2012</td>
<td>author</td>
<td>Anura</td>
<td>Bufonidae</td>
<td><em>Bufo</em></td>
<td>species</td>
<td>4</td>
<td>232</td>
<td>5</td>
<td>+</td>
</tr>
<tr>
<td><em>Tamias</em></td>
<td>Reid et al. 2012</td>
<td>author</td>
<td>Rodentia</td>
<td>Sciuridae</td>
<td><em>Tamias</em></td>
<td>species</td>
<td>22</td>
<td>232</td>
<td>5</td>
<td>+</td>
</tr>
<tr>
<td><em>Sitta</em></td>
<td>Walstrom et al. 2012</td>
<td>author</td>
<td>Passeriformes</td>
<td>Sittidae</td>
<td><em>Sitta</em></td>
<td>population</td>
<td>7</td>
<td>112</td>
<td>17</td>
<td>+</td>
</tr>
<tr>
<td>Cheirogaleidae</td>
<td>Weisrock et al. 2012</td>
<td>author</td>
<td>Primates</td>
<td>Cheirogaleidae</td>
<td>genus</td>
<td>20</td>
<td>65</td>
<td>12</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
excluded known introgressed loci. I also avoided publications whose express goal was species delimitation, because errors in species assignment are a clear violation of the multispecies coalescent. Each dataset was analyzed using the models of sequence evolution provided in the original manuscript with the exception of models requiring both a proportion of invariable sites and gamma-distributed rates across sites (RAS), because I occasionally observed problems with convergence when using them together. In those cases I used only gamma-distributed RAS. I retained any intra-locus partitioning schemes, and for the phylogenetic likelihood model checks treated each locus subset individually. I ran each dataset twice for at least as long as in the original publication. To conduct PPS I excised the first 10% of MCMC steps as burn-in, combined both runs, and thinned them to ~2000 MCMC samples. All analyses were conducted using custom scripts in the statistical language R (R Development Core Team, 2011), available on NMR’s website (https://sites.google.com/site/noahmreid/) in tandem with ms (Hudson, 2002), Seq-Gen (Rambaut & Grass, 1997), ape (Paradis et al., 2004), and phangorn (Schliep, 2011). I selected four datasets that fit the model poorly in some way and subjected them to further analysis. For two (Tamias and Cheirogaleidae) I removed single loci that showed poor fit, re-analyzed the rest of the data and compared the model estimates. For the other two (Ursus and Sistrurus) I eliminated the multispecies coalescent level of the model hierarchy, fit completely independent trees and clock models, and compared the gene tree estimates.

3.3. RESULTS

3.3.1. Datasets and Analyses

I analyzed 25 datasets from papers spanning 12 orders of Eukaryotes (Table 3.1). The average number of operational taxonomic units (OTUs) was 13.7, the average number of alleles per dataset was 67.2, and the average number of independently segregating loci was 9.6. Seventeen datasets utilized organellar as well as nuclear DNA. For *BEAST analyses, nearly all datasets had ESS values of over 200 for all parameters. Markov chains for a few datasets mixed poorly, but if parameter estimates from 2 independent runs were very similar after 200 million generations, I included them anyway. Results of my *BEAST analyses were consistent with published results for each dataset.

3.3.2. Overview of Results

At the level of the coalescent genealogies, I found evidence of poor model fit in 4 datasets (16%) considering all test quantities. Seven total loci across data sets deviated from expectations (2.9%; Table 3.2). Two of those datasets showed poor fit at only one locus (Certhia and Cheirogaleidae). Three of these datasets also had poor fit at the DNA sequence level (Aliatypus, Certhia and Tamias), although not always for the same loci (50%). I was unable to identify systematic trends in the observed deviations. Tamias had one locus with an excess of deep coalescences and low probability (mitochondrial Cyt b) and one with a deficit of deep coalescences (ACR; Fig 3.2). Aliatypus had three nuclear loci with deficits of deep coalescences and low probabilities (partial results in Fig. 3.3). Certhia and Cheirogaleidae each had one locus with an excess of deep coalescences, but no deviation in probability of coalescent genealogy.
Table 3.2. Summary of the results of the tests of the fit of the coalescent genealogies multispecies coalescent. Coal. Like refers to the coalescent likelihoods and Deep Coal. refers to the number of deep coalescence events.

| Dataset  | Loci | P(\text{u}|\text{S}) | \text{Deep coalescences across loci} | Individual Loci | % poorly fitting |
|----------|------|---------------------|----------------------------------|-----------------|------------------|
|          |      | Sum | Coef. | Sum | Coef. | Coal. Like. | Deep Coal. | Total | loci |
| Aliatypus| 5    | 1   | -     | 1   | -     | 1   | 3   | 3    | 3    | 60.0 |
| Certhia  | 20   | -   | -     | -   | -     | -   | 1   | -    | 1    | 5.0  |
| Cheirogaleidae | 12 | -   | -     | -   | -     | -   | -   | -    | 1    | 8.3  |
| Tamias   | 5    | 240 | 1     | 0   | 0     | 1   | 2   | 0    | 1    | 4    |
|          |      |     |       |     |       | 2   | 4   | 0    | 4    | 3    |
|          |      |     |       |     |       | 2   | 4   | 3    | 7    | 2.9  |

At the level of the sequence data I found evidence of poor model fit in 20 datasets (80%) with 45 partitions and 44 loci (16.9% and 18.3% respectively) deviating from expectations (Table 3.3). Deviations were apparent using all test statistics, but the GC statistic was the most frequent indicator (33/45 partitions). Again, there were no obvious systematic trends in the deviations, except that the empirical GC statistics tended to be smaller than expected (60% were smaller), although this was not significant under a binomial test.

While I have low sample size, mitochondrial genes (mtDNA) do not appear to be overrepresented among the loci that poorly fit in the coalescent genealogy tests. Two of the seven poorly fitting loci were mitochondrial (binomial test: successes=2, trials=7, probability=17/240, p=0.08). However, eleven of 44 of the loci that poorly fit in the sequence data tests were mitochondrial, a significant excess (binomial test: successes=11, trials=44, probability=17/240, p= 0.0002).

3.3.3. Case Studies—Removal of Genes Poorly Fitting Coalescent Assumptions

The *Tamias* and Cheirogaleidae datasets each contained one locus that fit poorly at the coalescent genealogy level. In order to determine whether data that do not fit the model affect the outcome of the analysis I elected to remove these loci (the mtDNA locus Cyt b from *Tamias* and the nuclear locus ABCA1 from Cheirogaleidae) from their respective datasets and reanalyze them. Both loci had an excess of deep coalescences, and the *Tamias* Cyt b locus also showed low probability given the species tree. After removal of Cyt b from *Tamias*, I did not observe substantial change in the posterior means of the relative mutation rate parameters. The species tree root height was 11% lower when Cyt b was included and the gene trees were on average 3% higher, although the 95% HPDs overlapped substantially (the result was the same if the gene trees were scaled to the same mutation rate or unscaled). By contrast, the scale parameter of a gamma-distributed prior on effective population sizes across branches in the species tree (the species.popMean hyperparameter), was 3.6 times larger in the dataset including Cyt b and the 95% HPDs for
the two analyses did not overlap. Most notably, the tree topologies between the two runs changed drastically, returning incompatible, highly supported nodes (Fig. 3.4). There was no evidence of poor fit using my PPS analysis after Cyt b was removed.

After removal of ABCA1 from the Cheirogaleidae dataset, I similarly observed few changes in the relative mutation rate parameters or root heights. The single exception was the Adora locus, whose mean relative mutation rate parameter was 1.5 times higher and whose unscaled root height was 1.4 times larger when ABCA1 was excluded. These changes apparently made the Adora tree a poorer fit to the sequence data, as all test statistics became more extreme, but none crossed the p=0.05 threshold. For all other loci, the mean root heights were 5% higher, the species tree root height was 10% higher and the popMean parameter was 10% larger when ABCA1 was included. The species tree relative branch lengths, topology and posterior probabilities were unchanged when ABCA1 was removed.

Figure 3.2. PPS model check results for Tamias. (A) Key for the figure. Distributions of test statistics are shown, where the dashed line is the expectation (0), and gray bars indicate the boundaries of the 95% and 99% highest posterior predictive density intervals. (B) and (C) give results for the test of the fit of coalescent genealogies; boxes with bold black lines indicate poor fit. A single gray bar indicates a one-tailed test. (B) Coalescent genealogy tests for all loci; (C) Coalescent genealogy tests for the two loci that fit poorly, ACR and Cyt b; (D) Sequence data tests for the same two loci.
3.3.4. Case Studies—Reanalysis with Independent Gene Trees

The *Sistrurus* and *Ursus* datasets did not show poor fit at the genealogy level, but each had several loci (3/20 and 4/14, respectively) that fit poorly at the sequence level. All poorly fitting loci had greater GC statistics than predicted. One locus from each dataset fit poorly using all test quantities. In order to test the hypothesis that poor fit at these loci stems from the multispecies coalescent-induced prior on the gene trees, I reanalyzed the data with no species tree and unlinked all parameters. As a result, all signs of poor fit vanished.

When comparing gene genealogy estimates between the analyses, results differed between the two datasets. There was no obvious reason why the independent-gene-trees model should be a better fit to two of the three *Sistrurus* loci, as they were primarily unresolved in both analyses. The third gene, Fibrinogen beta chain (FGB), however, had one strongly supported conflict among analyses: in the independent-gene-trees model, the placement of alleles from the outgroup taxon, *Agiostrodon contortrix* were polyphyletic with respect to the ingroup, instead of as sister to the other outgroup taxon, *A. piscivorus*. Trees sampled
Table 3.3. Summary of the results of the tests of the fit of the sequence data to the phylogenetic Likelihood.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Loci</th>
<th>Partition:</th>
<th>Variable sites</th>
<th>Phylogenetic likelihood</th>
<th>Multinomial likelihood</th>
<th>GC statistic</th>
<th>% poorly fitting partitions</th>
<th>% poorly fitting loci</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Aliatypus</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Brachycephalus</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Buarruemon</td>
<td>7</td>
<td>7</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bufo</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Certhia</td>
<td>20</td>
<td>20</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cettiidae</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cheiroleiidae</td>
<td>12</td>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ctenosura</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Etioestoma</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Herpystichum</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Lepus</td>
<td>14</td>
<td>14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Liolaemus</td>
<td>20</td>
<td>20</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Locustellidae</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Malurus</td>
<td>17</td>
<td>17</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Manacus</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Myotis</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Phocidae</td>
<td>16</td>
<td>40</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Phylomedusa</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Psittacidae</td>
<td>8</td>
<td>8</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Scoloporus</td>
<td>8</td>
<td>8</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Sistrurus</td>
<td>19</td>
<td>19</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Sitta</td>
<td>17</td>
<td>17</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Tamias</td>
<td>5</td>
<td>5</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Thomomys</td>
<td>7</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ursus</td>
<td>14</td>
<td>14</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>240</strong></td>
<td><strong>267</strong></td>
<td><strong>10</strong></td>
<td><strong>12</strong></td>
<td><strong>10</strong></td>
<td><strong>12</strong></td>
<td><strong>13</strong></td>
<td><strong>13</strong></td>
</tr>
</tbody>
</table>
in the MCMC for *Sistrurus* loci that fit the phylogenetic likelihood poorly tended to have very similar likelihoods whether the species tree was enforced or not.

Poorly fitting *Ursus* loci, by contrast, often had obvious topological differences across analyses. The locus nr11080, for example, yielded a paraphyletic *Ursus arctos*, with the bears from Admiralty, Baranof and Chichagof (ABC) islands being more closely related to *U. maritimus*. Under the species tree model, the ABC haplotypes were the sister group to *U. maritimus*, but under the independent-gene-trees model, they were scattered within the *U. maritimus* clade. Also in contrast to *Sistrurus*, each of the 4 poorly fitting loci showed average improvements of ~30 log-likelihood units for trees sampled during the MCMC under the independent-gene-trees model, while the other 10 loci improved ~5 log-likelihood units.

### 3.4. DISCUSSION

Gelman et al. (2009) argue that model checking is an essential part of Bayesian data analysis, on par with the initial formulation of models and fitting those models to data. Here I have developed the first general model-checking method for a commonly used multispecies coalescent phylogenetic inference model, and my results show that poor fit between model and data is detectable in a majority of sampled datasets. At the level of coalescent genealogies, it is relatively straightforward to suggest biological explanations for poor fit between processes generating gene tree topologies and the specified coalescent model. By contrast, poor fit at the level of the DNA sequence data could plausibly be explained by a variety of forms of model misfit. While the test quantities used here do not uniquely identify the biological processes that violate the multispecies coalescent model, the identification of loci that fit poorly in combination with relevant biological and

---

**Figure 3.4. Tamias species trees estimated in *BEAST* using 5 markers. (A) includes the mitochondrial locus Cyt b and (B) excludes it. Circles indicate 0.95 posterior probability. Red red indicates nodes are incompatible with the opposite tree.**
geographical context can suggest of directions for future analyses. Below, I discuss empirical examples of poor fit, their observed consequences, and possible approaches to take when poor fit is detected.

### 3.4.1. Empirical Examples of Poor Fit to the Multispecies Coalescent Model

Recent hybridization is a likely explanation for poor model fit when there is an excess of deep coalescences and closely related haplotypes are shared among species. The *Tamias* data provide a good example. Previous studies have detailed extensive mtDNA introgression between non-sister species in the genus (Reid et al., 2012, Good et al., 2008, Reid et al., 2010). In these analyses, the mtDNA is identified as having an excess of deep coalescences and low coalescent genealogy probability (Fig. 3.2). When the *Tamias* genealogies themselves are examined, statistically supported discordance among genes and species non-monophyly in the species tree are evident. When the mtDNA is removed and the data re-analyzed, all signs of poor fit, including at one nuclear locus that showed a deficit of deep coalescences, disappear. Additionally, the species tree topologies change drastically. Two species appearing as sister taxa in the tree including mtDNA are distantly related in the nuclear-only tree (Fig. 3.4). I expect recent hybridization to impact analyses because without modeling it, species divergences must always post-date the divergence of shared gene lineages.

Unmodeled population structure can also presumably cause poor model fit, as exemplified by the *Aliatypus* data. There is no supported discordance among loci in relationships between populations in this group, and yet poor fit is evident in 3 out of 5 loci, and also in the summaries across loci (Fig. 3.3). These statistics indicate a deficit of deep coalescences (1 locus also has low probability). I hypothesize that high genetic diversity within OTUs, as a result of unmodeled population structure, leads to overestimation of effective population size, and thus an overprediction of the amount of stochastic lineage sorting. This is consistent with what is known about *Aliatypus* biology: these are terrestrial spiders that live in subterranean burrows with limited dispersal ability (Coyle & Icenogle, 1994). Geographically close populations often have highly divergent mtDNA haplotypes, and those haplotypes tend to have highly restricted distributions (Satler et al., 2011).

The remaining two datasets that showed poor fit at the coalescent genealogy level, Cheirogaleidae and *Certhia*, each had one locus with an excess of deep coalescences, but did not have low probability. These issues are harder to resolve. An examination of the Cheirogaleidae locus, ABCA1, yielded a fairly well resolved genealogy that contained some species that were non-monophyletic, each with unique, highly divergent haplotypes not shared with other species. This could be a result of ancient hybridization, balancing selection, or gene duplication and extinction. Gene duplication and extinction seems unlikely, as the issue might have been expected to manifest itself in patterns of heterozygosity and been resolved through cloning. Distinguishing balancing selection from ancient hybridization would require analyses of Dn/Ds ratios and more detailed studies of population structure.
It is also possible for systematic error in coalescent genealogy estimation to cause poor model fit at this level, even if the model is correct. For this to be the case, misspecified models of sequence evolution and/or molecular clock models would have to prefer incorrect trees strongly enough to overcome the prior probability distribution on coalescent genealogies induced by the multispecies coalescent. This may be most likely in cases when misspecification is very serious (e.g. sequences with secondary structure where sites are non-independent) or when there is a lot of information and high complexity (e.g. mtDNA). Both Aliatypus and Tamias mtDNA are found to fit the model poorly at both levels, but I believe the genealogical patterns causing poor fit in those systems are clear enough to adequately explain observed patterns of misfit.

At the level of DNA sequence data, sources of poor fit are harder to distinguish. There are two main possibilities. First, models of sequence evolution and molecular clock models could be misspecified. Second, coalescent genealogies could be a poor fit to the multispecies coalescent, but the prior distribution on coalescent genealogies induced by the model might overwhelm the signal in the data. This could strongly favor topologies and branch lengths that fit the multispecies coalescent, but are a poor representation of the sequences. I speculate that both effects are evident in my analyses. The overrepresentation of mtDNA loci among those that poorly fit at this level may result from the first effect. In particular, mtDNA contains a large number of variable sites and thus more phylogenetic signal than most autosomal loci. It seems less likely that prior probability distributions on gene genealogies could push poorly fitting loci away from their preferred topologies. In support of this idea, of two datasets included here that partitioned their mtDNA, both had some partitions that fit the model and some that did not (e.g. Fig. 3.3). If the tree was the problem, I might expect all partitions of the same locus to fit poorly.

By contrast, I expect much of the poor fit I observe at nuclear loci to be a result of the second factor. My analyses of Sistrurus and Ursus support this idea. I would not expect to see improved fit when the species tree portion of the model was eliminated if the models of sequence evolution and mutation rate variation were causing the problem. Additionally, in Ursus I see changes in topology and posterior probabilities that are consistent with strong prior sensitivity. Interestingly, the pattern observed at the locus mentioned above, nr11080, is likely to be a previously unacknowledged signal of hybridization at nuclear loci in this dataset. *U. maritimus* mtDNA is thought to be a result of a fixed introgression from bears related to those from the ABC islands, so it is unsurprising to discover that the ABC bears may harbor DNA that has moved in the other direction (Edwards et al., 2011, Hailer et al., 2012, Miller et al., 2012).

It is important to note that there are potentially other non-examined sources of discordance, including inaccurate taxonomic knowledge of species limits. If individuals are inaccurately assigned to OTUs in a species tree analysis, one would expect that PPS would indicate that the species tree is a poor fit to the data, although it would likely be difficult to identify the cause of this poor fit.
3.4.2. Consequences of Poor Fit

My analyses suggest that poor fit to the multispecies coalescent model can mislead inference in empirical studies. In the case of recent hybridization, the consequences may be severe, as species divergences are forced to post-date gene divergences. For example, when the mitochondrial DNA were removed from *Tamias*, the species tree topology changed drastically. The topologies from both analyses conflicted at strongly supported nodes and two recently hybridizing species, *T. amoenus* and *T. ruficaudus* went from sister taxa, to being only distantly related (Fig. 3.4). Unexpectedly, the nuclear DNA did not fit either model poorly, which may be a sign that the data are insufficiently informative or that my approach does not have high power.

When topological conflict among coalescent genealogies is the result of ancient hybridization, balancing selection, or gene duplication and extinction, the consequences may be less severe. It seems possible that such conflicts may be resolvable by invoking deep coalescence. For example, when I removed the ABCA1 locus from the Cheirogaleidae, the changes to the topology, branch lengths, and posterior support were minimal. This flexibility of the multispecies coalescent may also make such processes difficult to detect using my framework. If detecting such processes were my primary goal, rather than identifying instances of poor model fit, development of new test quantities might be necessary.

If the coalescent genealogies themselves are of interest, my results suggest that the prior probability distribution induced by the multispecies coalescent can be quite informative. This is of concern because some recent studies have used species tree-based models to improve estimates of gene genealogies (Wu et al., 2012, Åkerborg et al., 2009). This may be useful when the prior distribution is appropriate, but if important processes are unmodeled, my results suggest analyses may be misled.

3.4.3. Strategies for Dealing with Poor Fit

When poor fit is detected, there are two main strategies that can be used to ameliorate it. First, remove data that violate the multispecies coalescent model. Many of the processes causing poor fit may be heterogeneous across the genome. Not every gene family is expected to be the focus of bouts of duplication and extinction, or intense selection, and not every region of the genome will introgress with equal ease. If a relatively small number of loci appear to fit poorly, it is easy to remove them and re-analyze the data. Second, the biological processes that generate variation in gene tree topologies should be explicitly modeled, as should relevant dynamics of molecular evolution. Increasingly complex multispecies coalescent models are being implemented, but there are tradeoffs. Some examine gene duplication and extinction (Rasmussen & Kellis, 2012) or migration (Pickrell & Pritchard, 2012) but cannot estimate divergence times.

I believe my results suggest that a concatenation approach to analyzing multilocus datasets with extensive inter-locus heterogeneity in topology may be even more perilous than simulation studies have shown (Kubatko & Degnan, 2007; Edwards, 2009). Those studies
assume that the only source of discordance among loci is coalescent stochasticity. Here I show that other factors contribute to heterogeneity among gene trees, exacerbating the issue.

3.5. CONCLUSIONS

The very act of data analysis requires researchers to make assumptions about the evolutionary processes that have shaped the data. I demonstrate that not all empirical data are consistent with the assumptions of the multispecies coalescent model. As the number and breadth of phylogenetic methods increases, it is far better to assess the fit between models and the data to which they are being applied than it is to assume that a certain method is appropriate to a given data set. Phylogenetics is no longer a data-poor enterprise, and I can afford to be choosy with the data that are analyzed. Posterior predictive simulation is an effective method for identifying data that violate important assumptions of analytical models.
Chapter 4.
Integrating Ecological And Evolutionary Methods For Inference Of Species Diversity And Community Structure: An Example Using A Tropical Alpine Fly Community

4.1. INTRODUCTION

While the fields of ecology and evolutionary biology share the common goal of elucidating the processes that generate regional patterns of biodiversity, they have historically taken disparate approaches in achieving this goal. Phylogeography, a subdiscipline of evolutionary biology, has traditionally concerned itself with inferring historical demography within and between closely related species (Avise et al., 1987, Hickerson et al., 2010) by inferring population divergence times, effective population sizes and rates of change, and interpopulation migration rates (Knowles & Maddison, 2002, Hey & Nielsen, 2007, Beerli & Palczewski, 2010, Minin et al., 2008). The field seeks to identify the impact of climatological and geological processes on the generation and distribution of evolutionary lineages. A natural expansion of this work is comparative phylogeography, which endeavors to understand the causes of incongruity in these coarse-grained demographic models among species with overlapping distributions (Huang et al., 2011, Hewitt, 2000, Taberlet et al., 1998).

Ecological approaches to understanding regional patterns of biodiversity include methods from macro- and community ecology (Leibold et al., 2004, Brown, 1995). These typically treat species as the fundamental unit of analysis, and take their limits as given by taxonomists. They seek to rigorously quantify the distribution of species richness, abundances and turnover across space and to identify correlational models between these quantities and explanatory environmental factors (Gotelli & Colwell, 2001, Jost, 2006, Willig et al., 2003, Hubbell, 2001). It has recently been acknowledged that species diversity may not be at an equilibrium distribution governed by ecological factors, and as such, historical factors may be an important determinant of contemporary patterns (Ricklefs, 2008). This has led ecologists to incorporate historical and phylogenetic information into various analyses, ranging from quantifying the geographic distribution of diversification rates to analyses of community phylogenetic structure to Hubbell’s unified neutral model (Jetz et al., 2012, Webb et al., 2002, Hubbell, 2001).

These disparate research communities have in part converged on addressing very similar questions, but the explanations they consider and the methods they apply are often profoundly different. For example, phylogeography has moved into an era in which many studies use almost exclusively process-based parametric evolutionary models (Nielsen & Beaumont, 2009, Lemey et al., 2009). This provides the advantage that results are intuitively interpretable as direct estimates of parameters of interest. Due to computational demands, however, these models, especially at the comparative level, are limited to relatively simple scenarios (Carnaval et al., 2009). They are also limited to examining
largely codistributed taxa and thus do not completely address the larger ecological and evolutionary questions. Ecologists, by contrast tend to use non-parametric approaches (Chao et al., 2000, Chao et al., 2005, Gotelli, 2000). These methods can be used to discover very complex patterns of community structure and covariation with environmental factors, but because they do not explicitly model community dynamics, the interpretation of these patterns can be limited (Gotelli et al., 2009).

Bridging the gap between these methodological approaches to explaining regional patterns of diversity has the potential to improve the pace and value of contributions to describing the world’s biodiversity by providing estimates of the composition, origins and connections between regional biotic communities. Here I explore the integration of these methods, using DNA sequence-based species delimitation, phylogenetic estimation, analyses of population genetic structure, ecological diversity statistics, and neutral models as a means to describe in broad strokes some of the factors governing the composition of a virtually unstudied tropical alpine insect community, the Diptera (Arthropoda:Insecta) of Central American páramos.

4.1.1. Study System

Central American páramos occur in the Cordillera Volcánica of Costa Rica and the Cordillera de Talamanca in Costa Rica and Panamá (Fig. 4.1A). These mountain ranges are areas of high endemism (Olson et al., 2001) and in their alpine reaches (above ~3200m) are fragmented páramo sky-islands. Páramo is a neotropical alpine habitat characterized by high levels of insolation, wide daily temperature fluctuation with frequent freezing temperatures and little seasonality (Luteyn, 1999). While the Talamancas are known for montane forest endemism at middle elevations, these small patches of alpine páramo also house a unique biota, including two endemic plant genera, a bird, and a now-extinct frog (Sklenar et al., 2005, Savage & Bolaños, 2009, Stiles & Skutch, 1989). Lineage diversification is known within and among these ranges in amphibians (García-París et al., 2000, Streicher et al., 2009) and at least one bird (Stiles & Skutch, 1989). This biota, as in other alpine systems (Moritz et al., 2008) is threatened by climate change through the process of mountaintop extinction.

The Talamancas are thought to be young, reaching elevations capable of sustaining alpine habitat between 1-2mya (Driese et al., 2007) but see recent controversies over the age of the Isthmian land bridge (Montes et al., 2012)). They are also highly isolated. The nearest extant alpine habitats are located in southern Guatemala (~900km distant) and Colombia (~650km distant). Intervening mountain ranges are low (~2000m) and separated by deep valleys (<1000m). During the Pleistocene climate oscillations, montane habitat zones fluctuated over 1000 vertical meters lowering the páramo zone to nearly 2000m elevation (Islebe & Hooghiemstra, 1997). This would have resulted in the physical connection of many currently isolated patches while others remained isolated, but the alpine system as a whole remained isolated from northern Central and South America.

The Páramo entomofauna is poorly known and not included in published accounts of several prominent Costa Rican biodiversity inventories (Colwell et al., 2008, Janzen et al.,
2009), but is thought to be far less diverse than that of forested lower elevations (Janzen, 1973, Kappelle & Horn, 2005). Owing to an extreme climatic regime, it is likely to be highly endemic, as observed in South American páramo systems (Campbell et al., 2006). I have focused on the Diptera of these habitats for several reasons. First, the Diptera represent one of the most poorly studied of the hyperdiverse insect orders, particularly in the tropics (Kingman, 1982a, Brown, 1995). They may compose 10% of global animal diversity, but estimates suggest that only 8% of dipteran species are currently described (Brown, 1995). Second, Diptera encompasses a broad range of life histories (typical sampling yields detritivores, herbivores, parasitoids and predators), and a large diversity of species from a given locality can be easily captured using standardizable sampling methods (Brown, 2005).

Figure 4.1. (A) Distribution of land above 2000m in Costa Rica (shaded area) and distribution of paramo (shapes nested within shaded area). Study sites indicated by yellow stars and three letter abbreviations - VIR, CLM and CHI. (B) Distribution of counts of individuals, sequences and species delimited by ABGD across sites: individuals collected (Ind), individuals sequenced (Seq), number of species (Spp: lumped/split), distance between sites (Dist) and the maximum height of mountain passes between sites (gray boxes).
In this study I sampled Diptera in three distinct páramo sky islands in order to ask questions about how many species might inhabit these fragments, how similar their communities are, and whether the valleys between the fragments pose substantial biogeographic barriers. I test three a priori models of páramo Diptera diversity:

1) Common community model: The fauna is either composed of highly dispersive alpine taxa, or similarly dispersive old-field edge species. In either case, genetic structuring will be low and communities will be highly similar. Underlying patterns of phylogenetic structure will be random.

2) Isolation-neutral divergence model: The fauna is composed of alpine or high elevation endemics that find intermontane habitat hostile. As such, communities will be similar, but genetic divergence will be high, and higher between VIR and CHI-CLM than between CHI and CLM.

3) Ecological divergence model: Community divergence will be moderate to high and communities will show non-random phylogenetic structure. Genetic structure within shared species will be low.

I do not expect that these models are mutually exclusive, and it is likely that species individualistic. However, the models are proposed to identify the unifying patterns in this poorly known ecosystem.

4.2. MATERIALS AND METHODS

4.2.1. Sampling Design

I sampled three habitat islands spanning ~60% of the distribution of páramo islands in Costa Rica (Fig. 4.1A). All samples were collected between July 25th and August 15th 2009. The three sampling sites are located between 3300m and 3400m at Volcán Irazú (VIR) in the Cordillera Volcánica, and at Cerro de la Muerte (CLM) and Cerro Chirripó (CHI) in the Cordillera de Talamanca. These sites are arrayed from northwest to southeast (VIR-46km-CLM-28km-CHI). VIR is separated from CLM and CHI by the central valley, and the pass between them is ~1500m. The pass between CHI and CLM is ~2700m. Sampling methods at each site include malaise traps, flight intercept traps, pan traps and sweep transects. I retrieved samples daily and stored them in lots by trap and day in 90% ethanol until processing.

4.2.2. Specimen Processing, DNA Extraction and Sequencing

I sorted individuals to morphospecies within lots and identified them to the family level (Brown, 2010, Brown, 2009). For DNA extraction I chose 5-10 specimens at random from each morphospecies within each lot. I extracted whole genomic DNA from entire individuals using either QiaGen DNeasy kits [cite] or the protocol of Ivanova et al. (2006), both of which are based on binding DNA to silica membranes in 96 well plates. Exoskeletons remain intact during this process and were retained as voucher specimens. A 648bp region of Cytochrome Oxidase I was amplified using standard primers (Vrijenhoek, 38
1994) and PCR products were sent to Beckman Coulter Genomics for Sanger sequencing. I edited chromatograms individually in Geneious version 5.04 (Drummond et al., 2011) and then aligned and verified each polymorphism by eye. I identified and discarded potential nuclear pseudogenes by checking each sequence for indel mutations that were not in groups of three nucleotides and stop codons. I discarded any sequences with multiple peaks causing ambiguity in base calls. Finally, I aligned all retained sequences using MUSCLE (Edgar, 2004) and constructed a neighbor-joining tree. I verified the identification of any specimen with a sequence that did not cluster with other specimens from its sampling locality or family. I generally was able to resolve these as cases of misidentification, mis-sorting, or contamination from an adjacent well in a 96-well plate. I also retrieved the closest BLAST hit in GenBank for each sequence using BioPython (Cock et al., 2009) to look for irregularities and to determine if any close relatives have been sequenced.

4.2.3. Genetic Analyses

First, I clustered DNA sequences into operational taxonomic units (referred to hereafter as species). I used ABGD (Automated Barcode Gap Detection) (Puillandre et al., 2012), a non-parametric method based on the idea that pairwise genetic distances among individuals within species should be smaller than those among species. It searches for such "gaps" in the distribution of pairwise distances using the slope of ranked pairwise distances. It does this iteratively by first partitioning the entire dataset using the largest detected gap and then searching each cluster for gaps.

Because these gaps are expected to appear by chance in intraspecific coalescent trees, the algorithm requires the user to provide a data-set wide estimate of maximum prior intraspecific divergence, $P$, that is used to calculate a value of $\theta = 4N_e\mu$. This value of $\theta$ is then used to calculate the maximum expected intraspecific divergence in order to cluster sequences. There is no objective means to select the "best" $P$ a priori, and different $P$ seemed to give different, but reasonable results. This is consistent with the expectation that there should be real uncertainty about assigning species limits from single-locus DNA sequence data, particularly in cases when genetic divergence among populations is high. Therefore I retained two sets of species partitions, a "lumped" dataset and a "split" dataset corresponding to reasonable large and small prior $\theta$ values (discussed in section 4.3.2. Genetic Analyses below). I conducted all further analyses on both sets of species limits to explore how this uncertainty might influence analyses of community structure.

I sampled the posterior distribution of the phylogeny uniting the sampled species using both the split and lumped datasets in BEAST (Drummond & Rambaut, 2007). The two datasets contain only one representative per species. I used the GTR+$\Gamma$ model of sequence evolution, a relaxed clock model assuming uncorrelated, lognormally distributed rates of molecular evolution among branches and the birth-death-sampling tree prior (Stadler, 2009), which accounts for extinction and the fact that not all descendants of the common ancestor of species in my sample were sampled. BEAST samples the posterior using Markov chain Monte Carlo. I ran the Markov chain for 100 million generations, sampling every 5000, discarding the first 10 million as burn-in. I used the maximum clade credibility
tree as a point estimate and subsampled 200 trees in order to test whether downstream analyses were sensitive to uncertainty in the phylogeny.

In order to test whether the fragmented nature of Central American páramos is associated with genetic divergence, I analyzed patterns of polymorphism within and among populations. Because many species shared among sites have very low or uneven sampling, (see below, but most are represented by 1-2 individuals at one or more of the sites) traditional measures such as AMOVAs and pairwise Fst’s are difficult to compute accurately and compare among populations. I chose instead to evaluate a simple measure using and estimate of Watterson’s $q$ (Watterson, 1975). $\hat{\theta}$ is an estimate of the population parameter $q=4N_{m} \mu$ based on the numbers of segregating sites, $K$, and alleles sampled, $N$. It is simply the quotient of $K$ and the N-1th harmonic number. For my test I compared pairs of populations, calculating $\hat{\theta}_L - \hat{\theta}_P$, where $\hat{\theta}_L$ is the value of $\hat{\theta}$ calculated by lumping the samples from the two populations together and $\hat{\theta}_P$ is the mean $\hat{\theta}$ of the two populations. If the two populations are really a single panmictic breeding population, the expectation of this value is zero. If they are divergent, then $\hat{\theta}_L$ will be greater than zero. If one population had only one individual sampled, I excluded that population from the mean calculation and used only it’s sister. I excluded population pairs represented by a single sequence each. This measure should have high variance with low sample size, so I also calculated the mean across all populations for each pair of localities.

### 4.2.4. Species Diversity Analyses

Because I only sequenced a few individuals per morphospecies per sample, and many diversity analyses require species abundances, I used a “pseudo-sample” of actual abundances. In order to get this pseudo-sample I assigned unsequenced individuals from each within-sample morphospecies to the associated sequenced-identified species. In the event that a morphospecies actually contained more than one species, I assigned unsequenced individuals proportionally with the sequenced species. I conducted this pseudo-sampling for both the lumped and split partitions.

I calculated three diversity indices for each site, the species richness, the exponential of Shannon entropy and the inverse Simpson index, each of which differ in how strongly relative abundance is weighed (Jost, 2006). I used rarefaction and extrapolation methods (Colwell et al., 2012) to compare these quantities among sites, which had uneven sample sizes. I estimated total species richness at each site using the Chao and ACE estimators (Chazdon et al., 1998, Colwell & Coddington, 1994), the number of species shared between sites using the Chao shared species estimator and a number of abundance-based estimate of similarity including the Chao-Jaccard and Chao-Sorensen measures (Chao et al., 2005), which estimate the true similarity of the communities, not the similarity of the community samples. Each statistic was calculated for both the lumped and split partitions. These analyses were conducted in EstimateS (Colwell, 1997).

Contingent on various assumptions, a number of analyses of patterns of community structure can give insight into processes structuring ecological communities. I first tested the use of neutral models of community assembly, that assume that local communities are
composed of immigrants and their descendants from a regional metacommunity. These models make a number of assumptions, including that all individuals have the same niche, regardless of species, and that a fixed number of individuals are playing a zero-sum game for survival and reproduction in a given site. I applied the $\text{G}_st(k)$ estimator of Munoz et al. (2008) that estimates rates of immigration ($m$) from the metacommunity using a statistic similar to Wright’s $F_{st}$ [cite] and a Maximum Likelihood approach that estimates Hubbell’s fundamental biodiversity number $\theta = 2Jm\nu$, where $J_m$ is the size of the metacommunity and $\nu$ is the speciation rate using the software package ‘unb’ in the statistical language R (Munoz et al., 2007, Hankin, 2007, R Development Core Team, 2011).

I also measured community phylogenetic structure using several metrics, the MPD and MNTD (weighted and unweighted) of Webb et al. (2002), PSV and PSE of Helmus et al. (2007), and two pairwise measures of phylogenetic beta diversity, PhyloSor of Bryant et al. (2008). For each metric I assessed significance by shuffling tip labels on the phylogeny for 100 replicates. This holds species distributions and the pattern of compositional dissimilarity among communities constant in order to test whether that pattern is non-randomly associated with phylogeny. I conducted significance tests on 100 samples from the posterior distribution of phylogenetic trees to test whether uncertainty in the phylogeny might affect conclusions about the observed patterns, and I conducted each analysis on both the split and lumped partitions.

4.3. RESULTS

4.3.1. Sampling and Sequencing

I sampled and sorted 2702 individuals. As the collecting period carried on, the rainy season began and my sampling effort at my final site, CHI, was limited by rain. At high elevations, activity of Diptera was markedly slower when it was foggy or raining, and this likely resulted in smaller trap samples. I submitted 1920 individuals in total for sequencing and 1728 sequences (90%) passed all filters and were included in the final dataset. Sequencing failures were assumed to be random and dropped from the dataset. 686 individuals were pseudo-sampled, yielding a final dataset containing 2414 individuals with 773 in VIR, 900 in CLM and 741 in CHI (Fig 4.1B). These individuals represent 30 families with Muscidae, Tachinidae and Sciaridae being the most diverse (Table 4.1). The uncorrected distances between my sequences and their best BLAST hits ranged from 0 to 16.5% with a mean of 8.5% and standard deviation of 2.7%. Most close BLAST hits were members of the same family, although this was not universally the case. Only ten sequences from 6 species had a hit less than 3% divergent and only two of those species had identical hits. One of these is an undescribed species of *Toxomerus* (Syrphidae) known from highland bogs in Costa Rica. The other is the only presently sequenced member of the genus *Copromyza* (Sphaeroceridae). It is likely *C. equina* a globally introduced synanthropic species (J. Kits pers. comm.).
I tested a range of prior intraspecific divergences in ABGD, from $P=0.001$ to $P=0.36$. At each extreme results seemed unrealistic, resulting in 740 and one species respectively. In the case of $P=0.001$, species in my estimated trees were not monophyletic and most
morphospecies were broken up. Between $P=0.0017$ and $P=0.0215$, the algorithm estimated between 334 and 283 species. I took this long area of relative stability to be an indication that these values were reasonable. These $P$ values result in estimates of $\theta$ from 0.0006 to 0.0033, which are small, but comparable to values published from across animals (Bazin et al., 2006, Wares, 2010). Because there is no straightforward way to assess which value in this range is optimal, I chose to conduct all analyses on two partitions, a “lumped” partition with 334 species at $\theta=0.0017$ and a “split” partition with 283 species at $\theta=0.0215$.

Results from BEAST yielded surprisingly well-supported trees for both the lumped and split datasets (Fig. 4.2). Shallower nodes tended to be well supported, except in a clade of Muscids with several recently diverged species, while deeper ones tended to be poorly supported. Many families were not recovered as monophyletic, but this is to be expected given the limited phylogenetic information contained in this short section of CO1. Because many of my specimens are extremely divergent from sequenced representatives of their families, I elected not to constrain families to be monophyletic in the event that they were in reality, not. The Bayesian approach to sampling credible trees should account for uncertainty in the phylogeny, so I tested the sensitivity of analyses using the phylogeny to that uncertainty below.

My measures of divergence using Watterson’s $\hat{\theta}$ were highly variable, but skewed positively, indicating genetic divergence among populations (Figs. 4.3, 4.4). Pairwise comparisons segregated by pairs of localities showed different patterns, with the positive skew for the CLM-VIR pairs being far greater than that of the other two, and substantially greater for the lumped than the split dataset.

### 4.3.3. Species Diversity Analyses

Sites varied greatly in observed lumped/split species richness with 92/102 species at CHI, 178/196 at CLM and 88/104 at VIR (Fig. 4.1b). There were 35/36 observed shared species between CHI and CLM, 15/17 between CHI, VIR and 28/34 between CLM and VIR and 10/12 shared at all three sites (Fig. 4.1B). Rarefaction curves indicated that CLM had the highest diversity and CHI the least (Fig. 4.5). The inverse Simpson measure for VIR was very close to that of CLM, indicating that CLM’s much greater richness is owed largely to rare species. The CHI sample has markedly lower diversity using the inverse Simpson index. This is likely a result of a single extremely common Dolichopodid species contributing over 300 specimens. Results were fundamentally similar for both the lumped and split datasets.

I estimated total site species richness using Chao and ACE statistics (Fig. 4.6). These results gave identical rankings for each site (CLM>CHI>VIR), which contrasted slightly with the raw measures of richness in which CHI and VIR are virtually identical. The lumped and split estimates gave different estimates of richness, but notably, the slopes of their rarefaction curves were also different. The lumped curves each appeared much closer to a stable asymptote, whereas the split curves suggest that increased sampling of individuals would yield increased estimates of species diversity.
Figure 4.2. Maximum clade credibility tree of single representatives of the split (334 species) dataset estimated with BEAST. Morphologically determined Family indicated by gray boxes on the right. Node support (posterior probability) shown as colored dots on internal nodes. Black bars immediately to the right of tree indicate at which sites (CHI, CLM, VIR) the species were sampled. 1%, 3% and 5% expected substitutions per site shown on the tree with dotted lines.
Figure 4.3. Mean pairwise $\theta_L - \overline{\theta}_P$ values between the three population pairs (CHI-CLM, CHI-VIR, CLM-VIR) using the lumped and split datasets. More positive values suggest stronger genetic divergence.

Figure 4.4. $\theta_L - \overline{\theta}_P$ values for all population pairs for all pairs of sites (CHI-CLM, CHI-VIR, CLM-VIR) for both the lumped (283 species) and split (334 species) datasets. More positive values indicate stronger genetic divergence.
Figure 4.5. Comparisons of site diversity indices for the lumped (283 species, left column) and split (334 species, right column) datasets. Extrapolated sample richness (top row), exponential of Shannon Entropy (middle row) and the Inverse Simpson index (bottom row). Curves are rarefied (and/or extrapolated) values obtained by averaging indices calculated from 200 rarefied replicates for each number of individuals.
I estimated the number of species shared between sites using the Chao estimator and the similarity of sites using the Chao-Sorensen estimator (Fig. 4.7). The closest sites CHI and CLM had the highest estimated number of shared species and similarity, while CHI and VIR, the most distant had the lowest. These numbers did not change markedly with the lumped and split dataset. The shared species and similarity estimates for CLM and VIR, however, did change substantially, increasing from 49.5 species in the split dataset to 83.2 in the lumped dataset, and from 0.375+/-0.048 to 0.546+/-0.079. Their standard deviations did not overlap. None of the 95% confidence intervals for any similarity indices in either dataset overlapped 1. There was a negative trend in similarity with increasing distance, but with only three pairwise comparisons I did not subject this to a statistical test.

Analyses with different estimators of neutral model parameters yielded very different results, though the lumped and split datasets were similar for both. The Likelihood
implementation estimated an extremely large $\theta = 2J_m v$ of 9700 and $m$ for each community approaching 1. This suggests that divergence in community composition results from the observed species being sampled randomly from an enormous metacommunity. The estimates of $m$ from $G_{ST}(k)$, by contrast ranged from 0.00541632 to 0.03958606, suggesting that community divergence is the result of very strong dispersal limitation.

Tests of phylogenetic community structure were found to be sensitive to phylogenetic uncertainty and to uncertainty in species limits (Fig. 4.8). The distribution of p-values produced from the sampled Bayesian credible trees sometimes spanned $\alpha = 5\%$ (two-tailed test). Analogous metrics of community phylogenetic structure from Webb et al. (2002) and Helmus et al. (2007) produced virtually identical results, with PSV=MPD and PSE=abundance weighted MPD, so I display only results from the Webb et al. metrics. CHI was never found to have a significant p-value in any test, while VIR was either significant or on the margin for phylogenetic clustering for both the MPD and the MNTD (incidence or abundance-based) for both the lumped and split datasets. CLM varied drastically depending on the statistic, whether it was abundance or incidence based and between the split and lumped datasets, ranging from significantly clustered (abundance-weighted MPD, split dataset) to marginally overdispersed (incidence-based MNTD, split dataset).

The PhyloSor index of Bryant et al. (2008) was also found to be sensitive to uncertainty in phylogeny and species limits (Fig. 4.9). Only one comparison was significant in both
datasets and that was CLM-VIR, in which phylogenetic similarity was less than expected given compositional similarity. The CHI-CLM comparison overlapped α for the split dataset.

**4.4. DISCUSSION**

**4.4.1. Inference Of Páramo Community Structure**

In this work I have combined analyses of species limits, phylogeny and ecological community structure in order to better understand the Diptera fauna of a little studied tropical alpine ecosystem. I can safely reject the common community model of the páramo fauna. I find large and significant differences among communities with estimates of compositional similarities among sites significantly different from 1. Estimates of shared species (including unobserved shared species) ranged as low as VIR sharing only 13% of its species with CHI. Neutral model estimates also suggest strong dispersal limitation to these communities.
Support for the other two models is equivocal. I find a signature of genetic divergence among populations of species shared across sites in the $\hat{\theta}$ statistics. The signal is strongest between CLM and VIR, which are separated by the deep Central Valley and weak between CHI and CLM. It is also weak between CHI and VIR, which span the same boundary, but the reason for this is unclear. It may be related to the fact that far fewer species are shared between that pair. In any case, this is strongly suggestive of population structuring and supported by the fact that the CLM-VIR positive skew increases drastically in the lumped dataset. This indicates that some of the uncertainty in species delimitation is associated with geographically partitioned genetic variation. This suggests lineage divergence is occurring in at least some species across sites and is consistent with the isolation-neutral divergence model. This divergence could be associated with environmental divergence and selection against immigrants among sites, however, so these hypotheses are difficult to separate.

I also observed non-random community phylogenetic structure, supporting the ecological divergence model. Again, the signal is strongest when considering the site VIR. This site shows strong evidence for phylogenetic clustering of species, and the PhyloSor index indicates that, given the pattern of compositional similarity, phylogenetic similarity between it and CLM is less than expected by chance. The deficit in expected shared branch length between these communities is $\sim 7\%$ of their average branch length. At a minimum, this represents, 60 million years of unshared evolutionary history. Phylogenetic clustering
is often interpreted as abiotic filtering (Webb et al., 2002) stemming from environmental differences among sites (but see Mayfield and Levine (2010)). This is consistent with intuition about the sites developed in the field, as VIR is clearly quite different in terms of soil and vegetation. CLM and CHI are dominated by the dwarf bamboo Chusquea subtessulata, while VIR has a much greater proportion of ericaceous shrubs.

I conclude from these analyses that it is likely that both ecological filtering processes and geographic isolation are working to structure páramo communities in Costa Rican highlands. Although most biogeographic work on Central American highlands has focused on much larger spatial scales (Casteo et al., 2009, Weir et al., 2008, Weir, 2009), isolation at this relatively scale in the southern Central American highlands is not heretofore unknown. A few bird populations are polymorphic among mountain ranges in Costa Rica, including the hummingbird Selaphorus flammula. This bird has diverged in gorget color, an important conspecific signal (Stiles & Skutch, 1989). Notably, several species of Bolitoglossa salamander exhibit exceptionally strong genetic structuring in apparently continuous habitat in sub-alpine oak (García-París et al., 2000).

There are a number of caveats that must be discussed with respect to some of my analyses. First, I used individual based methods to estimate diversity and similarity, but sample based statistics have much better performance (Gotelli & Colwell, 2001). My sampling was quite limited, and distributed among sampling methods, so I chose to use the individual approach. It is likely that some aspects of the patterns I observe would change with improved sampling. For example, CHI is by far the largest patch of Central American páramo, yet it had the lowest diversity. My sampling there was limited by the onset of the rainy season, which took away sampling days and decreased the activity of Diptera when rains ceased. This is a likely source of bias.

Second, while my use of neutral models is inconsistent with their typical use in ecology, my system clearly violates two of their important assumptions. First, the niche equivalency assumption cannot even be taken as a first approximation in Diptera. It is unlikely that a 2mm Sciarid fungus gnat is in any way competing with a 1.5cm parasitic Tachinid. Second, community samples are regarded as actual communities, not samples from a broader population, subject to measurement error, as ours are, yet that is how I treat them in my interpretation. My use of a statistic like GST(k), however, is in keeping with a long established practice in population genetics of attempting to gain insight through the analysis of strongly violated models.

Finally, my \( \hat{\theta} \) statistic is open to interpretations other than population genetic structure. In particular, rapid population expansion could result in a positive skew. In this case I regard that alternative as unlikely as páramo habitats have undergone vast contraction since the last glacial maximum. Additionally, the population expansion signature should be present in comparisons of all population pairs, but I observe it most strongly between the CLM-VIR pair, which is consistent with a priori predictions about genetic divergence.
4.4.2. Importance Of Species Limits And Phylogenetic Uncertainty

Most ecological studies that use trees, and all that use species limits use point estimates of these quantities provided by experts. Here I show that uncertainty in these estimates could influence the interpretation of empirical data. The approach I used to test this was to calculate each statistic using plausible estimates of each parameter. For phylogenetic trees, I calculated all metrics using the phylogeny using a sample of trees from the Bayesian posterior probability distribution. Each of these trees represents a draw from the posterior, and a collection of them represents a plausible range of topologies and branch lengths. While my topology was reasonably well resolved, branch length uncertainty was substantial, with a root height 95% posterior density interval of 0.15 substitutions per site. At a mutation rate of 2.5% per million years (Brower, 1994), this represents 6 million years of uncertainty. My results show that for the Webb et al. measures (Fig. 4.8), in at least three instances, the significance of the structuring of a community is dependent on the tree used to calculate the statistic. For the PhyloSor statistic (Fig. 4.9), this is true in one instance. I do not here provide a method of testing for significance while incorporating posterior distributions of trees (and advocate against posterior averaging of frequentist p-values), but this seems to be a straightforward way of testing whether such a dependency might exist.

In order to test the impact of choice of species limits on inference in community structure I used two partitions of my dataset, a lumped partition containing 283 species, and a split partition containing 334 species. Results suggest uncertainty in species limits has a potentially serious impact on inference. First, estimates of the total species richness are very different between datasets. Because the Chao estimator is based on the number of singletons and doubletons, this suggests that my split partition results in a notable increase in rare species. My rarefaction curves using the Chao estimator suggest very different trends in my sampling. The split dataset suggests that I do not yet have an accurate picture of species richness, while the lumped dataset suggests estimates of species richness are approaching an asymptote (Fig. 4.6).

Estimates of shared species are also impacted, but in a different way. In my case, population structure potentially associated with species formation results in uncertainty about species delimitation. Because this uncertainty affects the number of species shared among sites it also affects measures of community similarity. In my case, measures of similarity and shared species for CHI-CLM and CHI-VIR are unaffected by choice of species limits, but there is a fairly drastic impact on CLM-VIR. This is because six species in the lumped dataset split across this biogeographic boundary. Lineage divergence is a process widely recognized as contributing to community structure, but its nature as a process with fuzzy temporal boundaries can be confounding when analyses seek to model discrete lineages.

4.5. CONCLUSIONS

In this chapter I have shown that evolutionary methods (phylogenetic inference, species delimitation, population genetics) can be combined with analyses of ecological community
structure and diversity to rapidly make inferences about the history and diversity of poorly described faunas. In the face of rapid anthropogenic change, developing integrative approaches such as this that can speed the discovery of diversity are essential. The approach I have taken here is data limited, however, and I have shown that it requires a careful accounting of the uncertainty associated with various estimates. As genomic resources become more inexpensive and widely available, the DNA-based side of this approach will become more and more robust. Sampling, however, will not be subject to such technological economies. In order to increase confidence in my results here, a substantial increase in field collecting would be required.
Chapter 5.
Conclusions.

One of the primary goals of biology is to understand the origins and distribution of diversity. In recent years, model-based approaches to this problem have become widespread, and their continued development promises to facilitate more rapid and accurate progress toward this goal. Three important elements of these approaches are addressed in this document: accounting for uncertainty in nuisance parameters when generating model estimates, testing whether models are a good fit to empirical data, and the integration of ecological and evolutionary approaches.

In Chapter 2 I developed a Bayesian implementation of the general mixed Yule-coalescent model for species delimitation. This method produces estimates of the marginal posterior probability of species limits by integrating out uncertainty in the model parameters and in the underlying phylogeny. I used my implementation to show that inference of species limits from single-locus data can be accurate under a variety of demographic scenarios, but also that ignoring uncertainty in the phylogeny leads to inflated confidence in the results. This is a fundamentally useful tool in that the result is a sample from the marginal posterior distribution of species limits. This sample can be used to test whether downstream ecological and evolutionary analyses using an assumed taxonomic framework are sensitive to uncertainty underlying the evolutionary model. Indeed, in Chapter 4 I provide evidence that this can be the case. It can also be used to create posterior-averaged estimates of aspects of biological diversity that fully account for uncertainty.

In Chapter 3 I developed a novel posterior predictive check for a commonly used phylogenetic inference model: the multispecies coalescent. The multispecies coalescent represents a recent move in phylogenetics towards highly parameterized models that are more realistic representations of evolutionary processes. Though these models are complex, the genomic-scale datasets to which these models are applied may be more complex. The posterior predictive method is the only available means by which researchers may test the fit of their data to this popular model. I used the method to test the fit of 25 published datasets and found that poor model fit is evident in a majority of them. I also found that some cases of poor fit can yield positively misleading results. This tool will be highly useful to researchers, and the chapter emphasizes posterior predictive model checking as an essential step in Bayesian inference, on par with the selection of models and parameter estimation. Researchers to date have essentially ignored this step, even in cases when their results are unexpected and skepticism is warranted.

Chapter 4 is an empirical study of the diversity of the Dipteran fauna of a little-known, fragmented, tropical alpine ecosystem. Ecological and evolutionary methods that address diversity at regional scales often ask very similar questions, but take different approaches. Phylogeography typically examines the history and distribution of a set of closely related populations, while community ecology looks at structure at the community scale. In this chapter I explored an approach that exploits the complementarity of these fields by...
integrating DNA sequence-based species delimitation methods, phylogenetic inference and ecological analyses of diversity to try to understand the structure and history of this community. I found that different fragments harbor communities that are distinct in both composition and phylogenetic structure, and that shared species show strong evidence of genetic divergence among fragments. This suggests that both isolation and variation of abiotic conditions in these sky-islands contribute to their divergent community composition. I also find that my ecological analyses are somewhat sensitive to uncertainty in species limits and phylogeny. Species limits seem to be particularly important in this case, as whether or not divergent genetic lineages are ascribed species status has a substantial impact on estimates of compositional similarity.

5.1. FUTURE WORK

There are a number of prospects for the advancement of this work. Chapter 2 is based on single-locus sequence data. This is primarily because broad-scale surveys of diversity, such as Chapter 4, pose a number of barriers to traditional PCR-based approaches for developing multilocus datasets. With the rise of DNA sequence capture techniques and high throughput sequencing it may soon be possible to efficiently gather multilocus sequence data from thousands of individuals of distantly related species (Faircloth et al., 2012). These data have the potential to yield much more precise and accurate estimates of species limits. Conducting large-scale species delimitation analyses on these datasets, however, will require expanded models. It is possible to modify the GMYC to account for multilocus data, and such an approach may be preferable to expanding already parameterize rich models such as the one implemented in BPP (Yang & Rannala, 2010).

Fully parametric models such as BPP and *BEAST are exceptionally useful in that they make explicit their assumptions, have well-defined interpretations, and relatively straightforward techniques such as posterior predictive simulation can be implemented to test their fit. I have done this here, but I have not conducted power analyses on my method to see how sensitive it is to deviation from the true model, or how often it detects deviations of consequence. This would be a useful extension. Additionally, these methods are computationally expensive, particularly for genome-scale datasets. A number of approaches have been suggested that make use of summary statistics rather than full likelihood functions in order to save computation. Checking the fit of these models is much more difficult, and their performance when their assumptions are violated is poorly known. Because it is likely they will become widely used it would be useful to compare all approaches under various model violations to assess the risk posed by more expedient, less verifiable models.

Finally, here I have attempted to integrate ecological methods with evolutionary ones in biodiversity discovery, but much remains to be done. One major barrier is that the statistical paradigms utilized by each of these fields are fundamentally different. Frequentist null model analysis is common in ecology, but Bayesian approaches to model selection and fitting are widespread in evolutionary biology. Results of null model analyses cannot be validly conditioned on a Bayesian posterior distribution, and so I am limited to showing that uncertainty in evolutionary models could impact my ecological inferences. A
better approach would be to fit parametric models of community assembly that could be integrated into a fully Bayesian framework. With the exception of a few neutral models, which may not be particularly apt for this approach, these have not been developed. Likewise, because of my fairly limited genetic data (one genetic locus, small numbers of individuals per species) I was unable to fit parametric models of demographic history. My approach to using phylogeography to understand community connectivity would be bolstered by high throughput sequencing approaches mentioned above, which would allow the estimation of migration rates and divergence times. Interpreting patterns in these parameters would be much more useful than the simple measure of genetic divergence I used.

On the whole, further integration of evolutionary and ecological methods seems to have promise in allowing us to rapidly discover patterns of diversity and understand the processes that have generated it. This integration awaits a number of statistical and technical developments that I hope to contribute further to in my future career as a scientist.
References


Drummond, A. & Suchard, M. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology* **8**: 114.


Appendix A.

bGMYC Instructions For Use

Package written by:
Noah Reid
nreid1@tigers.lsu.edu (until summer 2013) or noah.reid@gmail.com

These instructions are for an R package that implements a Bayesian version of the general mixed yule-coalescent model of Pons et al. (2006). This work is in press at BMC Evolutionary Biology and should be released at some point in late 2012. The citation will be Reid and Carstens 2012. Phylogenetic estimation error can decrease the accuracy species delimitation: A Bayesian implementation of the General Mixed Yule Coalescent model.

The motivation for writing this package is that for the kind of data this model is usually applied to (DNA barcode sequences), a single point estimate of phylogeny is typically associated with large amounts of uncertainty. Unaccounted for, the model may lead to overconfidence in species limits, influencing the accuracy of downstream analyses of species diversity or community structure.

As such, this algorithm is meant to be applied to multiple trees sampled from a posterior distribution of ultrametric trees (approximated using some clock model, e.g. in BEAST (Drummond and Rambaut 2007)). Because the idea is to conduct multiple MCMC runs, if you use it, you should check the output of each MCMC to see that it looks to be sampling from a stationary distribution. I am not an experienced programmer, so this package is probably not the most efficient way of doing things, but I have run it on many data sets and conducted a simulation study testing it’s efficacy, so I know it is fairly robust, at least on a 2009 era Macbook Pro laptop. BUT PLEASE LET ME KNOW IF YOU HAVE PROBLEMS.

INSTRUCTIONS

bGMYC is distributed as an R package. It requires only the package "ape" to run. So that needs to be installed first. Once installed, bGMYC can be installed from a unix terminal window using the command:

    R CMD INSTALL path/bGMYC_1.0.tar.gz

Once R is opened, it can be loaded using:

    library(bGMYC)

Documentation for each function is included and can be gotten by typing:

    ?functionname
The data necessary to run bGMYC are phylogenetic trees. It can be run with a single tree, but the main strength of this implementation is to integrate over uncertainty in tree space by analyzing many trees. In practice, I have found that around 100 trees sampled at intervals from a posterior distribution of trees is sufficient to capture that uncertainty, but it may vary from dataset to dataset.

Trees can be loaded using the ape command:

```r
read.nexus(file="path/treefile.tre") -> trees
```

Once the trees are in loaded, the first step is to run the algorithm on a single tree to make sure it’s working and see what the burn-in should be. 50k generations is a good start. The threshold parameter priors and starting parameter values need to be set. The threshold parameter is essentially the number of species, and is given a uniform prior, so the priors, t1 and t2 need to be set to possible values (e.g. between 2 and the number of leaves in the tree) and the starting value for that parameter (the third value in the "start" vector) needs to be t1<start<t2 or else there will be an error.

```r
bgmyc.singlephy(trees[[1]], mcmc=50000, burnin=1, thinning=10, t1=2, t2=100, start=c(1,1,25)) -> result.single
```

When this finishes, plot the output of the MCMC to see if the chain is mixing and how long it takes to reach stationarity. I am recommending visual checks here rather than a more rigorous check for convergence because the model is fairly simple, with only three parameters. Use the function:

```r
plot(result.single)
```

You can check whether the uniform priors are placing unreasonable bounds on the chain, whether mixing is adequate, and decide what the burn-in should be. If it looks like the MCMC is converging on the number of tips in your input trees, this may be an indication that the model does not fit your data (MCMC evaluation of trees with no discernable threshold tend to favor delimiting every sequence as a species).

Once you decide what the burn-in needs to be, you can decide how many trees you want to run (I usually use 100) and the thinning interval (I usually use 100). If you run 100 trees, keeping 10k post-burnin generations and use a thinning interval of 100, that yields 10k samples. That usually seems to be more than enough (in that adding more trees doesn’t really change the final results), but results will vary from dataset to dataset. I should also point out that as you start adding many tips, things slow down a bit, the most I’ve ever run was ~400 tips and it took about a day to complete. Now, in order to run many trees you can use:

```r
bgmyc.multiphylo(trees, mcmc=50000, burnin=40000, thinning=100) -> result.multi
```

When this finishes, again use:
plot(result.multi)

to check the behavior of the MCMC. These plots may look a little funny, however, because they are the results of several sequential MCMC runs, each with their own burn-in, etc. You may find that one or two of the trees fails to reach stationarity or winds up in a weird area of parameter space. This may suggest that you should use a longer burn-in.

If things look fine, then you can run

bgmhc.spec(result.multi)->result.spec

to get a table that lists each species sampled in the course of the MCMC runs and its posterior probability. If you specify "filename" it will output this information as a table.

If you want to visualize the posterior distribution in the context of the tree, you should do:

spec.probmat(result.multi)->result.probmat
plot(result.probmat, trees[[1]])

generates a matrix of tree leaves by tree leaves with each cell containing the probability that the two corresponding tree leaves are members of the same species. Plotting this matrix along with a single phylogenetic tree will order the matrix and display a heatmap-like image with a legend that shows the probabilities of conspecificity, as shown in our paper.

This is all I've got for now. Please let me know if you need assistance or have recommendations for how better to process and display the results. Thanks for using my package!

Noah Reid

References:


Reid and Carstens 2012. Phylogenetic estimation error can decrease the accuracy species delimitation: A Bayesian implementation of the General Mixed Yule Coalescent model.
Appendix B.  
starbeastPPS Instructions For Use

Instructions for package 'starbeastPPS'
Written by Noah Reid
2 December 2012

Introduction:

This package implements a posterior predictive model check for the multispecies coalescent model implemented in *BEAST (Heled and Drummond 2010). This check is described in a publication submitted to Systematic Biology in November of 2012 (Reid et al. 2013?). It is meant to assess whether an empirical dataset is a good fit to the model. The model assumes that individual gene genealogies are independent, identically distributed draws from a stochastic coalescent process in a species tree. It further assumes that all species are panmictic populations and that sampled individuals have been assigned to the correct species. This is not always the case, as we know that migration, selection, gene duplication and extinction, in addition to population substructure and uncertainty in identification can all be important in specific datasets. When these processes are operating but unmodeled, phylogenetic inference can be mislead.

The goal of this package is to identify, using posterior predictive simulation (Gelman et al. 2003) cases when processes other than stochastic lineage sorting are influencing a phylogenetic dataset. Identification of such issues can help users of *BEAST figure out if their phylogenetic estimates are being mislead, or give them hints that previously unconsidered biologically interesting processes are at work in their data. In any case, Gelman et al. (2003), who have written a really awesome text on Bayesian statistics, say that model checking is a fundamentally important part of Bayesian data analysis, of equal importance to model specification (what Heled and Drummond 2010 have done) and model fitting (what you as a *BEAST user have presumably done). So if you have run *BEAST for phylogenetic inference, you should use this package to see if the model is a good fit for your data! You might learn something interesting!

I have tried to make this package user friendly because I think people should use it. I have tried, at least cursorily, to document all functions in the R help system, so try ?function if you need help. As for interpretation, I will post a copy of the manuscript, which will hopefully be helpful. Otherwise, if something isn’t clear, or you think something is wrong, please contact me at noah.reid@gmail.com and I will try to help sort it out.

Package preface and excuses:

I am still in a learning phase with this whole scripting thing, so I know that I have not always found the most efficient or elegant solutions in the creation of these scripts. In particular, reading the data in for a lot of loci with large numbers of alleles takes much,
much more time than I would have hoped. This is because I was uncertain of the best way to get the branch information (rate variation and effective population sizes) out of the tree and into a vector. If anybody out there sees this and has suggestions, I will be happy to implement them. Also, I wrote all this on a mac, and some scripts need to be pointed to executables of ms and seq-gen, so I’m not sure it will work on a pc.

How the analysis works:

I have included a test dataset from my M.S. thesis work on the phylogeny of the western North American chipmunks (Sciuridae: Tamias; Reid et al. 2012). There is rampant mtDNA introgression in Tamias, so the test data kicks out the mtDNA is poorly fitting the *BEAST multispecies coalescent model. For space-saving purposes, the dataset only includes 100 samples from the posterior distribution of the model. The test data is ready to be read into R (i.e., no file re-naming is necessary).

BEFORE STARTING:

You need copies of two pieces of software: Hudson’s ms (Hudson 2002) and Rambaut and Grass’s seq-gen (Rambaut and Grass 1997). Simulations of coalescent genealogies and sequence data, respectively, are conducted using these programs. This is because I was most familiar with them and they are fast. As I said before, I’m not an expert programmer, so I don’t know how to actually physically incorporate them in the package, so you need to point R to them. I’ve included their source code so you can compile them on your machine.

Then you need to install the package. In the mac terminal you can type:

R CMD INSTALL path/to/starbeastPPS_1.0.tar.gz.

The package depends on phangorn (Schliep 2011) and ape (Paradis and Strimmer 2004).

*BEAST STEPS:

The only important analytical consideration is that the piecewise constant model of population sizes must be used. The simulation machinery can’t deal with population size change on branches. This is because *beast uses a linear size change model and ms only simulates under exponential population size change. It seems unlikely that most datasets will have large amounts of information about change in Ne on interior branches anyway.

Naming is important for loci. names with spaces, starting with numbers, or that have sections.separated.by.periods cause errors in my scripts. sorry about that.

Taxa cannot have parentheses in names either.
if you change BEAST output file names after running it, you also need to change them inside the xml file, because the scripts use the XML file to identify files to be read.

A note on missing data and gaps: sequences with all missing data are removed in the simulations and calculations of summary statistics at the sequence level. After those sequences are removed, all alignment columns with missing data are removed. This means that loci for which each alignment column has missing data across two or more individuals will be entirely removed and this causes an error (e.g. if sample 1 has bases 1:50/100 and sample 2 has 51-100/100). Missing columns are removed because I don't know if you can, or how, to calculate the multinomial likelihood when data are missing.

FILE ORGANIZATION AND NAMING:

First, run the included perl script "beast.tree.log.combiner.pl". This script is basically the equivalent to the "logcombiner" program included with BEAST, in that it will discard burn-in, thin your MCMC samples if you wish, and combine the results from different runs, but it will be faster and easier and rename your files so the R package can find and read them.

To do this, put all the tree files and the log file from one or more *BEAST runs in one directory.

Then execute the perl script.

Enter your burn-in and if you want the samples thinned. In the publication we only used 2k MCMC samples split between 2 runs.

That's it.

If you don't run the perl script:

Make sure names are consistent with XML file (in case you've altered them)

all tree files and the log file have to have the same number of samples, and they need to correspond with each other.

all gene tree files must be called "mylocus.combined.trees", the species tree must be called "species.combined.trees", and the log file must be called "combined.log"

All the relevant files should now be in a single directory...

R STEPS:

library(starbeastPPS)
load the package.

read.starbeast("mydata.xml", "path/directorycontainingbeastoutput")-&gt;mydata
reads in all necessary information

analyze.coalescent(mydata, "path/directorycontainingHudsonsms")-&gt;mydata.coal
analyzes fit of coalescent genealogies to species tree using coalescent likelihood and deep coalescences

```r
summary(mydata.coal)
```
gives a summary of analysis including posterior predictive p-values and the mean of the test statistic distributions (the expected mean for all test statistic distributions is 0).

```r
plot(mydata.coal)
```
visualizes summary statistic distributions. red lines indicate boundaries of 95% (solid) and 99% (dashed) posterior predictive density intervals. lines only on one side of a distribution indicate a one-tailed test. the black line indicates 0, the expected mean of all test statistic distributions.

```r
analyze.sequences(mydata, "path/directorycontainingseqgen") -> mydata.seq
```
analyzes fit of sequence data to gene trees (with branch lengths in substitutions per site).

```r
summary(mydata.seq)
```
gives summary of analysis including posterior predictive p-values and the mean of the test statistic distributions (the expected mean for all test statistic distributions is 0).

```r
plot(mydata.coal)
```
visualizes summary statistic distributions. red lines indicate boundaries of 95% (solid) and 99% (dashed) posterior predictive density intervals. lines only on one side of a distribution indicate a one-tailed test. the black line indicates 0, the expected mean of all test statistic distributions.

**SUPPORT FUNCTIONS**

read.gtree.samples.R

reads in a *BEAST gene tree file, including among-branch rate variation values.

read.sptree.samples()

reads in a *BEAST species tree file, including demographic values.

mstree()

used to simulate gene trees within a species tree. available to the user.

**REFERENCES:**


Appendix C.
License For Reproducing Chapter 2

Below is the full BioMed Central Open Access license agreement (Identical to the ‘Creative Commons Attribution License’), under which Chapter 2 was published in BMC Evolutionary Biology. It grants permission to reproduce Chapter 2 in this document.

Full BioMed Central Open Access license agreement
(Identical to the 'Creative Commons Attribution License')

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS BIOMED CENTRAL OPEN ACCESS LICENSE ("LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions
1. "Collective Work" means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.
2. "Derivative Work" means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance of doubt, where the Work is a musical composition or sound recording, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered a Derivative Work for the purpose of this License.
3. "Licensor" means the individual or entity that offers the Work under the terms of this License.
4. "Original Author" means the individual or entity who created the Work.
5. "Work" means the copyrightable work of authorship offered under the terms of this License.
6. "You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License.
despite a previous violation.

2. Fair Use Rights
Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3. License Grant
Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

1. to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;
2. to create and reproduce Derivative Works;
3. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works;
4. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission Derivative Works;
5. For the avoidance of doubt, where the Work is a musical composition:
   1. **Performance Royalties Under Blanket Licenses.** Licensor waives the exclusive right to collect, whether individually or via a performance rights society (e.g. ASCAP, BMI, SESAC), royalties for the public performance or public digital performance (e.g. webcast) of the Work.
   2. **Mechanical Rights and Statutory Royalties.** Licensor waives the exclusive right to collect, whether individually or via a music rights agency or designated agent (e.g. Harry Fox Agency), royalties for any phonorecord You create from the Work ("cover version") and distribute, subject to the compulsory license created by 17 USC Section 115 of the US Copyright Act (or the equivalent in other jurisdictions).
6. **Webcasting Rights and Statutory Royalties.** For the avoidance of doubt, where the Work is a sound recording, Licensor waives the exclusive right to collect, whether individually or via a performance-rights society (e.g. SoundExchange), royalties for the public digital performance (e.g. webcast) of the Work, subject to the compulsory license created by 17 USC Section 114 of the US Copyright Act (or the equivalent in other jurisdictions).

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4. Restrictions
The license granted in Section 3 above is expressly made subject to and limited by the
following restrictions:

1. You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this License or the recipients’ exercise of the rights granted hereunder. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this License Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested. If You create a Derivative Work, upon notice from any Licensor You must, to the extent practicable, remove from the Derivative Work any reference to such Licensor or the Original Author, as requested.

2. If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Derivative Works or Collective Works, You must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied; to the extent reasonably practicable, the Uniform Resource Identifier, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and in the case of a Derivative Work, a credit identifying the use of the Work in the Derivative Work (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Derivative Work or Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5. Representations, Warranties and Disclaimer
UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.
6. Limitation on Liability
EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination
1. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Derivative Works or Collective Works from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.
2. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous
1. Each time You distribute or publicly digitally perform the Work or a Collective Work, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.
2. Each time You distribute or publicly digitally perform a Derivative Work, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.
3. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
4. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.
Vita

Noah Mattoon Reid was born in Rochester, New York in 1981, where he grew up on the edge of a city park that fostered his love of the natural world. He began his undergraduate career at Boston University in the fall of 2000, where he majored in Biology. He attended a transformative semester abroad in Ecuador in 2001 with Kelly Swing, where tropical biodiversity became a major research interest for him. He was the Teaching Assistant for the same course in 2003. During the summer of 2002, he was a participant in the National Science Foundation’s Research Experience for Undergraduates under Polly Campbell, where they studied the roosting ecology of Malaysian bats. After graduating in 2003, he worked as a field assistant on various ecological and evolutionary projects ranging from bird banding in Massachusetts to community ecology of tropical leaf litter ants.

Noah started a Master’s degree at the University of Idaho with Jack Sullivan in 2005. He investigated hybridization and phylogenetics of western Chipmunks. In 2008, he began a Doctoral degree with Bryan Carstens at Louisiana State University. In 2013, he will begin a position as post-doctoral researcher with Andrew Whitehead at the University of California, Davis.