GENOMIC AND PHYLOGENETIC CONSEQUENCES OF DIVERGENCE WITH GENE FLOW IN CHIPMUNKS (SCIURIDAE: *TAMIAS*)

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctorate of Philosophy

with a

Major in Biology

in the

College of Graduate Studies

University of Idaho

by

Brice A. J. Sarver

August 2014

Major Professor: Jack Sullivan, Ph.D.

Authorization to Submit Dissertation

The dissertation of Brice A. J. Sarver, submitted for the degree of Doctorate of Philosophy with a Major in Biology and titled "Genomic and Phylogenetic Consequences of Divergence with Gene Flow in Chipmunks (Sciuridae: *Tamias*)," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____
Jack Sullivan, Ph.D.

Committee Members: _____ Date: _____
Luke J. Harmon, Ph.D.

_____ Date: _____
Paul A. Hohenlohe, Ph.D.

_____ Date: _____
David C. Tank, Ph.D.

Department Administrator: _____ Date: _____
James J. Nagler, Ph.D.

Discipline's College Dean: _____ Date: _____
Paul Joyce, Ph.D.

Final Approval and Acceptance

Dean of the College of Graduate Studies: _____ Date: _____
Jie Chen, Ph.D.

Abstract

Speciation, the means through which new species arise, is of central interest to biology. Recent theoretical, methodological, and computational advances have allowed researchers to study speciation at scales that were once impossible. This dissertation builds on existing literature to arrive at conclusions regarding the estimation of diversification rates from molecular phylogenies and characterize nuclear and mitochondrial genetic patterns in western North American chipmunks (*Tamias*, subgenus *Neotamias*), a system that has experienced rapid diversification in the face of gene flow.

Rates of diversification and extinction can be inferred from phylogenies using a series of statistical approaches. However, parameter estimates should be affected by errors introduced during phylogenetic estimation. I conducted a simulation study to assess the impact of molecular clocks and tree priors on the estimates of diversification rates. I found that the choice of molecular clock and choice of tree prior do not impact estimates of diversification rates except in circumstances of extreme mismatch (e.g., assuming a single-rate clock when extreme rate heterogeneity is present).

Previous work describes widespread mitochondrial introgression in chipmunks. To generate data sufficient to characterize evolutionary patterns in central and southern Rocky Mountains chipmunks, I use an exon capture technique to sequence thousands of loci from 51 individuals across 6 species of chipmunks. I assemble and characterize mitochondrial genomes from all individuals. Phylogenetic analyses indicated rampant mitochondrial introgression; subsequent tests suggest that selection at protein-coding genes does not appear to be governing introgression. I conclude that demographic factors, such as population expansion, provide a likely explanation of the patterns of introgression in these species.

I conducted a nuclear phylogenomic analysis, using data from thousands of nuclear loci and resolved the relationships among the six species using a mix of traditional and species-tree estimation approaches. I characterized the genomics of this system using several population genetic approaches and find little nuclear introgression, suggesting that the nuclear genome is resistant to introgression in the face of widespread mitochondrial introgression. I conducted these analyses in the absence of a reference genome and provide pipelines and suggestions for genomic inferences in non-model systems.

Acknowledgements

I thank everyone who provided guidance throughout my graduate career. Specifically, I would like to thank my major professor, Jack Sullivan, for helping with everything from data generation through analysis. I also thank my committee members Luke Harmon, Paul Hohenlohe, and David Tank. In particular, Luke provided mentoring and advice on a variety of topics even though I was not his immediate student, for which I am grateful.

Matt Settles and Sam Hunter convinced me that computational and programming skills are necessary in modern biology. I cannot thank them enough for pushing me to develop a computational skillset. I also thank Kerry O'Donnell. His willingness to grant independence to a young researcher in his lab was my impetus for pursuing a graduate degree, and he showed me what research was really like outside of the classroom.

Finally, I would like to thank my family, friends, and colleagues who have shown me support and/or helped with projects over the past six years: Jim and Sheree Sarver, Jim and Delores Slater, Jim and Kris Slater, Lauren Sarver, Susan Kologi, Zev Kronenberg, Kayla Hardwick, Matt Pennell, Travis Hagey, Kyle Tolbert, Matt Rappaport, Zak Rosemore, Martin Kuffel, Dane Wilkes, and all professors, students, and staff of the Department of Biological Sciences and IBEST.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speciation biology is synthetic, and it incorporates data from across many sub-disciplines of biology to draw conclusions about how species form and, ultimately, explain life's diversity (Coyne and Orr 2004; Seehausen et al. 2014). A PubMed search reveals that 1,657 papers were published in 2013 that included the term "speciation," and the number of papers published each year has been steadily increasing since 2000. This is not surprising. The early 2000s marked the transition to the "Genomic Era" of modern biology (Guttmacher and Collins 2003), and with it came the ability to generate datasets that were once impossible or cost-prohibitive.

Prior to the advent of high-throughput sequencing technologies, models had been developed with the goal of explaining and characterizing genomic patterns of divergence. These models, coupled with evidence that hybridization and introgression appears frequently in natural systems (Funk and Omland 2003; Mallet 2005, 2007), focused attention on the notion that reproductive isolation between lineages does not necessarily occur instantaneously. Commonly referred to as "divergence (or speciation) with gene flow" models (Wu 2001; Pinho and Hey 2010), these conceptualizations emphasize that reproductive isolation between incipient lineages occurs gradually. Furthermore, species can still be characterized as "good" species and not be reproductively isolated. Under these models, genomic divergence increases with time, driven by genetic hitchhiking, until isolation is complete. However, characterizing genome-wide patterns of divergence requires sequencing technologies and computational approaches that were, until recently, unavailable outside of model systems.

It is now possible to obtain data from hundreds or thousands of markers from multiple individuals and test hypotheses in a system of interest. This can be done with or without a reference genome (e.g., Hodges et al. 2007, 2009; Bi et al. 2012) and is therefore applicable to non-model systems. Genomic datasets provided the basis for two new fields: population genomics (Jorde et al. 2001; Charlesworth 2010) and phylogenomics (Eisen 1998; Delsuc et al. 2005). This dissertation draws from these fields and explores speciation using phylogenetic simulations and empirical characterization of a natural system.

Chapter 2 focuses on estimating diversification rates from phylogenies. A series of models have been developed that take information present in phylogenies, such as branching times, and estimate diversification parameters (e.g., net rate of diversification or relative rate of extinction) from them (e.g., Yule 1925; Kendall 1948; Nee et al. 1994a,b; Gernhard 2008; Stadler 2013). However, parameters are closely tied to the phylogenies from which they are estimated; if error is introduced while estimating the phylogeny, results may be incorrect or biased. Only a handful of studies have investigated the impact of phylogenetic error on inferring diversification rates, focusing on model misspecification (Revell et al. 2005), error in branch length estimates (Wertheim and Sanderson 2011), and incomplete taxonomic sampling (Höhna 2014).

This chapter assesses the impact of tree priors and molecular clocks on estimates of diversification parameters. Specifically, I simulated trees by sampling from the prior under several combinations of tree priors and molecular clocks using BEAST v1.7.5 (Drummond et al. 2012). I then simulated data on these trees and estimated trees from the sequence data under all combinations of tree priors and molecular clocks. Finally, I estimated diversification parameters from the resulting posterior distributions. This approach allowed

me to assess the impact of phylogenetic misspecification (of the tree prior, molecular clock, or both) on estimates of diversification parameters by comparing parameter estimates from the original trees to trees generated under match and mismatch conditions.

Chapters 3 and 4 investigate evolutionary patterns in chipmunks (Sciuridae: *Tamias*). Chimpunks consist of 25 species. One species, *Tamias sibiricus* (subgenus *Eutamias*), is distributed throughout eastern Asia. The remaining 24 are distributed throughout North America, with one species, *T. striatus* (subgenus *Tamias*) inhabiting eastern North America and the remaining 23 (subgenus *Neotamias*) inhabiting western North America. Chipmunks are stark niche partitioners with narrow zones of contact (Heller 1971; Heller and Gates 1971), and species distributions can be governed by competitive exclusion (e.g., Brown 1971). Furthermore, individuals can be assigned to species based on a morphological character, the baculum or *os penis*, providing a reliable, non-genetic means of species identification.

Over a decade of work has used a single mitochondrial locus (cytochrome b) in concert with nuclear loci and microsatellites to describe genetic patterns in chipmunks (Good et al. 2003, 2008; Hird and Sullivan 2009; Hird et al. 2010; Reid et al. 2010, 2012; Sullivan et al. 2014). One of the most surprising outcomes from this work is the extent of mitochondrial introgression. Mitochondrial introgression is pervasive and, as a result, hinders systematic resolution within the genus (Good et al. 2008; Reid et al. 2012). To combat these issues, Reid et al. (2012) sequenced reproductively-associated proteins in an attempt to resolve species relationships. While the analysis produced the most resolved chipmunk phylogeny to date, several nodes presented with low statistical support.

These two chapters focus on a subgroup of western North American chipmunks, the *T. quadrivittatus* group, consisting of six species (*T. canipes*, *T. cinereicollis*, *T. dorsalis*, *T. quadrivittatus*, *T. rufus*, and *T. umbrinus*). Using a targeted exon capture protocol designed for chipmunks using a pooled-tissue transcriptome (Bi et al. 2012), I was able to obtain data for thousands of loci spanning the diversity of this group. The analysis of this data was broken into two separate studies.

Chapter 3 focuses on describing patterns of mitochondrial introgression using whole mitochondrial genomes. A draft mitochondrial genome was included as part of probes designed for exon capture, and an iterative assembly approach (ARC; Hunter et al. in prep; https://github.com/ibest/ARC) was used to assemble a mitochondrial genome for each individual. Here, I focus on investigating patterns of introgression within the *T. quadrivittatus* group; specifically, I am interested in characterizing the mitochondrial genomes, estimating a phylogeny, and testing for evidence of positive selection using a myriad of approaches. If there is evidence for selection, it could provide an explanation of the widespread introgression in this system. If, however, I fail to detect signatures of selection, I would favor an alternative hypothesis: demographic factors, such as population expansion, explain patterns of introgression (e.g., Klopfstein et al. 2006; Currat et al. 2008; Excoffier et al. 2009).

Chapter 4 provides a contrast to work in Chapter 3, focusing on the nuclear genome. In light of extensive mitochondrial introgression, the next logical step is to document the extent of nuclear introgression. In addition, there are still outstanding questions concerning the systematic relationships among the six species. In order to investigate these questions, I use ARC to generate a set of contigs for each individual that are based on the sequences

used to design exon capture probes. The resulting set of contigs was processed to generate two datasets: a reduced dataset consisting of approximately 220,000 base pairs across ~1000 contigs to be used for phylogenetic inference, and a set of variants called against the assembly from a single individual to be used in population genomic analyses. I estimated phylogenies using a series of classical (i.e., concatenation) and species-tree (e.g., Maddison 1997; Edwards 2009) approaches that explicitly account for discordance among loci. Furthermore, I estimated population genetic statistics (e.g. F_{ST}, F_{IS}, H_O, etc.) for species or species pairs in addition to individual coancestry to describe patterns of admixture among species.

In conclusion, this dissertation consists of three chapters. Chapter 2 uses phylogenetic simulations to assess the impact of tree prior and molecular clock misspecification on phylogenetic estimates of diversification rates. Chapter 3 describes the assembly of chipmunk mitochondrial genomes, their characteristics, and tests for selection as the driving force of mitochondrial introgression. Chapter 4 characterizes nuclear patterns of introgression, estimates population genetic parameters, and uses a series of phylogenetic approaches to resolve the systematics of central and southern Rocky Mountains chipmunks.

Chapter 2

The Choice of Tree Prior and Molecular Clock Does Not Substantially Affect Phylogenetic Inferences of Diversification Rates

Abstract

Comparative methods allow researchers to make inferences about evolutionary processes and patterns from phylogenetic trees. In the majority of studies, the phylogeny is assumed to be estimated without error. However, estimating trees can be error prone for a variety of reasons. Sources of error introduced throughout the process, and their impact on comparative parameters, has only been investigated in a handful of studies. Additionally, this error may systematically bias phylogenetic estimation and, therefore, estimation of parameters. Here, we focus on the impact of priors in Bayesian phylogenetic inference and evaluate how it affects the estimation of parameters in macroevolutionary models of lineage diversification. Specifically, we use BEAST to simulate trees under combinations of tree priors and molecular clocks, simulate sequence data, estimate trees, and estimate diversification parameters (e.g., speciation rates and extinction rates) from these trees. We find that the choice of tree prior and molecular clock has relatively little impact on the estimation of diversification rates, insofar as the sequence data are sufficiently informative and rate heterogeneity among lineages is low to moderate. When rate heterogeneity is large, parameter estimates deviate substantially from those estimated under the simulation conditions. Therefore, the impact of priors on phylogenetic analyses should be assessed before using phylogenetic trees to infer rates of diversification.

Introduction

Statistical comparative methods use phylogenetic trees to gain insight into large scale, macroevolutionary patterns (Felsenstein 1985; Harvey and Pagel 1991; Pennell and Harmon 2013). Branch lengths and node ages provide information about the rate of lineage accumulation throughout time (e.g., Nee et al. 1994b; Nee 2006; Ricklefs 2007; Pyron and Burbrink 2013). Approaches use a point estimate of a phylogenetic tree or a distribution of trees to estimate macroevolutionary parameters, such as the rate of lineage accumulation (speciation) or extinction, which are often compared across groups to provide insight into diversification rates and the tempo of evolution (Nee et al. 1992; Magallón and Sanderson 2001; Alfaro et al. 2009). However, parameter estimates are dependent on the tree from which they are inferred (Felsenstein 1985). Most inference procedures assume that a tree is estimated without error (Felsenstein 1985, 2004), but, because branching times and branch lengths are critical to estimates of diversification parameters, inaccurate phylogenies can be expected to yield unreliable estimates. A handful of studies have focused on the causes of parameter misestimation when fitting diversification models to trees (e.g., Nee et al. 1994b; Barraclough and Nee 2001; Revell et al. 2005; Cusimano and Renner 2010; Rabosky 2010; Wertheim and Sanderson 2011), but few studies have evaluated error in phylogenetic estimation explicitly (but see Revell et al. 2005).

While a definitive characterization of the impact of phylogenetic error remains at large, recent theoretical advances have expanded the scope of phylogenetic comparative methods. Previously, only models that assumed a constant rate of lineage diversification or extinction existed. Current methods can use phylogenies to determine where shifts in the rates of speciation and extinction take place (e.g., Rabosky 2006a, 2006b, 2014; Alfaro et al.

2009) or estimate rates that depend on species' traits (e.g., Maddison et al. 2007; FitzJohn et al. 2009; FitzJohn 2010). However, phylogenetic error can directly affect results. For example, Revell et al. (2005) demonstrated that underparameterization of the model of nucleotide sequence evolution as part of the process of phylogenetic estimation can produce apparent slowdowns in the rate of diversification as quantified by Pybus and Harvey's (2000) gamma statistic (Revell et al. 2005). Additionally, errors in branch lengths (Wertheim and Sanderson 2011) and biased taxonomic sampling can both affect estimates of diversification rates (Höhna 2014). These studies suggest that phylogenetic error can affect the estimation of comparative parameters.

Bayesian methods of inference produce posterior distributions of trees, and comparative parameters can be estimated across such distributions to quantify error. Furthermore, the use of Bayesian approaches in phylogenetics has increased in recent years due in part to the availability of efficient software, including BEAST (Drummond et al. 2012) and MrBayes (Ronquist et al. 2012). However, the impact that the choice of priors governing the molecular clock and branching processes has on the estimate of comparative parameters has not been thoroughly investigated. Two commonly used tree priors are the Yule (Yule 1925) and Birth-Death (BD; Kendall 1948; Nee et al. 1994b; Gernhard 2008; Stadler 2013) models. The Yule model is the simplest of a group of continuous-time branching processes; it has one parameter, λ, which is the instantaneous per-lineage rate of speciation that is constant across the tree. The BD model is also a continuous-time process but includes a probability that a lineage will go extinct (and, therefore, leave no descendants); thus, model has two parameters, λ and μ, the instantaneous per-lineage rates of speciation and extinction (both of which are constant across the tree). In practice, many

approaches re-parameterize the model and estimate $r = (\lambda - \mu)$ and $\varepsilon = (\mu / \lambda)$, the net diversification rate and relative extinction rate, respectively. In general, estimates of r have greater precision than ε (Nee et al. 1994a, 1994b; FitzJohn et al. 2009). In BEAST, researchers must specify a prior distribution on λ or on r and ε, depending on the choice of tree prior.

Molecular clock models must also be specified; BEAST v1.7.5 gives users the choice of using a strict (or global) molecular clock or an uncorrelated log-normal relaxed molecular clock, among others (Drummond et al. 2012). The strict clock assumes a constant, global rate of sequence evolution across the tree (Zuckerkandl and Pauling 1962), and the uncorrelated log-normal relaxed clock (UCLN) assumes branch-specific rates are drawn from a log-normal distribution (Drummond et al. 2006). Priors are placed on the mean rate of evolution for the strict clock and the mean and standard deviation of the log-normal distribution for the uncorrelated log-normal relaxed clock.

Wertheim and Sanderson (2011) focus on trees generated only under the Yule process with a range of λ values. They simulated sequences under a simple model of sequence evolution (HKY85), and trees were estimated using BEAST assuming a strict clock and narrow prior or range of prior widths on the root age. Their study assessed the impact of sequence length and nodal calibrations on estimating posterior distributions of λ, and they found that increasing sequence length leads, as expected, to narrower 95% HPD widths of speciation rates. Additionally, broader calibration priors were shown to increase posterior widths of these estimates.

As a result, there is precedence for the choice of priors affecting the estimation of diversification parameters. Similar conclusions can be drawn from first principles in

Bayesian statistics: strong priors influence posterior distributions. In a phylogenetic sense, it is plausible that forcing estimation of a tree under a particular branching process (such as a Yule process) may produce an inaccurate tree if the true generating process was different (such as a BD process); this could systematically affect diversification parameter estimates. Since branch lengths play an important part in estimating diversification parameters, it is also the case that a mismatch of clock models could similarly affect results.

Here, we quantify the effect of tree prior and clock misspecification on parameter estimation for diversification models by comparing estimated values inferred under the generating conditions to those inferred from mismatched conditions. In order to accomplish this, we simulate phylogenetic trees and associated sequence data under a range of combinations of tree priors and molecular clock models. We re-estimate trees using BEAST and use these reconstructed trees to calculate maximum likelihood estimates of diversification rate parameters. We compare these trees to estimates from the original trees to evaluate whether or not prior misspecification contributes to error in estimating diversification rates.

Materials and Methods

We take advantage of existing and newly-developed applications to simulate trees under a variety of conditions, simulate nucleotide sequence data on these trees, estimate a tree from the nucleotide data, and estimate comparative parameters. The workflow is illustrated in Figure 1.1. All scripts are written in the R programming language (R Core Team 2013) and are available on GitHub (http://github.com/bricesarver/priorsims).

Generation of initial distributions of trees

We simulated trees of two sizes, 25 and 100 taxa, both with a tree depth of 5 arbitrary time units. We simulated initial trees using BEAST v1.7.5 with XML input files from BEAUti v1.7.5 (Drummond et al. 2012). Because BEAUti requires data in NEXUS format consisting of DNA sequences for each individual, we created a 'dummy' tree using TreeSim (Stadler 2011). DNA sequence data were simulated on this tree using SeqGen v1.3.2 (Rambaut and Grass 1997). These dummy data were only used to fix the number of taxa; sequence data were replaced by an empty alignment before execution by sampling from the prior (see below).

The simulation process itself consisted of two steps. First, a tree prior was selected for each round of simulations. Of the possible choices, Yule and BD were used in this study. In order to avoid improbable combinations of parameters such that tree shapes were non-randomly sampled (Pennell et al. 2012), initial parameter values were calculated using the expectation relating the net diversification rate, the number of taxa, and the tree height:

$$E[N_t] = N_0 e^{rt} \qquad [1]$$

where N_t is the number of taxa at t, N_0 is the initial number of taxa, r is the net diversification rate ($\lambda - \mu$), and t is the height of the tree (Nee 2006). For BD cases, ε is fixed at 0.5.

BEAST produces chronograms and phylograms and requires the specification of a type of molecular clock. For the strict case, the prior on the clock rate is fixed to a log-normal distribution of the form log-normal(1.5, log(0.125)). For the UCLN case, the prior on the mean of the distribution is of the form U(1.45, 1.55), and the prior on the standard

deviation of the distribution is U(0.17, 0.18). This log-normal prior distribution represents a modest-to-low amount of per-lineage rate heterogeneity.

For the 100-taxa case, we also investigated the impact of estimating under Yule and BD tree priors when rates are sampled from an 'extreme' log-normal distribution. We placed a U(45, 55) prior on the mean and a U(0.95, 1.05) prior on the standard deviation, resulting in trees with a wide range of rates on each branch. Rates were, therefore, more variable and greater than in the standard UCLN case above. Other than these prior probabilities, all other simulations and analyses were identical to the non-extreme cases.

We then generated a distribution of trees under these conditions using BEAST, sampling only from the priors. Operators were removed for fixed parameters, such as root height, r, ε, and λ, and left in place for parameters associated with clocks. Operators are the implementations of the MCMC proposal mechanisms; in other words, they governed the distributions from which new parameter values were drawn. Removing them fixed parameters at their initial value.

Simulation of nucleotide datasets

We generated a posterior distribution of 10,001 phylograms by sampling from the prior. Ten trees were selected at random without replacement. Trees were rescaled by multiplying branch lengths by 0.01 before simulation of nucleotide sequence data, effectively reducing the substitution rate to more realistic values. Alternatively, the distributions from which rates were drawn could be characterized by smaller values. Regardless, trees were rescaled to the appropriate size after estimation in BEAST. 5000 bp of sequence data (see Wertheim and Sanderson 2011) were simulated under a GTR+Γ model of nucleotide sequence evolution with parameters estimated in Weisrock et al. (2005) for

nuclear rRNA (π_A: 0.1978, π_C: 0.2874, π_G: 0.3403, π_T: 0.1835; r_{AC}: 1.6493, r_{AG}: 2.9172, r_{AT}: 0.3969, r_{CG}: 0.9164 r_{CT}: 8.4170, r_{GT}: 1.0; α: 2.3592). Sequences were simulated on selected topologies using Seq-Gen v1.3.5 (Rambaut and Grass 1997) with randomly generated seeds.

Estimation under tree prior and clock combinations

The resulting NEXUS data files were processed using BEASTifier v1.0 (Brown 2014). BEASTifier takes a list of NEXUS files and generates BEAST XML input files under conditions specified in a configuration file. Each combination of tree priors and clock types was used for each dataset. For example, the sequences generated using a 100 taxon tree that is simulated under a Yule tree prior and strict molecular clock ultimately produced four XML files for analysis: the condition matching the simulation conditions [e.g., a posterior distribution of trees using a Yule tree prior and a strict clock (1)] and all mismatch conditions [e.g., a posterior distribution of trees using a Yule tree prior and a UCLN clock (2), a BD tree prior and a strict clock (3), and a BD prior and UCLN clock (4)]. Each file was then processed using BEAST v1.7.5 (Drummond et al. 2012). Chains were run for 50,000,000 generations, sampling every 5000, with a burn-in of 5,000,000 generations. Convergence was assessed through visual inspection of traces and verification that the ESS of all parameters was approximately 200 or greater. A maximum clade credibility tree was generated for each analysis using TreeAnnotator v1.7.5 and assuming median node heights and a posterior probability limit of 0.5.

Analysis of posterior distributions and maximum clade credibility trees

We analyzed each combination of the four possible simulation/estimation cases (Yule:Strict, Yule:UCLN, BD:Strict, and BD:UCLN) and number of taxa (25 or 100). First, each distribution of trees was rescaled to the exact root height of the original tree using

functions in ape (Paradis et al. 2004). This served to remove a small amount of error when estimating parameters because, in almost every case, the root height was not exactly 5 but was extremely close (e.g., 4.9997). The first 1000 trees of the posterior distribution were removed as a burnin. For each tree in the posterior, we estimated λ and r using the pureBirth() and bd() functions in the library LASER (Rabosky 2006a). The means of λ and r were calculated for each posterior distribution. These ten point estimates were then plotted for each simulation case.

In addition, we produced lineage-through-time (LTT) plots for each replicate. The LTT plot of the maximum clade credibility tree produced from each analysis was plotted on the same graph as the original tree from which that data was simulated. Each plot, then, consists of LTT plots for the 10 original trees and consensus trees from the corresponding 10 estimated posterior distributions.

Results and Discussion

The goal of this study is to determine the impact the choice of tree prior and molecular clock have on the estimation of comparative phylogenetic parameters. We focused our efforts on estimating λ, the rate of lineage formation, and r, the net diversification rate, under all combinations of two tree priors (Yule and BD) and two flavors of molecular clocks (strict and UCLN). While previous work describes a relationship between parameter estimation and misspecification of the model of nucleotide sequence evolution during phylogenetic estimation (Revell et al. 2005), as well as sequence length and nodal calibrations (Wertheim and Sanderson 2011), no studies to our knowledge have directly focused on the impact of tree priors and choice of molecular clocks under which trees are estimated.

We found that the combination of tree prior and clock did not substantially impact diversification parameter estimates. Across our simulation conditions, parameters from trees estimated under all combinations of tree priors and clocks were concordant with parameter estimates produced from the trees on which nucleotide data is simulated. When original trees were simulated under a Yule process, all combinations of tree priors and clocks produced extremely similar estimates to the parameters estimated from trees on which data were simulated (Fig. 1.2). Distributions overlapped across all combinations of tree priors and molecular clocks. Interestingly, there appeared to be some inflation of estimates as indicated by a noticeable increase in medians when estimating λ (Fig. 1.2, 100:Yule:Strict and 25:Yule:Strict λ cases). These trends were not replicated when estimating r. Additionally, there was a slight decrease in median estimates in the 100-taxon Yule:UCLN cases. LTT plots of maximum clade credibility trees indicated that the estimated trees generally coincide with the original trees, though the Yule:UCLN case showed greater discordance at nodes deeper in the tree (Fig. 1.3). This is likely attributable to sampling error associated with selecting 10 trees on which to simulate data, and we would expect that this discrepancy would be reduced if we were to perform simulations using thousands of starting trees. Computational limitations prohibit this in practice.

When trees were simulated under a BD process, estimates were also concordant with the original trees. Medians were nearly identical among many simulation conditions (Fig. 1.4). Parameters were slightly underestimated in a single case, the BD:UCLN estimates of λ, though not extremely so. LTT plots revealed that maximum clade credibility trees were, again, approximately equivalent to the original. There were some exceptions, again in the deep nodes of the trees (most clearly illustrated in the BD:UCLN case indicating non-

overlap of the original trees and estimated trees), though these did not affect parameter estimation (Fig. 3).

Extreme clock simulations revealed that it is possible for the choice of clock to have an impact on the estimates of results (Fig. 1.5). It is important to recognize that, in these circumstances, rates of evolution will be both large and highly variable among branches. Parameters estimated from trees inferred under conditions identical to which they were originally simulated are the most accurate, with estimates overlapping those produced from the original trees. Mismatch conditions produced less accurate results. In particular, the strict clock, which assumes a single rate of evolution across the tree, did not accurately capture rate heterogeneity, resulting in inflated or depressed estimates (Fig. 1.5). The uncorrelated log-normal clock better captured per-branch rate heterogeneity, as expected.

Therefore, it appears that reasonable parameter estimates can be achieved with either prior. This is somewhat surprising given that posterior estimates can be influenced by the choice of priors. At least in these cases, either choice of tree prior appears to capture the underlying branching process on which data was simulated; the same holds for molecular clocks (at least for the non-extreme cases). While estimates are concordant across tree priors and clock models, previous studies have shown that the accuracy of the estimates depends on the amount of data available; here, this refers to the number of taxa. In one example, trees of 1000 taxa produce more accurate estimates of diversification rates than trees of 100 taxa (Stadler 2013). Adding more informative data produces more accurate phylogenetic estimates (assuming no signal conflict) and should reduce the impact of stochasticity on parameter estimation.

The assumption of a single rate of evolution across a tree is often violated and can severely impair phylogenetic estimation (e.g., Shavit et al. 2007; Penny 2013). This study assumed rates with a modest amount of heterogeneity, and it appears that a strict clock produces reasonable results in the face of this violation. In other words, a dataset with a small to moderate amount of heterogeneity may have rates that are reasonably captured by a single, global rate. However, it may not be known *a priori* whether a dataset has disparate rates of evolution among lineages. It would be advisable, then, to assume a clock model that has the potential to model heterogeneity accurately, and this is partially why the uncorrelated log-normal relaxed clock has seen such widespread use and success in systematic analyses (Drummond et al. 2006). Furthermore, should rates of evolution be extreme among some lineages, it would make sense to attempt to capture any heterogeneity using appropriate priors as opposed to assuming it is absent. Rate homogeneity among lineages, or the absence of a clock altogether, may represent a poor prior given our current understanding of molecular biological processes (Drummond et al. 2006).

There are several caveats to these conclusions. First, our original trees are fully resolved, and nucleotide sequence data are simulated under parameters estimated from a quickly evolving nuclear intron. This indicates that there will be a large number of phylogenetically informative sites per individual. Therefore, these trees will be easier to estimate than those that lack signal and/or contain unresolved nodes. Second, there is not extreme rate heterogeneity among lineages. Third, the datasets only contain 25 and 100 taxa, each with only 5000 bp of nucleotide sequence data, following the protocol of Wertheim and Sanderson (2011). Datasets of this size are considered modest in the current era of high-throughput sequencing, where the generation of hundreds of thousands or millions of base

pairs of sequence per individual is possible. It is also reasonable to assume that some systems may be best explained through more complex models, i.e., models that specifically assume multiple, independent diversification rates across a dataset (e.g., Alfaro et al. 2009; Rabosky 2014). Our analyses only assume a single rate of diversification, and this assumption may be violated in larger datasets with greater levels of taxonomic divergence. It is important to select among models in order to produce accurate, interpretable results for each dataset.

Conclusion

The choice of tree priors and molecular clock has little impact on the estimation of diversification parameters under these simulation conditions. Parameters can be estimated reliably from any combination of tree prior and molecular clock with informative data and reasonable clock rates. The choice of molecular clock has little impact on diversification parameters unless there is extreme rate heterogeneity.

References

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. U. S. A. 106:13410–4.

Barraclough T.G., Nee S. 2001. Phylogenetics and speciation. Trends Ecol. Evol. (Personal Ed. 16:391–399.

Brown J.W. 2014. BEASTifier. Available at http://github.com/josephwb/BEASTifier.

Cusimano N., Renner S.S. 2010. Slowdowns in diversification rates from real phylogenies may not be real. Syst. Biol. 59:458–64.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A.J., Suchard M. a, Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–73.

Felsenstein J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1–15.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland, Massachusetts: Sinauer Associates.

FitzJohn R.G., Maddison W.P., Otto S.P. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595–611.

FitzJohn R.G. 2010. Quantitative traits and diversification. Syst. Biol. 59:619–33.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–78.

Harvey P.H., Pagel M.D. 1991. The Comparative Method in Evolutionary Biology. Oxford Ser. Ecol. Evol. 1:239.

Höhna S. 2014. Likelihood inference of non-constant diversification rates with incomplete taxon sampling. PLoS One. 9:e84184.

Kendall D.G. 1948. On the Generalized “Birth-and-Death” Process. Ann. Math. Stat. 19:1–15.

Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a binary character’s effect on speciation and extinction. Syst. Biol. 56:701–10.

Magallón S., Sanderson M.J. 2001. Absolute diversification rates in angiosperm clades. Evolution. 55:1762–80.

Nee S., Holmes E.C., May R.M., Harvey P.H. 1994a. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 344:77–82.

Nee S., May R.M., Harvey P.H. 1994b. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 344:305–11.

Nee S., Mooers A.O., Harvey P.H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl. Acad. Sci. U. S. A. 89:8322–8326.

Nee S. 2006. Birth-Death Models in Macroevolution. Annu. Rev. Ecol. Evol. Syst. 37:1–17.

Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 20:289–290.

Pennell M.W., Harmon L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. N. Y. Acad. Sci. 1289:90–105.

Pennell M.W., Sarver B.A.J., Harmon L.J. 2012. Trees of unusual size: biased inference of early bursts from large molecular phylogenies. PLoS One. 7:e43348.

Penny D. 2013. Rewriting evolution--"been there, done that". Genome Biol. Evol. 5:819–21.

Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. Biol. Sci. 267:2267–72.

Pyron R.A., Burbrink F.T. 2013. Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. Trends Ecol. Evol. 28:729–36.

R Core Team. 2013. R: A Language and Environment for Statistical Computing.

Rabosky D. 2006a. LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. Evol. Bioinform. Online.247–250.

Rabosky D.L. 2006b. Likelihood methods for detecting temporal shifts in diversification rates. Evolution. 60:1152–64.

Rabosky D.L. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution. 64:1816–24.

Rabosky D.L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One. 9:e89543.

Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics. 13:235–238.

Revell L., Harmon L., Glor R. 2005. Under-parameterized Model of Sequence Evolution Leads to Bias in the Estimation of Diversification Rates from Molecular Phylogenies. Syst. Biol. 54:973–983.

Ricklefs R.E. 2007. Estimating diversification rates from phylogenetic information. Trends Ecol. Evol. 22:601–10.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M. a, Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–42.

Shavit L., Penny D., Hendy M.D., Holland B.R. 2007. The problem of rooting rapid radiations. Mol. Biol. Evol. 24:2400–11.

Stadler T. 2011. Simulating trees with a fixed number of extant species. Syst. Biol. 60:676–84.

Stadler T. 2013. How can we improve accuracy of macroevolutionary rate estimates? Syst. Biol. 62:321–9.

Weisrock D.W., Harmon L.J., Larson A. 2005. Resolving deep phylogenetic relationships in salamanders: analyses of mitochondrial and nuclear genomic data. Syst. Biol. 54:758–77.

Wertheim J.O., Sanderson M.J. 2011. Estimating diversification rates: how useful are divergence times? Evolution. 65:309–20.

Yule G.U. 1925. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. Philos. Trans. R. Soc. B Biol. Sci. 213:21–87.

Zuckerkandl E., Pauling L.B. 1962. Molecular disease, evolution, and genic heterogeneity. In: Kasha M., Pullman B., editors. Horizons in Biochemistry. New York City: Academic Press. p. 189–225.

Figures

Figure 2.1: Simulation workflow. λ is the instantaneous speciation rate, and r is the net diversification rate. Both are estimated for each set of simulation conditions.

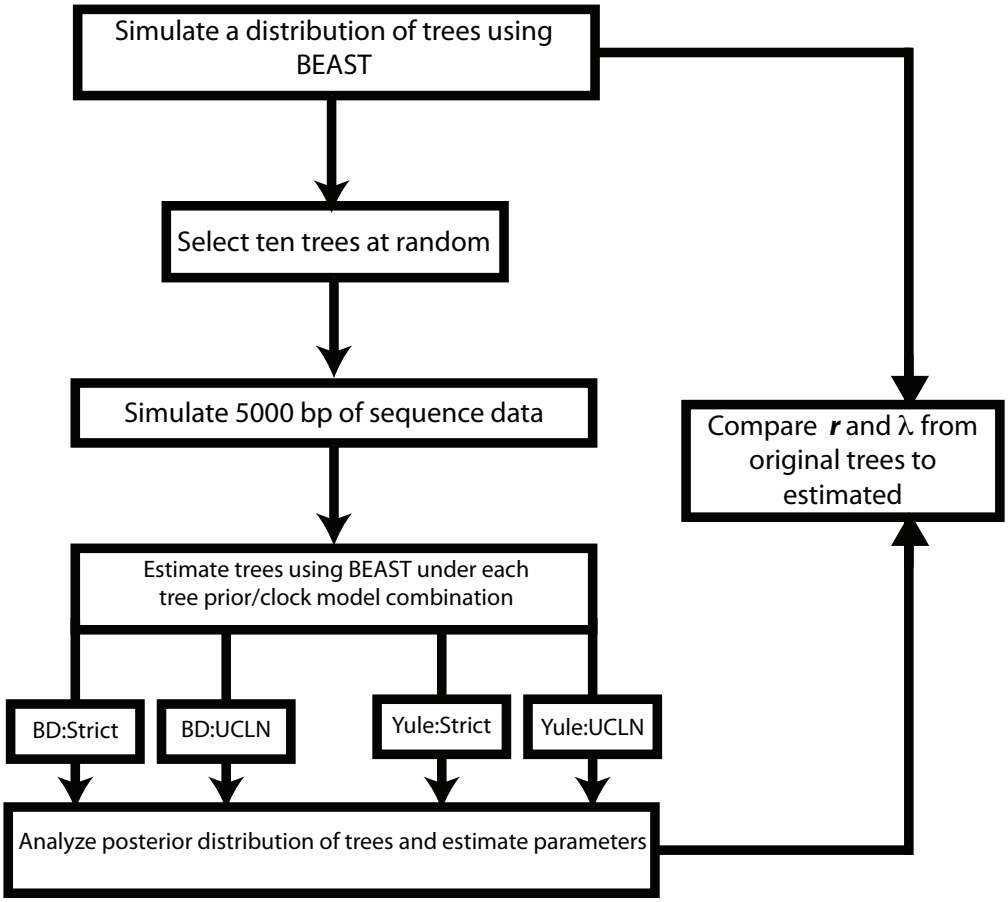


Figure 2.2: Yule simulations. The top row of plots refers to the 100-taxa cases, whereas the bottom row refers to the 25-taxa cases. Median estimates of λ or r, estimated from the 10 original trees, are displayed on each plot. The title of each subplot refers to the simulation conditions. Each combination of tree priors and molecular clocks under which trees are estimated is listed on the x-axis. The distribution of estimates from the original trees is also displayed. Parameter estimates are generally consistent with the original trees with slight deviations in some cases.

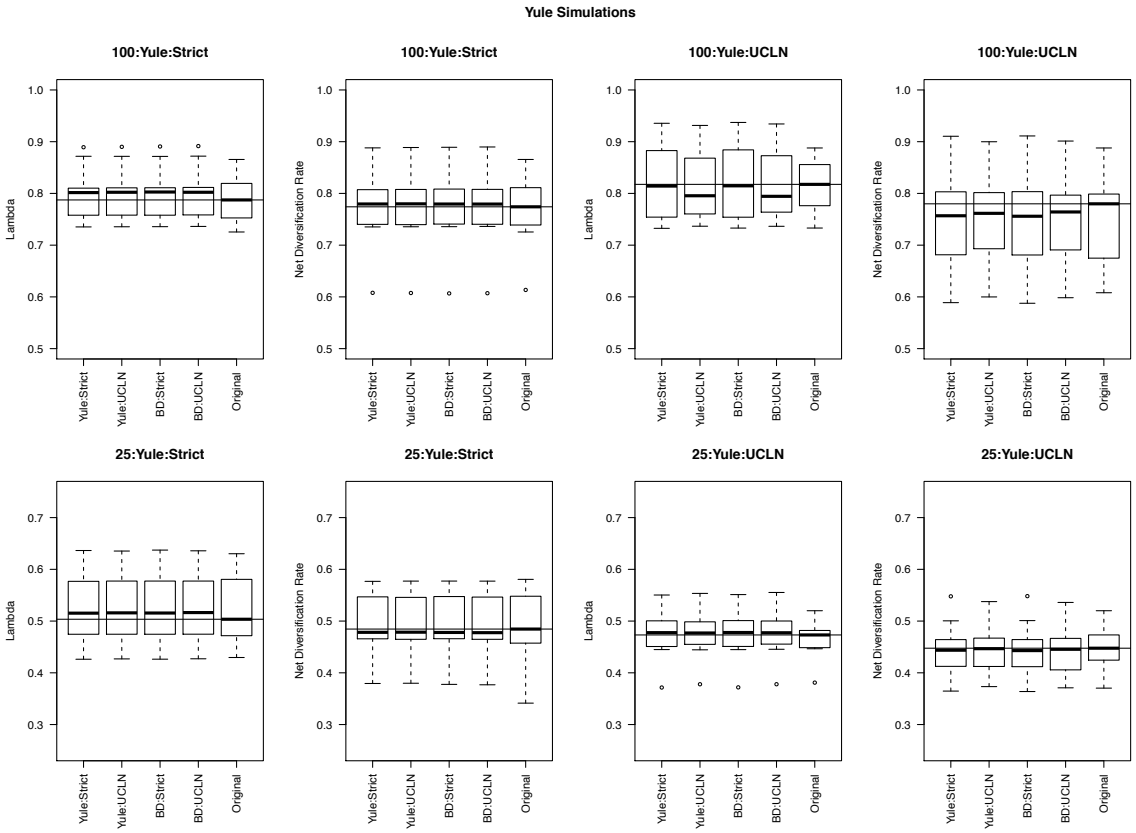


Figure 2.3: Lineage-through-time plots. The y-axis of each plot is natural-log transformed. Rows refer to conditions under which original trees are simulated, and columns refer to conditions under which trees are estimated. Thick gray lines represent the original trees and are, therefore, identical across each row of plots. Thin dark lines refer to the maximum clade credibility trees summarized from the posterior distribution of trees under the specified combination of tree prior and molecular clock. There is a significant amount of concordance, indicative of accurate phylogenetic estimation, though some discordance (indicated by non-overlapping lines) is present.

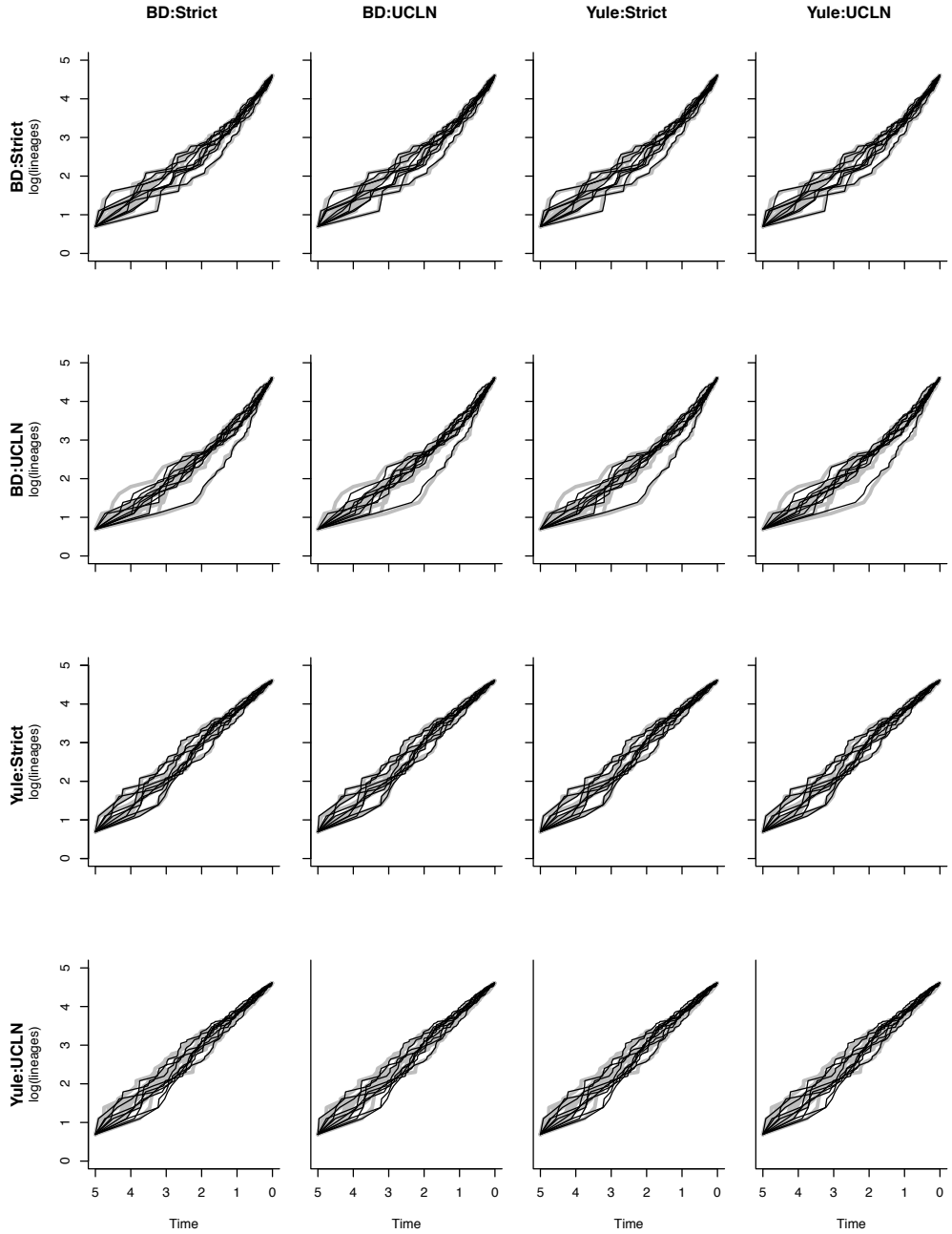


Figure 2.4: Birth-Death simulations. The top row of plots refers to the 100-taxa cases, whereas the bottom row refers to the 25-taxa cases. The median estimates of λ or r, estimated from the 10 original trees, is displayed on each plot. The title of each subplot refers to the simulation conditions. Each combination of tree priors and molecular clocks under which trees are estimated is listed on the x-axis. The distribution of estimates from the original trees is also displayed. Parameter estimates are highly congruent with the original trees under each set of simulation conditions.

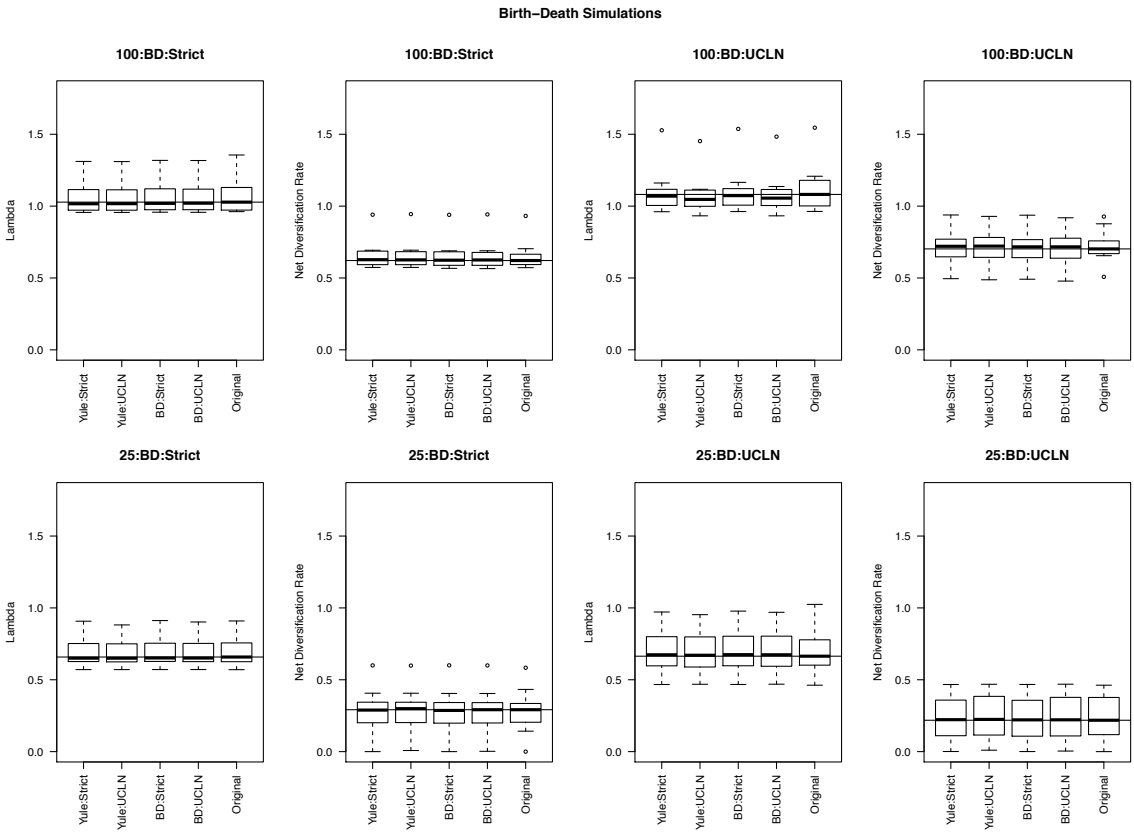
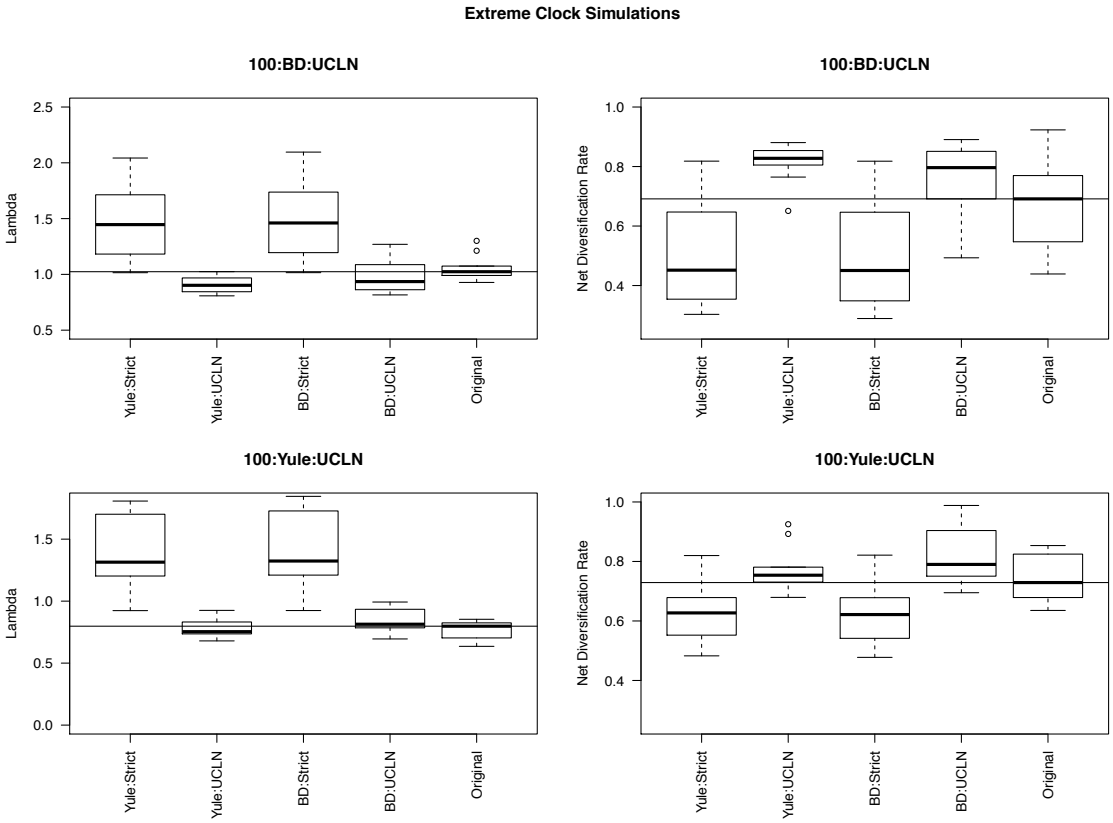


Figure 2.5: ‘Extreme’ clock prior simulations. A prior distribution of U(45, 55) was placed on the mean of the log-normal distribution from which rates are estimated under an uncorrelated log-normal relaxed molecular clock, and a prior of U(0.95, 1.05) was placed on the standard deviation. This results in more substantial rate heterogeneity per lineage than under the original simulation conditions. Misspecification of the tree prior and molecular clock reveals substantial differences in the parameter estimates produced using the original trees (simulated using an uncorrelated log-normal relaxed molecular clock) and the trees estimated using a strict clock. We attribute these trends to the inability of a single-rate strict clock to accurately account for substantial clock rate heterogeneity.



Chapter 3

Comparative Mitochondrial Phylogenomic Assessment of Introgression Among Several Species of Chipmunks

Abstract

Many well-characterized species are not completely reproductively isolated and hybridize, resulting in introgressive hybridization and offspring of mixed ancestry. Organellar genomes, such as those derived from mitochondria (mtDNA) and chloroplasts, introgress frequently in natural systems; however, the forces shaping patterns of introgression are not always clear. Here, we investigate extensive mtDNA introgression in western chipmunks, focusing on species in the *Tamias quadrivittatus* group from the central and southern Rocky Mountains. Specifically, we investigate the role of selection and demographic factors are factors in driving patterns of introgression. We sequenced 51 mtDNA genomes from six species and combine these sequences with other published genomic data to yield annotated mitochondrial reference genomes for nine species of chipmunks. Genomic characterization was performed using a series of molecular evolutionary and phylogenetic analyses to test protein-coding genes for positive selection. We fit a series of maximum likelihood models using a model-averaging approach, assessed deviations from neutral expectations, and performed additional tests to search for codons under positive or diversifying selection. We found no evidence for positive selection among these genomes and extensive evidence for negative or purifying selection, indicating that selection has not been the driving force of introgression in these species. Thus, extensive mtDNA introgression among several species of chipmunks likely reflects genetic drift of introgressed alleles in historically fluctuating populations.

Introduction

Interspecific hybridization occurs frequently in natural systems. Approximately 23% of animals exhibit mitochondrial DNA (mtDNA) polyphyly (Funk & Omland 2003), and approximately 10% of species are estimated to hybridize (Gray 1972; Mallet 2005). Although hybridization may promote genetic homogenization by collapsing incipient lineages into hybrid swarms (e.g. Taylor *et al.* 2006; Behm *et al.* 2010; Gilman & Behm 2011), it is becoming clear that gene flow between lineages may not prevent lineage divergence (Wu 2001; Pinho & Hey 2010). Models detailing this process are captured by the overarching term 'divergence-with-gene-flow' (e.g., Rice & Hostert, 1993) and describe the speciation process as an accumulation of reproductive barriers over time.

One of the most strongly recurring patterns in hybridizing systems is that organellar genomes (i.e., mitochondrial and chloroplast genomes) often introgress. Indeed, much of the evidence for hybridization in animals derives from studies of mtDNA introgression, which has been documented in many groups, including lizards, amphibians, birds, and mammals (e.g., Funk & Omland 2003; Mallet 2005; McGuire *et al.* 2007; Ryan *et al.* 2009; Rheindt & Edwards 2011; Johanet *et al.* 2011). The high frequency of mitochondrial introgression may result from selection on haplotypes in a novel genetic or ecological background (e.g., Llopart et al. 2014). Alternatively, genetic drift can promote haplotype fixation in small populations either through chance or in concert with selection following introgression (Ballard & Whitlock 2004). A special case of genetic drift is allele surfing, in which rare alleles present on the front of expanding populations may become fixed (Edmonds et al. 2004; Hallatschek et al. 2007; Hallatschek & Nelson 2008; Excoffier & Ray 2008); such

"surfing" alleles can either be deleterious, neutral, or advantageous (Klopfstein et al. 2006; Travis et al. 2007; Excoffier et al. 2009).

Yet another form of neutral introgression occurs when diverged populations come into contact due to expansion of one (or perhaps both); the alleles of the resident population can introgress into the expanding population (Currat & Excoffier 2004; Currat et al. 2008). This results in an asymmetric pattern of introgression, and empirical results provide support for this prediction; 36 of the 44 studies examined by Currat *et al.* (2008) involve allelic introgression from the resident population into the invading population. In this case, the extent of introgression appears to be governed jointly by allelic fitness, the rate of migration, and demographic stochasticity (Petit & Excoffier 2009).

Here, we focus on characterizing whole mitochondrial genomes in a group of hybridizing chipmunks (Sciuridae: *Tamias*), a diverse group of ground squirrels composed of 25 species (but see Piaggio & Spicer 2000, 2001). Of these, one species is restricted to eastern Asia (*T. sibiricus;* subgenus *Eutamias*), and one is restricted to eastern North America (*T. striatus*; subgenus *Tamias*). The remaining 23 species (subgenus *Neotamias*) are distributed throughout western North America. Assignment to species has relied on variation in the male genital bone, the baculum or *os penis*, with variation of other phenotypic characters (e.g., pelage and body size) showing considerable overlap among species.

Previous work in this system characterizes frequent mitochondrial introgression, which has been most thoroughly documented through two cases. First, several studies have documented asymmetric introgression of red-tailed chipmunk (*T. ruficaudus*) mtDNA genome into yellow-pine (*T. amoenus*) chipmunks (Good *et al.* 2003, 2008; Hird *et al.* 2010;

Reid *et al.* 2010, 2012), even though these species are rather distantly related phylogenetically (Reid *et al.* 2012). Second, more recent work documented widespread mtDNA introgression within the *T. quadrivittatus* species group, a group of six species that appears to have diverged within the last ca. 1.7 MY (Sullivan *et al.* 2014).

These previous studies used variation at a singe mtDNA gene (cyt *b*) to characterize introgression between species and therefore did not assess the molecular evolution of chipmunk mtDNA genome in the context of introgression. Here, we combine published genomic data from *T. amoenus* and *T. ruficaudus* (Bi et al. 2012) with sequencing of complete mtDNA genomes sampled from 52 individuals across the six species in the *T. quadrivittatus* group to characterizing representative mitochondrial genomes in these species. We use these data to assess the roles of selection and drift in explaining the distribution of introgressed mitochondrial genomes among species. Specifically, we investigate whether widespread introgression in *Tamias* is mediated by selection by explicitly testing for positive selection across the mtDNA genome.

Materials and Methods

Sample selection and extractions.

To date, approximately 1800 mtDNA Cyt *b* sequences have been generated (reviewed in Sullivan *et al.* 2014), nearly 300 of which have introgressed haplotypes. We selected 56 individuals from taxa that have been demonstrated to exhibit extensive mtDNA introgression (Good *et al.* 2003; Sullivan *et al.* 2014). These included *T. ruficaudus ruficaudus, T. r. simulans, T. amoenus luteiventris, T. a. canicaudus,* and *T. striatus* (one each, which were sequenced as part of a separate study; Bi et al. 2012), as well as 51 individuals from the *T. quadrivittatus* group (*T. canipes, T. rufus, T. quadrivittatus, T.*

cinereicollis, T. dorsalis, and T. umbrinus; see Supplementary Table 3.2 for collection localities). The *T. quadrivittatus*-group sample included both introgressed and non-introgressed individuals. DNA was isolated from heart or liver tissue using Qiagen DNEasy DNA extraction kits and eluted into ~ 50 *u*L of 10 mM Tris-Cl. Extractions were stored at -20°C prior to use.

Obtaining a draft Tamias *mitochondrial genome*

Prior to this study, no reference mitochondrial genome was available for *Tamias*. We therefore generated a preliminary reference via primer walking and Sanger sequencing. In order to design primers (using Primer3; Untergasser *et al.* 2007), we generated a consensus sequence from *Sciurus vulgaris* (a sciurid; Reyes *et al.* 2000), and *Glis glis* (a glirid; Reyes *et al.* 1998). Pairwise combinations of primers for PCR were utilized in 50 μL amplifications consisting of 39.3 μL of ddH_2O, 5 μL of 10X buffer, 1 μL of 10 mM dNTPs, 1.5 μL of 50 mM $MgCl_2$, 1 μL of a 100 μM solution of each primer, and 2 μL of genomic DNA. Amplification was performed on a BioRad MyCycler with the following parameters: 94°C for 2:00 followed by 45 cycles of 94°C for 0:30, 55°C for 0:45, and 72°C for a variable period depending on the size of the region amplified (assuming 1:00/1000bp). Each reaction was subject to a 5:00 final extension at 72°C and a 4°C final hold. Amplicons were cleaned using Qiagen PCR purification kits. Purified amplicons were prepared for cycle sequencing in 8 μL reactions consisting of 2 μL Big Dye, 1.6 μL 100 μM primer solution, and 4.4 μL purified amplicon. Cycle sequencing was performed on the same thermocycler under the following conditions: 96°C for 1:00, followed by 25 cycles of 96°C for 0:15, 50°C for 0:15, and 60°C for 4:00. Each sequencing reaction was subject to a 4°C hold. Sequencing reactions were cleaned using Sephadex spin columns, dried, and

suspended in 10 μ L of formamide for sequencing on an Applied Biosystems 3130 capillary sequencer. The final genome was checked for sequence quality and assembled into a working draft genome, using the *Sciurus* mitochondrial genome as a backbone. This draft mitochondrial genome was included as part of a set of approximately 12,000 exons that were used as baits for a targeted capture experiment.

Targeted capture and mitochondrial genome assembly

Fifty-one samples across of six species of chipmunks (*Tamias canipes, T. rufus, T. quadrivittatus, T. cinereicollis, T. dorsalis, and T. umbrinus*: the *T. quadrivittatus* group, Figure 1*)* were captured on an Agilent SureSelect microarray using a previously described protocol (Bi et al. 2012). Samples were sequenced on two lanes of an Illumina HiSeq 2000 at the Vincent Coates Genome Sequencing Laboratory at the University of California–Berkeley. In addition, raw data from five additional individuals (*T. r. ruficaudus, T. r. simulans, T. a. luteiventris, T. a. canicaudus,* and *T. striatus*) captured on the same microarrays as part of another study (Bi et al. 2012) were included in this analysis. Libraries were cleaned using SeqyClean (Zhbannikov et al. *in prep*.; http://bitbucket.org/izhbannikov/seqyclean) in order to remove low-quality reads, low-quality bases, and residual Illumina adapter sequences.

Cleaned reads were used as the starting point for *de novo* assembly using Assembly by Reduced Complexity v0.1 (ARC; Hunter et al. in prep.; current version available from http://github.com/ibest/ARC). Briefly, ARC identifies reads that are similar to a target of interest, places these reads into a reduced-representation pool, and performs a *de novo* assembly on the reduced pool of reads. This process is then iterated until no new reads are incorporated. This approach reduces biases that may be introduced through *de novo*

assembly of the entire library and/or by calling variants relative to a divergent reference in a different species (data not shown; Hunter et al. in prep.). The *Sciurus vulgaris* mitochondrial genome was used as the target sequence for assembly of a randomly selected *T. canipes* mitochondrial genome. This mitochondrial genome was then used as the target for all other assemblies in order to reduce reference bias and expedite assembly.

Following assembly, the resulting contig(s) were oriented to a common start site (tRNA-Phenylalanine/12S rRNA) by mapping and manual reorientation using Geneious Pro v6.1.7 (created by Biomatters; available from www.geneious.com). Protein-coding genes and rRNAs were identified by free-end alignment to *Sciurus* sequences downloaded from Genbank. Two datasets were produced: a complete dataset consisting of all data available across all individuals, and a pruned dataset where individuals missing all or a large portion of a gene were removed for that gene. Any individuals that had sequence removed for one gene may have been included for other genes if other gene sequences were complete. Due to incomplete sequencing of the control region in the draft genome, each genome was trimmed to 16,500 bp after multiple sequence alignment; as a result, the working genomes may have less than 16,500 bp. The final mitochondrial genome assemblies were also included as part of the complete dataset. A representative mitochondrial genome from each of the nine species (including the two subspecies of *T. ruficaudus* and two of *T. amoenus*) was used to calculate nucleotide frequencies using Geneious Pro. Sequences will be uploaded to Genbank before publication. An annotated version of the draft genome is included as part of the supplementary material (Supplementary Figure S1).

In order to provide an independent validation for the *de novo* assemblies of the mitochondrial genomes, the lysine transfer RNA (tRNA-Lys) was extracted from each

genome. This tRNA has an interesting evolutionary history within Mammalia (Dorner *et al.* 2001; see Results). A representative from each species and subspecies was annotated, extracted, and transcribed using Geneious Pro v6.1.7. The secondary structure was generated using the RNAfold web server (Hofacker 2003). The secondary structure of the *Mus musculus* tRNA-Lys was also visualized using the same approach.

Phylogenetic analyses

The 13 protein coding and two rRNA loci from both datasets were aligned in Mafft v6.86b using the G-INS-i (global homology) algorithm (Katoh & Toh 2008). A model of nucleotide sequence evolution was inferred for each locus using the decision theoretic approach implemented in DT-ModSel (Minin *et al.* 2003). Maximum-likelihood phylogenetic estimation under the inferred model were performed using Garli v1.0, with a score threshold of 0.01 and a requirement of 500,000 stable generations prior to termination (Zwickl 2006). We performed ten replicate searches, and convergence among searches was assessed by identification of identical or highly similar trees and similar log likelihood values. Trees from the complete dataset were used to assess topological incongruence among genes using the Shimodaira-Hasegawa (Shimodaira & Hasegawa 1999) test implemented in PAUP* v4.10b (Swofford 2003). These tests used the model of sequence evolution selected by DT-ModSel and calculated likelihoods using RELL bootstrapping with 1000 replicates. The tree generated from each locus was compared to the tree generated using the mitochondrial genome that was included in the set of alternative topologies among topologies estimated from each individual locus.

A maximum clade-credibility tree for the entire mitochondrial genome of the complete dataset was generated from a posterior distribution of trees estimated using

BEAST v1.7.5 (Drummond & Rambaut 2007; Drummond et al. 2012). Two identical runs were performed using a GTR+I+G model of nucleotide sequence evolution, a birth-death tree prior, and an uncorrelated lognormal clock prior. Chains were run for 50 million generations with samples taken every 5000. Convergence between runs was assessed through visual inspection of traces, comparison of mean parameter estimates, and verification that all ESS ~200 or greater. The posterior was summarized into a maximum clade credibility tree after removing 10% of samples as a burn-in using TreeAnnotator v1.7.5, with median node heights and a nodal probability cutoff of 0.5.

Finally, we estimated the mtDNA gene tree from 989 *Tamias* cytochrome *b* sequences obtained from Genbank via the PhyLoTA Browser (Sanderson *et al.* 2008). This tree includes individuals sampled from across the diversity of chipmunks, including the nine species in this study and the Siberian chipmunk (*Tamias sibiricus*). Where necessary, sequences were trimmed to include the 632 bp that are common across all individuals. The analysis was performed using RAxML v8.0.1 under the GTR+G model of sequence evolution with 1000 bootstrap replicates (Stamatakis 2014). All trees are deposited in TreeBASE.

Tests for selection

The pruned dataset was used to infer patterns of selection among loci using codeml in PAML v4.6 (Yang 2007) and a range of site models (M0, M1a, M2a, M3, M4, M5, M6, M7, M8, M8a). The simplest of these models is M0, which assumes a single ω (dN/dS) across all sites. The other models estimate the proportion of sites assigned to a range of site classes when the omegas corresponding to these classes are constrained. For example, M1a estimates the proportion of sites assigned to two classes, one where ω is fixed at one and

another where ω < 1. Alternatively, model M2a estimates the proportion of sites assigned to three classes, one where ω is fixed at one, another where ω is less than one, and a third where ω is greater than one. These models, and the others, are described in greater detail within the PAML manual.

Two approaches are commonly used to test for selection using these sets of models. The first is to select among the models using an information theoretic criterion, typically the Akaike Information Criterion. The second is to perform a series of likelihood-ratio tests (LRT) among nested models that allow sites to be placed into a class with a ω greater than one. Three likelihood-ratio tests were performed: M1a vs. M2a, M7 vs. M8, and M8 vs. M8a. Models M2a and M8 allow for assignment of codons to a site class where ω is greater than one and, therefore, allow for the detection of positive selection. M8a includes a site class where ω is fixed at one instead of being allowed to vary. The null distribution of the test statistic using these two models is a mixture and therefore differs from the standard chi-square distribution used to calculate p-values for the other two comparisons (see details in the PAML manual).

However, the LRT approach is not applicable to tests among other models (e.g. M3, M4, M5, and M6). In order to accommodate uncertainty in model choice into our assessment of selection, we implemented a model-averaging approach. (e.g., Sullivan & Joyce 2005). Although some authors (e.g., Posada & Buckley 2004) have advocated model averaging based on dAIC, the posterior probabilities calculated by Bayesian approaches represent a more explicit treatment of the model as a random variable (Sullivan & Joyce 2005). The posterior probability of a model can be approximated as:

$$P(M_i \mid D) \approx \frac{e^{(-BIC_i/2)}}{\sum_i^m e^{(-BIC_i/2)}},$$

where the summation in the denominator is across all *m* models in the candidate pool and $BIC_i = -2\ln(L_i) + k_i \ln(n)$.

Here, k_i is the number of free parameters in model *i*, and *n* is the sample size (Raftery 1995; this is usually approximated by sequence length). This approach assumes uniform (or vague; Schwarz 1978) prior probabilities across models. Furthermore, because ln*L* is typically calculated at its highest point (and therefore estimates a joint rather than marginal probability), this use of the BIC to approximate posterior probabilities assumes that the joint MLEs approximate the marginal likelihoods. Evans and Sullivan (2010) used reversible-jump MCMC to estimate model probabilities directly and assessed the usefulness of the BIC as an approximation of probabilities for models of nucleotide substation. They found that the approximation works well when there is much information in the data regarding model preference (Evans & Sullivan 2011). Therefore, we created a custom script (codemlMA.py, available from www.github.com/bricesarver/codemlma) to perform LRTs, calculate BICs, and approximate posterior probabilities of the models available in PAML. We then used this approach to derive model-averaged estimates of model parameters (κ and ω).

We also performed pairwise, multilocus McDonald-Kreitman tests (McDonald & Kreitman 1991) for each protein-coding gene combined using the Generalized and Standard McDonald-Kreitman test website (Egea et al. 2008). Mantel-Haenszel tests of homogeneity were performed to confirm equal rates among loci. If the rates were homogenous, multilocus McDonald-Kreitman p-values were calculated to determine whether there was a significant deviation from neutral expectations. Furthermore, we calculated three population-genetic statistics commonly used to test for neutrality: Tajima's D (Tajima 1989), Fu and Li's D_2*, and Fu and Li's F* (Fu & Li 1993). Each statistic was calculated per-species and per-locus,

including a combined alignment of all loci. These analyses were performed using the Intrapop Neutrality Tests web server (located at http://wwwabi.snv.jussieu.fr/achaz/neutralitytest.html).

Finally, we performed additional tests to look for positive or negative selection across all protein-coding loci in the *T. quadrivittatus* group. Five models were fit to attempt to identify codons under selection: SLAC (Single-Likelihood Ancestor Counting; Kosakovsky Pond & Frost 2005), FEL (Fixed Effects Likelihood; Kosakovsky Pond & Frost 2005), IFEL (Internal Fixed Effects Likelihood; Kosakovsky Pond et al. 2006), REL (Random Effects Likelihood; Kosakovsky Pond & Frost 2005), and FUBAR (Fast, Unconstrained Bayesian AppRoximation; Murrell et al. 2013). When applicable, these analyses utilized neighbor-joining trees with distances corrected using the model of nucleotide sequence evolution selected by DT-ModSel (see above). Furthermore, each codon determined to be under positive or diversifying selection, regardless of the method used, was subject to a series of molecular characterizations to identify amino acid changes, shifts in biochemical properties (using PRoperty Informed Models of Evolution, PRIME; Pond et al. 2005), and placement of the amino acid within the protein. All analyses were performed using the Datamonkey web server (Kosakovsky Pond & Frost 2005), a publically-accessible front-end to a cluster computing system running HyPhy (Kosakovsky Pond et al. 2005).

Results

Characteristics of the chipmunk mitochondrial genome

Chipmunk mtDNA genomes exhibit syntenic conservation with other mammalian mtDNA genomes. 36% (+/- 0.9%) of the positive strand is composed of cytosine or guanine

nucleotides. This strand also has a 46:54 purine:pyrimidine bias with a standard deviation of 0.006 across all species. Nucleotide frequencies are similar across species (Table 3.1). Lengths of protein-coding genes are largely conserved, but there is some interspecific length variation in the control region and ribosomal RNAs associated with indels in loop regions. Several genes are missing one or two nucleotides that make up the final stop codon (see Table 3.2; complete vs. pruned lengths). In *Sciurus* and other mammals, the stop codon appears to be completed through the polyadenylation of pre-mRNAs (Reyes et al. 2000; Chang & Tong 2012). *Tamias* also appears to require this modification, as several exons do not end in the appropriate stop codon for mammalian species.

Twelve of the 13 protein-coding genes have estimates of κ between 8 and 17, implying that the number of transitions is around an order of magnitude higher than the number of transversions across all samples. Interestingly, the κ of ND3 is estimated at 123.8, slightly more than an order of magnitude higher than the other loci (Table 3.2). ND3 is one of the shortest genes, consisting of 345 nucleotides without the stop codon present. There are seven amino-acid substitutions in *T. striatus* relative to *T. ruficaudus ruficaudus* at this locus, higher than the mean among all species (2.2 +/- 1.8; Table 3). All *T. striatus* loci exhibit a greater number of substitutions, as expected given the p-distance of *T. striatus* relative to the other species (~0.16, Table 3.4).

One interesting case involves tRNA-Lys, the transfer RNA sequenced as part of a validation of mitochondrial assembly described above. Within Mammalia, monotremes and eutherians possess a functional tRNA-Lys that is synthesized from the mitochondrial genome and then lysylated by a lysine-tRNA ligase. In metatherians, the mitochondrial genome lacks a functional tRNA-Lys (Dorner et al. 2001; Axel et al. 1994). Instead, this

tRNA is synthesized from nuclear DNA. In these cases, it acts as a functional pseudogene in the mitochondrial genome due to a relaxation of selective constraint. We found that, as expected, chipmunks possess functional mitochondrial tRNA-Lysines. The secondary structures exhibit the canonical cloverleaf shape, and the appropriate anticodon for tRNA-Lys (UUU) is present in the anticodon loop (Figure 3.2). Furthermore, substitutions and insertions are present in areas that do not affect the secondary structure of the RNA, as expected for a locus under strong purifying selection. *Tamias* tRNA-Lysines have one, two, three, or four bp inserted in the first loop (the "D-loop") relative to *Mus*, but all other intramolecular base pairings are maintained. Positional probabilities are high, indicating good structural estimates. Additionally, resolution of this tRNA provides evidence of accurate assembly.

Phylogenetic inference

DT-ModSel selected two- or three-parameter models with or without gamma-distributed rate heterogeneity or a proportion of invariable sites. More complex models (TIM+I+ Γ and GTR+I+ Γ) were selected for rRNAs and the complete genome. The same models were selected within the complete or pruned datasets and regardless of whether the stop codon (or trailing bases) was removed from the sequence. Phylogenetic analysis of complete mtDNA genomes is broadly congruent with the analysis of 989 cytochrome *b* sequences and the tree estimated in Sullivan et al. (2014) (Figure 3.3, Supplementary Figure 3.2, Supplementary Figure 3.3).

The yellow-pine (*T. amoenus*) and red-tailed (*T. ruficaudus*) chipmunks are recovered as sister to the six *T. quadrivittatus* group species. Within this clade, *T. canipes* and *T. rufus* are each recovered as monophyletic (PP = 1). The other four species (*T.*

umbrinus, T. cinereicollis, T. dorsalis, and *T. quadrivittatus*) have genomes dispersed throughout the tree with high posterior nodal support. Because species were assigned using morphology, the complete mtDNA genome tree is in agreement with other studies (e.g., Good *et al.* 2003; Sullivan *et al.* 2014) in indicating extensive mitochondrial introgression.

SH tests reveal that each gene tree is not significantly different from the combined tree (Table 3.2). These results support the interpretation that mitochondrial genomes can be treated as a single marker due to linkage, a hypothesis is rarely tested in empirical datasets. Since a single underlying topology is supported among all loci, using the complete genome for phylogenetic inference is appropriate. It also confirms that previous studies that used cytochrome b or cytochrome oxidase subunit II as the sole marker should have recovered the correct underlying mitochondrial tree. This provides support for previous, single-locus studies in chipmunks.

Selection analyses

Likelihood-ratio tests between models M1a and M2a detect no signatures of selection. In contrast, a significant difference between M7 and M8 is detected for COII ($p < 0.01$), and marginally significant differences occur between M8 and M8a for ATP8 ($p = 0.087$), COII ($p = 0.092$), and ND3 ($p = 0.0548$), though these differences disappear when corrections for multiple testing are applied (Supplementary Table 3.1). Estimating the posterior probability of each model is illuminating. In many cases, high posterior probabilities are assigned to models M1, M5, and M7, models that do not include a site class where $\omega > 1$. In cases where a site is assigned to a class with $\omega > 1$, the posterior probability of that model is no more than 0.03. Investigating model fit by traditional likelihood-ratio tests, in concert with Bayesian model averaging, allows for a more confident inference of

patterns of selection; this approach could be used in other studies where the posterior probabilities of a particular site being under selection can be inferred with greater confidence.

Pairwise McDonald-Kreitman tests reveal no significant deviations from neutrality except in two cases involving *T. canipes* (Supplementary Table 3.3). The significant *p*-values are most likely artifacts resulting from small sample size, and Tajima's D, Fu and Li's D_2*, and Fu and Li's F* provide support for this claim (Supplementary Table 3.4). For many *T. canipes* and *T. rufus* loci, there are not enough substitutions to produce accurate estimates. No species has a significant deviation from neutrality at an individual locus or with all loci combined (with one exception: ND4L for *T. dorsalis* for Tajima's D).

Many of our analyses (SLAC, FEL, IFEL, REL, and FUBAR) detected pervasive negative or purifying selection across all genes (Supplementary Table 3.5). In a handful of cases, codons were classified as being under positive or diversifying selection. Upon further inspection, including biochemical characterization, many appear to be false-positives (Supplementary Material). There is marginal evidence for a single codon under positive or diversifying selection in *T. dorsalis* at the COII gene, though three out of five methods do not identify this as a site of interest.

Discussion

Mitochondrial introgression is often described in natural systems, and the frequency at which it occurs is surprising. Since mitochondrial genomes contain protein-coding genes that act as part of the oxidative-phosphorylation pathway, as well as rRNAs that bind with ribosomal proteins encoded in the nuclear genome (and imported into the mitochondria) for proper ribosomal assembly, it is reasonable to assume that antagonistic epistatic interactions

with nuclear-encoded proteins from divergent lineages could reduce fitness. Selection, then, ought to purge substitutions that lead to deleterious mitonuclear interactions (reviewed in Dowling et al. 2008). However, the mitochondrial genome has the potential to act as a source of novel genetic variation, often resulting in the presence of multiple haplotypes within a population (e.g., Dowling et al. 2007).

The population genetic dynamics of mtDNA depend on relative fitness in concert with demographic factors. Haplotypes that confer high fitness may rise to a high frequency in a population following a selective sweep (e.g., Haldane 1924; Barton 2000). Alternatively, genotypes may invade resident populations through demographic means such as population expansion. Theory has shown that this results in asymmetric introgression of alleles from the resident population into the invading population (Excoffier et al. 2009). Furthermore, alleles at the front of an advancing population may also become fixed due to local effects (Excoffier & Ray 2008; Hallatschek & Nelson 2008). Because these effects are at least partly stochastic, it may be impossible to predict the resulting dynamics *a priori*.

We did not detect positive selection acting on mitochondrial protein-coding genes. While these genes do have polymorphic sites, consistently low values of model-averaged estimates of omega suggest that purifying or negative selection plays a principal role in preserving amino acid similarity. McDonald-Kreitman tests, Tajima's D, Fu and Li's D_2*, and Fu and Li's F* detect almost no significant deviations from neutrality, and any significant results can be explained by multiple testing artifacts. Additionally, support for pervasive negative or purifying selection is widespread across five additional analyses (SLAC, FEL, IFEL, REL, and FUBAR), with little-to-no support for positive or diversifying selection. However, it could be the case that positive selection does act on the mitochondrial

genome, just not on the 13 protein-coding genes (e.g., Melo-Ferreira et al. 2014). Ribosomal RNAs interact with a number of nuclear proteins to construct a functional ribosome, suggesting that selection could act on these genes. In addition, the control region contains several promoter regions and an origin of replication, any of which could viably be targets for positive selection.

In the absence of evidence for positive selection, we accept demographic factors as a likely explanation that governs the distribution of mtDNA. *T. dorsalis* provides support for this conclusion. It is broadly distributed, and its range overlaps with several other species. Interestingly, whenever it co-occurs with another species, it receives a mitochondrial genome from its congener. It is reasonable that *T. dorsalis* expanded into populations consisting of other species and, as a result, inherited those species' mitochondrial genomes. Ancestral niche modeling by Waltari and Guralnick (2009) revealed that *T. dorsalis* has undergone a northward range expansion since the Last Glacial Maximum resulting in recent contact with other species; *T. dorsalis*-specific mtDNA is only found in southern refugial areas. Furthermore, *T. dorsalis* has consistently negative values of Tajima's D at each locus and all loci combined, suggestive of population expansion.

However, the northern Rocky Mountains introgression of *T. ruficaudus ruficaudus* into *T. amoenus luteiventris* (Good et al. 2003) may not follow this pattern. In this case, it appears that the *T. ruficaudus ruficaudus* mtDNA is introgressing into *T. amoenus luteiventris*. However, because the location of this introgression (along the crest of the Canadian Rockies) did not support forested habitats until just a few thousand years ago (Mack et al. 1978). *T. amoenus luteiventris* has been undergoing population expansion in a similar fashion to *T. ruficaudus ruficaudus*; differential rates of expansion could produce the

observed pattern of introgression. A complete characterization of the dynamics of this introgressive event would require mitochondrial genome sequencing from several individuals within and around this contact zone.

It may also be the case that we cannot accurately estimate ω with a dataset of this size and across such recently diverged genomes (~2.5 MYa; Sullivan et al. 2014). These analyses are likely influenced by the length of the sequence, the number of individuals, and the amount of sequence divergence among individuals (Anisimova et al. 2001, 2002; Wong et al. 2004). Thus, it may not be possible to confidently estimate ω with modestly divergent sequences relative to analyses of taxa that span deeper divergences.

Additionally, we develop a model-averaging approach that can be used to investigate patterns of selection in genomic-scale datasets. Estimating posterior probabilities for a range of models provides more information than discriminating among models using δAIC or likelihood ratio tests alone. Furthermore, posterior probabilities can be used to weight parameters of interest, such as kappa or omega, estimated under several models and mitigate biases that may be introduced from using a single model. We suggest use of the model-averaging approach described above account explicitly for uncertainty in model selection, especially with regard to detecting positive selection when using likelihood-based approaches.

Conclusion

We sequence and characterize mitochondrial genomes for several species of chipmunks. Protein-coding genes are analyzed and used to test demographic vs. selection hypotheses governing introgression. We find no evidence for positive selection at mitochondrial loci and conclude that introgression is mediated by demographic factors in this system. Future analyses will focus on quantifying the amount of nuclear introgression taking place in this complex system, building on the characterizations in this study.

References

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19:950–8.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–92.

Axel J, Feldmaier-Fuchs G, Thomas WK, von Haeseler A, Paabo S. 1994. The marsupial mitochondrial genome and the evolution of placental mammals. Genetics. 137:243–256.

Ballard JWO, Whitlock MC. 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13:729–744.

Barton NH. 2000. Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 355:1553–62.

Behm JE, Ives AR, Boughman JW. 2010. Breakdown in postmating isolation and the collapse of a species pair through hybridization. Am. Nat. 175:11–26.

Bi K et al. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 13:403.

Chang JH, Tong L. 2012. Mitochondrial poly(A) polymerase and polyadenylation. Biochim. Biophys. Acta. 1819:992–7.

Currat M, Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. PLoS Biol. 2:e421.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. Evolution. 62:1908–20.

Dorner M, Altmann M, Paabo S, Morl M. 2001. Evidence for Import of a Lysyl-tRNA into Marsupial Mitochondria. Mol. Biol. Cell. 12:2688–2698.

Dowling DK, Friberg U, Hailer F, Arnqvist G. 2007. Intergenomic epistasis for fitness: within-population interactions between cytoplasmic and nuclear genes in Drosophila melanogaster. Genetics. 175:235–44.

Dowling DK, Friberg U, Lindell J. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. Trends Ecol. Evol. 23:546–54.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–73.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. Proc. Natl. Acad. Sci. U. S. A. 101:975–9.

Egea R, Casillas S, Barbadilla A. 2008. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. Nucleic Acids Res. 36:W157–62.

Evans J, Sullivan J. 2011. Approximating model probabilities in Bayesian information criterion and decision-theoretic approaches to model selection in phylogenetics. Mol. Biol. Evol. 28:343–9.

Excoffier L, Foll M, Petit RJ. 2009. Genetic Consequences of Range Expansions. Annu. Rev. Ecol. Evol. Syst. 40:481–501.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol. Evol. 23:347–51.

Fu Y-X, Li W-H. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly : Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.

Gilman RT, Behm JE. 2011. Hybridization, species collapse, and species reemergence after disturbance to premating mechanisms of reproductive isolation. Evolution. 65:2592–605.

Good JM et al. 2008. Ancient hybridization and mitochondrial capture between two species of chipmunks. Mol. Ecol. 17:1313–27.

Good JM, Demboski JR, Nagorsen DW, Sullivan J. 2003. Phylogeography and introgressive hybridization: chipmunks (genus Tamias) in the northern Rocky Mountains. Evolution. 57:1900–16.

Gray AP. 1972. *Mammalian hybrids: a check-list with bibliography*. 2nd ed. Commonwealth Agricultural Bureau: Slough.

Haldane JBS. 1924. A mathematical theory of natural and artificial selection—I. Trans. Cambridge Philos. Soc. 23:19–41.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR. 2007. Genetic drift at expanding frontiers promotes gene segregation. Proc. Natl. Acad. Sci. U. S. A. 104:19926–30.

Hallatschek O, Nelson DR. 2008. Gene surfing in expanding populations. Theor. Popul. Biol. 73:158–70.

Hird S, Reid N, Demboski J, Sullivan J. 2010. Introgression at differentially aged hybrid zones in red-tailed chipmunks. Genetica. 138:869–83.

Hofacker IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31:3429–3431.

Johanet A, Secondi J, Lemaire C. 2011. Widespread introgression does not leak into allotopy in a broad sympatric zone. Heredity. 106:962–72.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9:286–98.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. Mol. Biol. Evol. 23:482–90.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208–22.

Kosakovsky Pond SL, Frost SDW. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 21:2531–3.

Kosakovsky Pond SL, Frost SDW, Muse S V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21:676–9.

Kosakovsky Pond SL et al. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput. Biol. 2:e62.

Llopart A, Herrig D, Brud E, Stecklein Z. 2014. Sequential adaptive introgression of the mitochondrial genome in Drosophila yakuba and Drosophila santomea. Mol. Ecol. 23:1124–36.

Mack RN, Rutter NW, Bryant VM, Valastro S. 1978. Reexamination of postglacial vegetation history in northern Idaho: Hager Pond, Bonner Co. Quat. Res. 10:241–255.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–37.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 351:652–4.

McGuire JA et al. 2007. Mitochondrial introgression and incomplete lineage sorting through space and time: phylogenetics of crotaphytid lizards. Evolution. 61:2879–97.

Melo-Ferreira J et al. 2014. The elusive nature of adaptive mitochondrial DNA evolution of an arctic lineage prone to frequent introgression. Genome Biol. Evol. 6:886–96.

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. Syst. Biol. 52:674–683.

Murrell B et al. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol. Biol. Evol. 30:1196–205.

Petit RJ, Excoffier L. 2009. Gene flow and species delimitation. Trends Ecol. Evol. 24:386–93.

Piaggio A, Spicer G. 2000. Molecular phylogeny of the chipmunk genus Tamias based on the mitochondrial cytochrome oxidase subunit II gene. J. Mamm. Evol. 7.

Piaggio AJ, Spicer GS. 2001. Molecular phylogeny of the chipmunks inferred from Mitochondrial cytochrome b and cytochrome oxidase II gene sequences. Mol. Phylogenet. Evol. 20:335–50.

Pinho C, Hey J. 2010. Divergence with Gene Flow: Models and Data. Annu. Rev. Ecol. Evol. Syst. 41:215–230.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.

Raftery AE. 1995. Bayesian Model Selection in Social Research. Sociol. Methodol. 25:111–163.

Reid N, Demboski JR, Sullivan J. 2012. Phylogeny estimation of the radiation of western North American chipmunks (Tamias) in the face of introgression using reproductive protein genes. Syst. Biol. 61:44–62.

Reid N, Hird S, Schulte-Hostedde A, Sullivan J. 2010. Examination of nuclear loci across a zone of mitochondrial introgression between Tamias ruficaudus and T. amoenus. J. Mammal. 91:1389–1400.

Reyes A, Gissi C, Pesole G, Catzeflis FM, Saccone C. 2000. Where Do Rodents Fit? Evidence from the Complete Mitochondrial Genome of Sciurus vulgaris. Mol. Biol. Evol. 17:979–983.

Reyes A, Pesole G, Saccone C. 1998. Complete mitochondrial DNA sequence of the fat dormouse, Glis glis: further evidence of rodent paraphyly. Mol. Biol. Evol. 15:499–505.

Rheindt FE, Edwards S V. 2011. Genetic Introgression : An Integral but Neglected Component of Speciation in Birds. The Auk 128:620–632.

Rice WR, Hostert EE. 1993. Laboratory Experiments on Speciation: What Have We Learned in 40 Years? Evolution 47:1637.

Ryan ME, Johnson JR, Fitzpatrick BM. 2009. Invasive hybrid tiger salamander genotypes impact native amphibians. Proc. Natl. Acad. Sci. U. S. A. 106:11166–71.

Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A. 2008. The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. Syst. Biol. 57:335–46.

Schwarz G. 1978. Estimating the Dimension of a Model. Ann. Stat. 6:461–464.

Shimodaira H, Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. 16:1114–1116.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2010–2011.

Sullivan J et al. 2014. Divergence-with-gene-flow within the recent chipmunk radiation (Tamias). Heredity. In press.

Sullivan J, Joyce P. 2005. Model Selection in Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.

Swofford DL. 2003. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–95.

Taylor EB et al. 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (Gasterosteus aculeatus) species pair. Mol. Ecol. 15:343–55.

Travis JMJ et al. 2007. Deleterious mutations can surf to high densities on the wave front of an expanding population. Mol. Biol. Evol. 24:2334–43.

Untergasser A et al. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 35:W71–4.

Waltari E, Guralnick RP. 2009. Ecological niche modelling of montane mammals in the Great Basin, North America: examining past and present connectivity of species across basins and ranges. J. Biogeogr. 36:148–161.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–51.

Wu C-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–91.

Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

Figures

Figure 3.1: Sampling localities of the 51 *T. quadrivittatus* group individuals sequenced as part of this study. Individuals are color-coded based on their bacular species assignments.

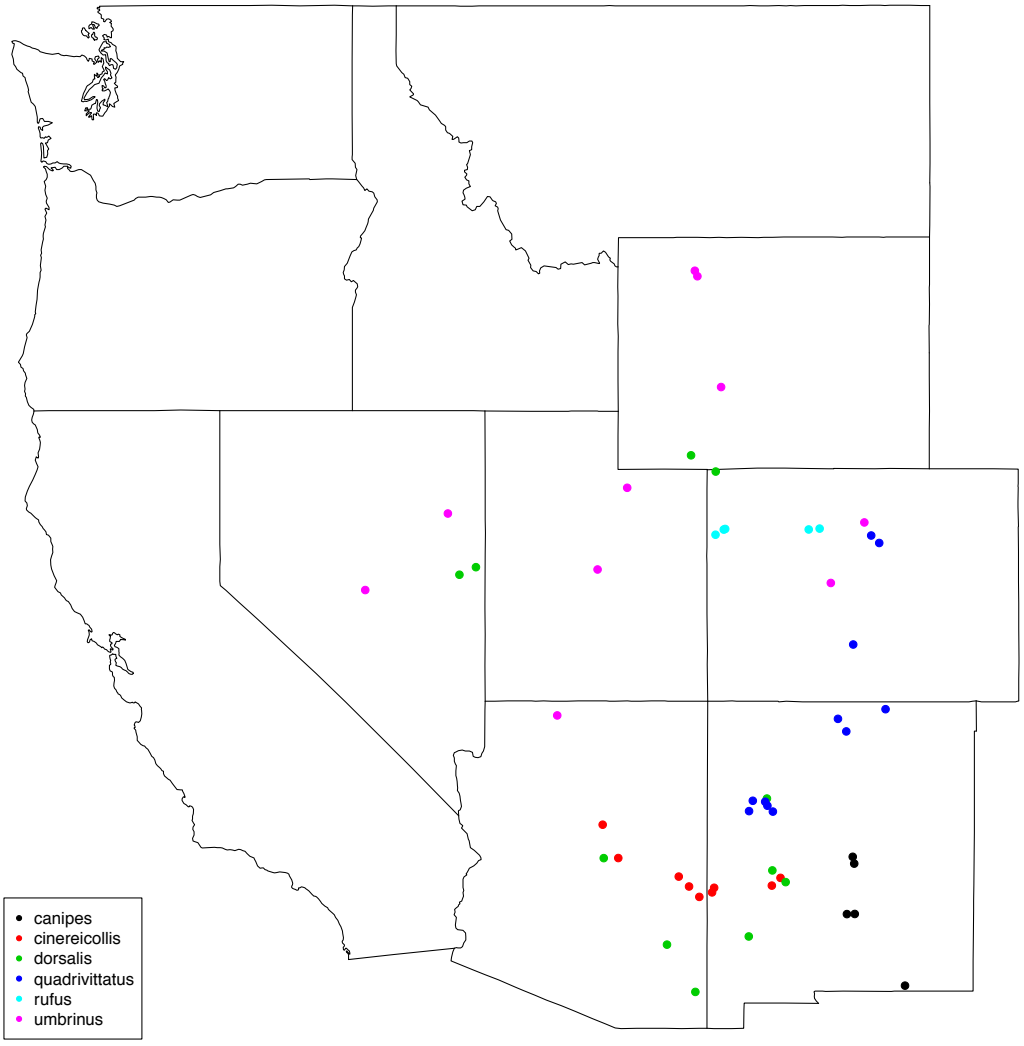


Figure 3.2: Secondary structures of tRNA-Lysine for each of the nine species in this study. Secondary structures are visualized using the RNAfold server (Hofacker 2003). Colors refer to per-base positional probabilities with blue representing 0 and red representing 1. The secondary structure estimated for *Mus* is also included for reference.

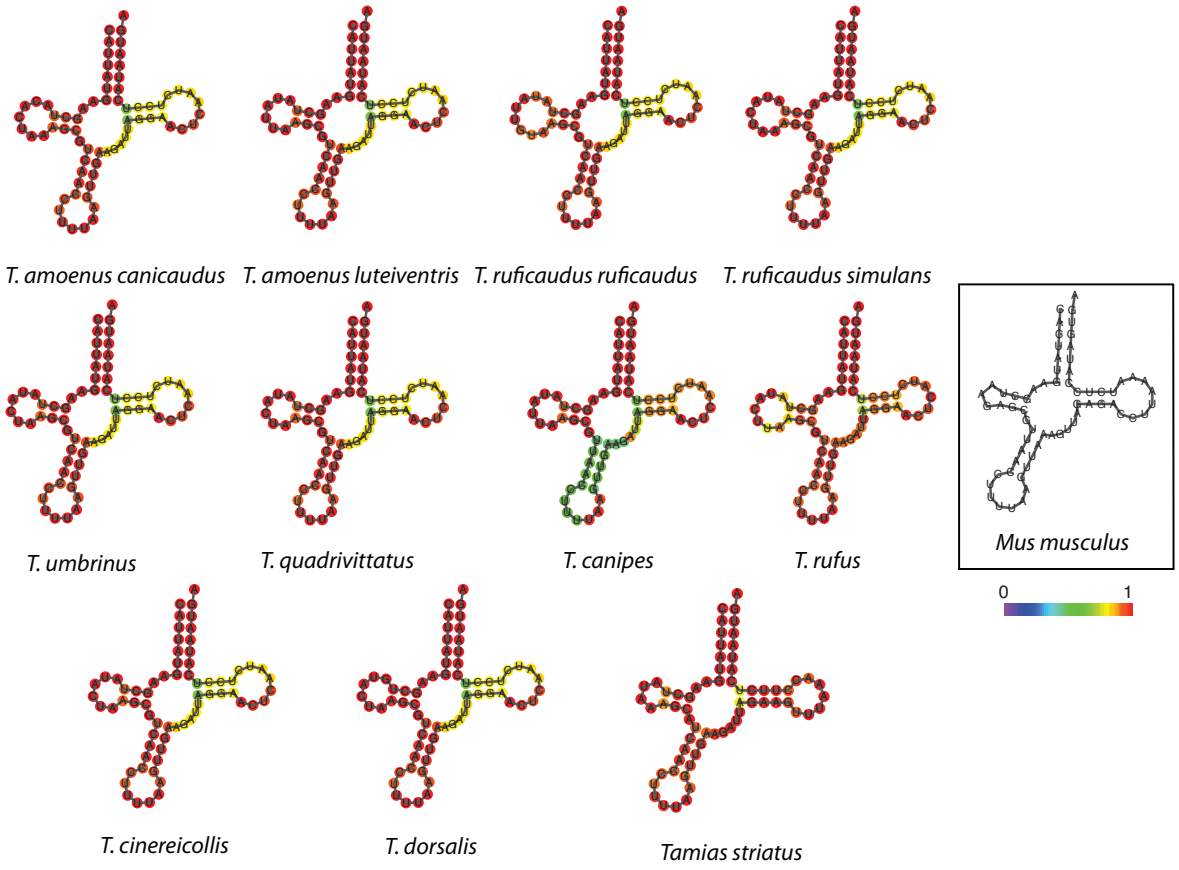
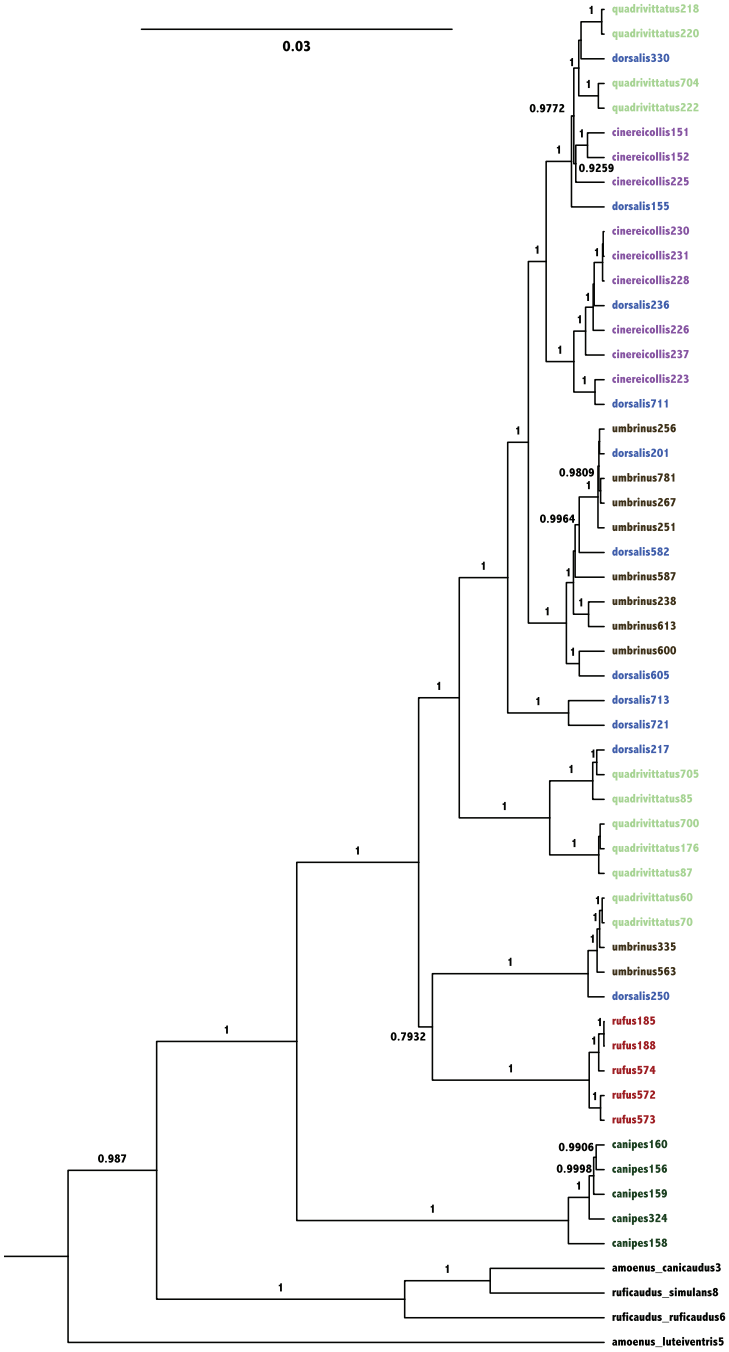


Figure 3.3: Maximum clade credibility tree estimated for the *T. quadrivittatus* group using all mitochondrial sequence data. Each node with a posterior probability greater than 0.7 is annotated. Colors correspond to bacular species assignments, which are also indicated as part of the individual identifier.



Tables

Table 3.1: Descriptive characteristics of chipmunk mitochondrial genomes. A representative individual of each species is used and is identified by the ID column. Genomes are trimmed to conservatively account for assembly errors. Nucleotide frequencies are listed in concert with G/C bias and purine/pyrimidine bias on the positive strand.

Species	ID	A	C	G	T	Total Length	G/C	A/G
T. canipes	324	5516	3963	1993	4974	16446	0.36	0.46
T. cinereicollis	223	5501	3968	2016	4959	16444	0.36	0.46
T. dorsalis	250	5491	3998	2022	4934	16445	0.37	0.46
T. quadrivittatus	220	5505	3979	2008	4951	16443	0.36	0.46
T. rufus	573	5493	3978	2010	4958	16439	0.36	0.46
T. umbrinus	238	5498	3980	2014	4952	16444	0.36	0.46
T. a. canicaudus	3	5495	3915	2008	5027	16445	0.36	0.46
T. a. luteiventris	5	5496	3232	2010	4987	15725	0.33	0.48
T. r. ruficaudus	6	5477	3959	2022	4989	16447	0.36	0.46
T. r. simulans	8	5475	3944	2026	5000	16445	0.36	0.46
T. striatus	11	5469	3746	2025	5108	16348	0.35	0.46
Mean		5492.364	3878.364	2014.000	4985.364	16370.091	0.360	0.459
Standard Deviation		13.894	225.245	9.706	48.486	215.901	0.009	0.006

Table 3.2: Per-locus characterization. Model refers to the model of nucleotide sequence evolution selected by DT-ModSel. The number of sites, with and without the complete stop codon, refers to the number of nucleotides per-locus. Shimodaira-Hasegawa p-values are listed for the comparison to the complete mitochondrial genome. Model-averaged estimates of κ and ω are weighted relative to the posterior probability of the model under which they were estimated. Both estimates of ω are averaged across the entire locus.

Locus	Model (both datasets)	Number of Sites (complete)	Number of Sites (pruned)	SH p-value	Kappa (Weighted)	Omega (one-rate)	Omega (model-averaged)
12s rRNA	TIM+I+G	974	-	0.925	-	-	-
16s rRNA	GTR+I+G	1595	-	0.933	-	-	-
ATP6	HKY+G	680	678	0.921	8.097	0.02258	0.02519
ATP8	HKY+G	204	201	0.752	12.595	0.16823	0.20920
COI	HKY+G	1542	1539	0.93	13.351	0.00186	0.00220
COII	HKY+G	684	681	0.966	17.587	0.00471	0.00725
COIII	HKY+G	784	783	0.986	12.892	0.01302	0.02068
CytB	HKY+I	1140	1137	0.89	13.612	0.01492	0.01986
ND1	HKY+I	956	954	0.939	11.033	0.01347	0.01434
ND2	HKY+G	1042	1041	0.907	15.95	0.04409	0.06193
ND3	HKY+I	347	345	0.896	123.816	0.03093	0.06398
ND4H	HKY+G	1378	1377	0.982	12.349	0.02592	0.03088
ND4L	K81uf+I	297	294	0.92	11.312	0.04385	0.05129
ND5	K81uf+I	1818	1815	0.988	8.101	0.03971	0.04662
ND6	HKY+G	525	522	0.871	12.966	0.08626	0.08955
Complete	GTR+I+G	16500	-	-	-	-	-

Table 3.3: Pairwise nonsynonymous amino acid substitutions. Differences were calculated relative to *T. ruficaudus ruficaudus*. AA Gap refers to the number of amino acids that were not able to be resolved in *T. striatus* due to a lack of sequencing depth at some positions; these were conservatively counted as mismatches.

Locus	T. canipes	T. cinereicollis	T. dorsalis	T. quadrivittatus	T. rufus	T. umbrinus	T. amoenus canicaudus	T. amoenus luteiventris	T. ruficaudus simulans	T. striatus	Mean	SD	AA Gap
ATP6	1	2	1	1	2	2	0	1	1	7	1.8	1.93	-
ATP8	2	2	2	3	3	3	1	4	1	7	2.8	1.75	-
COI	0	0	0	0	0	0	0	1	0	2	0.3	0.67	-
COII	1	1	1	2	2	1	0	1	0	8	1.7	2.31	3
COIII	3	1	2	2	1	1	2	2	2	9	2.5	2.37	15
CytB	5	4	7	5	5	4	4	6	2	14	5.6	3.24	-
ND1	5	3	4	2	2	2	2	2	1	27	5	7.82	-
ND2	12	9	14	9	11	9	4	12	2	47	12.9	12.53	-
ND3	2	2	2	2	3	2	1	2	0	7	2.3	1.83	-
ND4H	4	6	3	4	5	4	6	7	3	33	7.5	9.06	-
ND4L	3	1	2	1	2	1	1	2	1	10	2.4	2.76	-
ND5	17	18	17	15	17	17	10	13	10	55	18.9	13.03	-
ND6	10	14	10	13	11	11	3	12	3	28	11.5	6.92	-
Total	65	63	65	59	64	57	34	65	26	254	75.2	64.35	-

Table 3.4: Pairwise *p*-distances among representative mitochondrial genomes for each species.

	T. canipes	T. cinereicollis	T. dorsalis	T. quadrivittatus	T. rufus	T. umbrinus	T. amoenus canicaudus	T. amoenus luteiventris	T. ruficaudus ruficaudus	T. ruficaudus simulans	T. striatus
T. canipes	-										
T. cinereicollis	0.049	-									
T. dorsalis	0.048	0.03	-								
T. quadrivittatus	0.05	0.013	0.03	-							
T. rufus	0.047	0.029	0.026	0.03	-						
T. umbrinus	0.049	0.014	0.028	0.014	0.027	-					
T. amoenus canicaudus	0.064	0.062	0.06	0.063	0.062	0.062	-				
T. amoenus luteiventris	0.07	0.068	0.068	0.068	0.067	0.068	0.068	-			
T. ruficaudus ruficaudus	0.063	0.061	0.059	0.061	0.059	0.061	0.034	0.067	-		
T. ruficaudus simulans	0.064	0.061	0.059	0.062	0.06	0.061	0.022	0.067	0.032	-	
T. striatus	0.155	0.155	0.155	0.156	0.155	0.155	0.154	0.156	0.156	0.154	-

Supplementary Material

This section describes the rationale behind ruling out sites as part of the selection analyses described in the text. It also includes supplementary figures and tables.

ATP8:

REL detects three codons under positive selection: codons 18, 35, and 47.

Codon substitutions:

Codon 18: Two of the shifts take place on a branch leading to a *canipes* and *amoenus canicaudus*. There is no amino acid substitution. *T. ruficaudus ruficaudus* has a shift from phenylalanine to leucine. *T. striatus* has a shift from leucine to alanine. This codon is explicitly negatively selected in other analyses. Chemical composition is conserved but is not significant ($p = 0.192$).

Codon 35: The shift takes place on a branch leading to *T. striatus*. No amino acid substitution. Chemical composition changes but is not significant ($p=0.850$).

Codon 47: The shift takes place on branches leading to a single *T. dorsalis* and *T. striatus*. The *dorsalis* has a shift from histidine to asparagine. The striatus has a shift from histidine to tyrosine. Chemical composition and polarity changes are present but are not significant ($p = 0.747$ and $p = 0.610$).

Codon 18 encodes an amino acid in a membrane-spanning domain. The others are outside the membrane.

Conclusion: These are false-positives.

COII:

REL and FUBAR detect a single codon under positive selection: 129. Two internal nodes, *T. striatus*, and two *T. dorsalis* experience a shift from serine to asparagine.

There is conservation in the chemical composition of this codon, though it is not significant ($p = 0.190$).

This codon encodes an amino acid in a topological domain in the mitochondrial intermembrane.

Conclusion: Two out of five analyses support this change in the outgroup and in two *T. dorsalis*. Possible site under positive selection in *T. dorsalis*.

CytB:

FEL and FUBAR detect a single codon under positive selection: 238. An internal node, *T. amoenus luteiventris,* and *T. striatus* experience an amino acid shift from valine to alanine, threonine, and leucine, respectively.

This codon has a shift in hydropathy, though it is not significant ($p = 0.325$).

This amino acid is in a transmembrane domain.

Conclusion: Since this is divergent in the outgroup and also in a transmembrane domain, chances are the substitution has little effect on functional protein structure especially since the amino acid side chains are of approximately equal size. The shift in *amoenus canacaudis* to threonine does replace a methyl group with a hydroxyl group, but this is the only change. Unlikely to be under positive selection for these reasons.

ND6:

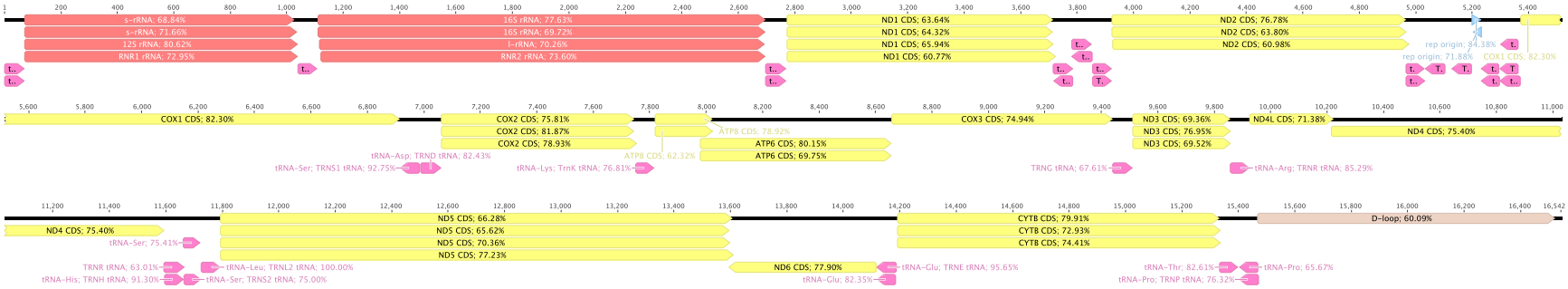
REL detects a single codon under positive selection: 171.

There is no polymorphism at this codon. This is a false positive.

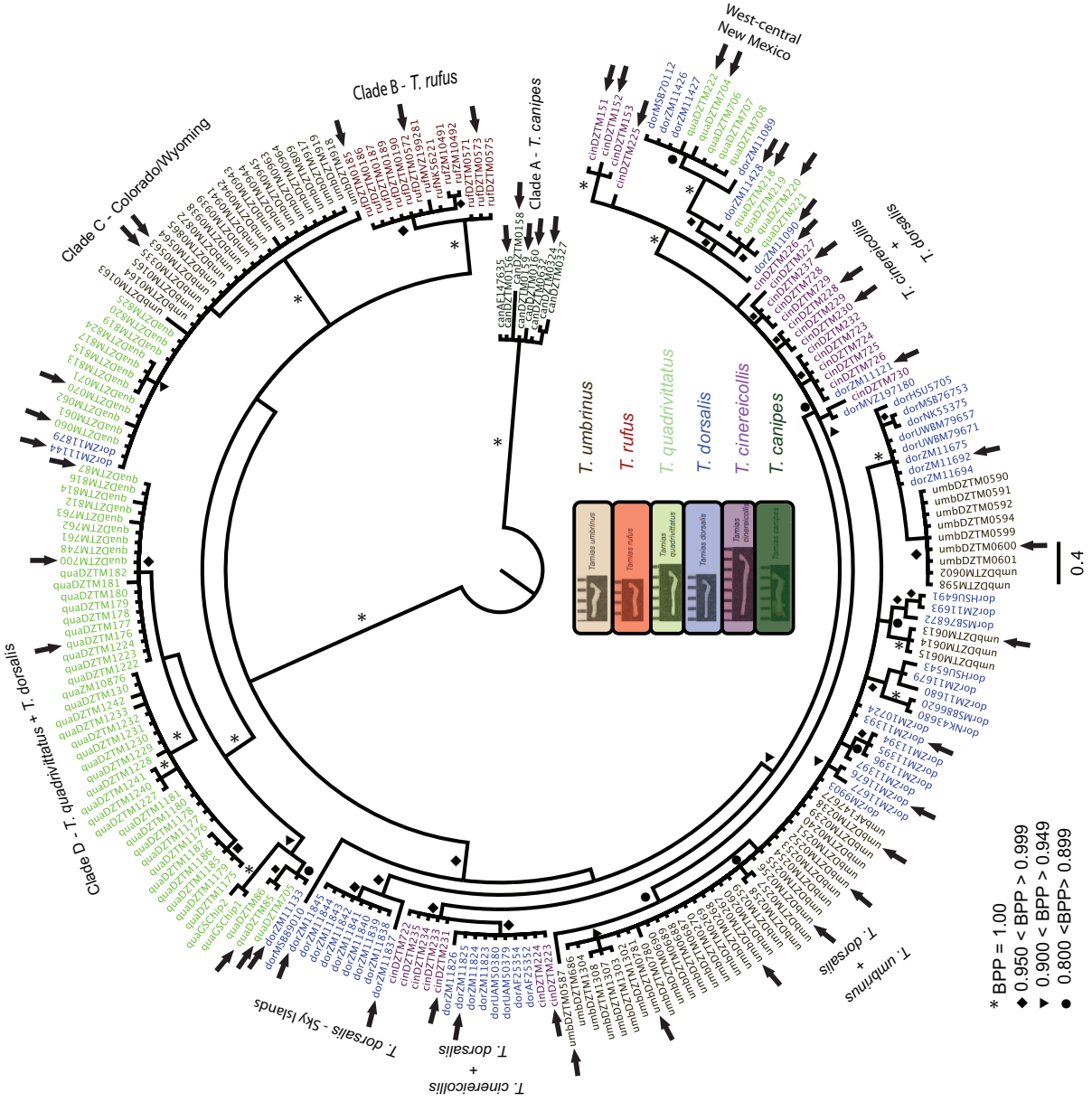
Remarks

Since the NJ trees constructed by DataMonkey are unrooted, and we have prior knowledge that confirms *T. striatus* is the appropriate outgroup, many of these represent substitutions in *striatus* relative to the *quadrivittatus* group. Possible positive selection at COII within *dorsalis*, but only marginal evidence at this point. More sampling and sequencing would be needed to confirm that this is a truly adaptive substitution and not a sampling-based artifact. No more than two of five methods ever agree that there is selection at a codon, and SLAC, the most conservative, never detects any.

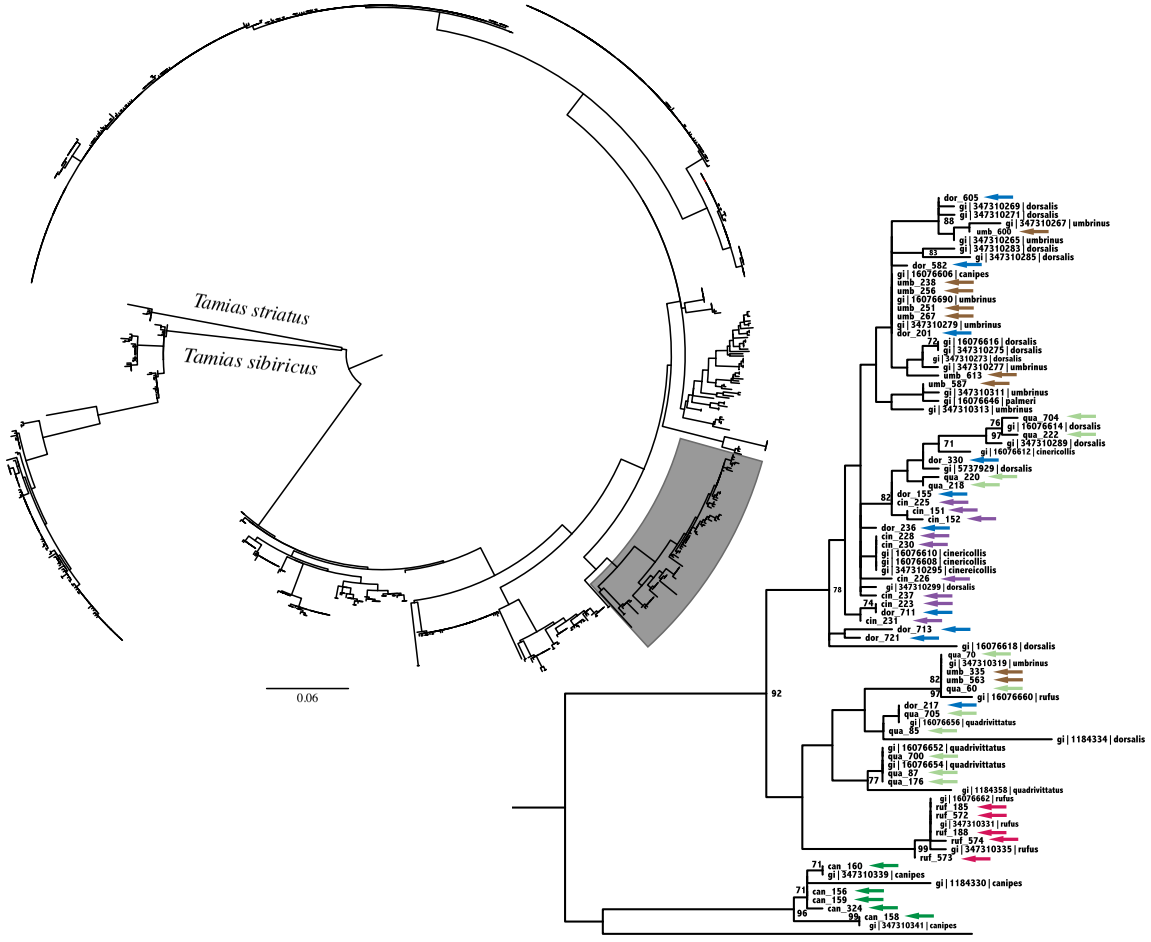
Supplementary Figure 3.1: Annotated *Tamias* draft mitochondrial genome reference. Genes were annotated using Geneious Pro v6.1.7 using *Mus musculus* and *Sciurus vulgaris* references.



Supplementary Figure 3.2: Mitochondrial genomes sequenced in this study with respect to cytochrome b sequences generated for Sullivan et al. (2014). Arrows indicate individuals whose mitochondrial genomes were sequenced as part of this study.


Clade A - T. canipes
Clade B - T. rufus
Clade C - Colorado/Wyoming
Clade D - T. quadrivittatus + T. dorsalis
West-central New Mexico
T. umbrinus
T. rufus
T. quadrivittatus
T. dorsalis
T. cinereicollis
T. canipes
T. cinereicollis + T. dorsalis
T. dorsalis - Sky Islands
T. dorsalis + T. umbrinus
0.4
* BPP = 1.00
◆ 0.950 < BPP > 0.999
▼ 0.900 < BPP > 0.949
● 0.800 <BPP> 0.899

Supplementary Figure 3.3: Cytochrome *b* tree estimated from 989 *Tamias* individuals downloaded using the PhyLoTA browser. Individuals included as part of this study are identified in the insert tree with arrows. The tree is midpoint rooted.



Supplementary Table 3.1**:** Posterior probabilities of models fit using codeml. Posterior probabilities are estimated per protein-coding gene. *P*-values of likelihood ratio tests for nested models are listed below each gene.

Model	ATP6	ATP8	COI	COII	COIII	CytB	ND1	ND2	ND3	ND4H	ND4L	ND5	ND6
M0	0.00488	0.00001	0.44924	0.00000	0.00000	0.00000	0.20482	0.00000	0.00000	0.00000	0.00022	0.00000	0.00000
M1	0.07112	0.25418	0.18572	0.91773	0.00270	0.00002	0.01148	0.00062	0.15037	0.00000	0.07642	0.00000	0.00003
M2	0.00031	0.00379	0.00036	0.01001	0.00001	0.00000	0.00004	0.00000	0.00131	0.00000	0.00078	0.00000	0.00000
M3	0.00013	0.00124	0.00003	0.00067	0.00064	0.00022	0.00007	0.00034	0.00031	0.00006	0.00050	0.00005	0.00056
M4	0.00072	0.00892	0.00010	0.00202	0.00912	0.00058	0.00000	0.00000	0.01747	0.00000	0.00439	0.00000	0.00169
M5	0.44471	0.46990	0.17709	0.00007	0.47560	0.48559	0.38012	0.50298	0.37913	0.48897	0.43317	0.49321	0.42331
M6	0.00197	0.00765	0.00035	0.00000	0.00218	0.00132	0.00124	0.00269	0.00021	0.00114	0.00443	0.00087	0.00712
M7	0.44462	0.21654	0.17716	0.00007	0.47539	0.48565	0.37979	0.40890	0.37913	0.48607	0.43206	0.48493	0.52391
M8	0.00197	0.00884	0.00042	0.00970	0.00218	0.00132	0.00119	0.00440	0.01805	0.00106	0.00441	0.00084	0.00310
M8a	0.02957	0.02892	0.00953	0.05973	0.03217	0.02531	0.02125	0.08006	0.05402	0.02269	0.04364	0.02011	0.04029
M1 vs. M2	1.00000	1.00000	1.00000	0.41066	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
M7 vs. M8	0.99005	0.36422	0.81873	0.00003	1.00000	1.00000	0.99005	0.28083	0.18268	1.00000	1.00000	1.00000	1.00000
M8 vs M8a	0.50000	0.08748	0.50000	0.09197	0.50000	0.50000	0.50000	0.50000	0.05480	0.50000	0.50000	0.50000	0.50000

Supplementary Table 3.2: Sampling information for the central and southern Rocky Mountains chipmunks sequenced as part of this study.

Species	Sample	Identifier	Latitude	Longitude	Sex	Genbank ID
canipes	DZTM.156	ZM.11091	33.4469	-105.782	female	JN042434
canipes	DZTM.158	ZM.11093	32.1097	-104.7475	male	JN042436
canipes	DZTM.160	ZM.11095	34.2207	-105.796	male	JN042435
canipes	DZTM.324	ZM.11424	33.3946	-105.7246	male	KJ139460
canipes	DZTM.159	ZM.11094	34.2207	-105.796	female	KJ139457
rufus	DZTM.185	ZM.11203	39.8629	-106.642	female	JN042432
rufus	DZTM.188	ZM.11206	39.8629	-106.642	male	NA
rufus	DZTM.574	ZM.11807	39.8189	-108.715	male	NA
rufus	DZTM.572	ZM.11805	39.8189	-108.715	male	KJ139470
rufus	DZTM.573	ZM.11806	39.8189	-108.715	male	KJ139464
dorsalis	DZTM.713	ZM.11837	31.8782	-109.2229	female	KJ139550
dorsalis	DZTM.721	ZM.11845	32.6661	-109.8747	male	KJ139558
dorsalis	DZTM.711	ZM.11825	32.9773	-108.2181	male	KJ139568
dorsalis	DZTM.582	ZM.11676	39.224	-114.5662	female	KJ139582
dorsalis	DZTM.605	ZM.11692	39.1784	-114.2833	male	KJ139571
dorsalis	DZTM.330	ZM.11428	34.01	-107.1995	male	NA
dorsalis	DZTM.155	ZM.11090	33.9363	-107.515	female	KJ139561
dorsalis	DZTM.154	ZM.11089	33.9363	-107.515	female	JN042410
dorsalis	DZTM.201	ZM.11393	40.8295	-108.7349	female	JN042396
dorsalis	DZTM.217	ZM.11133	35.25	-107.6758	female	JN042427
dorsalis	DZTM.236	ZM.11121	34.3838	-111.265	female	JN042416
dorsalis	DZTM.250	ZM.11144	41.2851	-109.335	male	JN042426
umbrinus	DZTM.563	ZM.11625	39.0995	-106.1548	male	KJ139596
umbrinus	DZTM.238	ZM.11379	36.699	-112.2752	male	JN042406
umbrinus	DZTM.251	ZM.11160	40.607	-110.9933	male	JN042394
umbrinus	DZTM.256	ZM.11165	44.2976	-109.2544	female	JN042395
umbrinus	DZTM.267	ZM.11147	44.2976	-109.2544	female	JN042397
umbrinus	DZTM.587	ZM.11681	38.9219	-116.8647	female	KJ139613
umbrinus	DZTM.613	ZM.11700	39.1511	-111.557	male	KJ139614
umbrinus	DZTM.600	ZM.11687	40.1755	-114.8556	male	KJ139611
umbrinus	DZTM.781	ZM.11881	42.53891	-108.79631	male	KJ139624
umbrinus	DZTM.335	ZM.11433	40.0205	-105.5142	male	KJ139595
quadrivittatus	DZTM.60	ZM.11031	39.7603	-105.3296	male	JN042423
quadrivittatus	DZTM.70	ZM.11024	39.7362	-105.248	female	JN042424
quadrivittatus	DZTM.85	ZM.11078	36.6776	-106.014	male	JN042428
quadrivittatus	DZTM.87	ZM.11085	36.4443	-106.007	female	JN042429
quadrivittatus	DZTM.176	ZM.11096	38.0222	-105.6799	female	JN042430
quadrivittatus	DZTM.218	ZM.11134	35.2092	-107.6274	female	JN042409
quadrivittatus	DZTM.704	ZM.11818	35.2165	-108.1309	male	KJ139475
quadrivittatus	DZTM.222	ZM.11138	35.2092	-107.6274	female	KJ139474
quadrivittatus	DZTM.705	ZM.11819	35.2165	-108.1309	male	KJ139492
quadrivittatus	DZTM.700	ZM.11814	36.7981	-105.0714	male	KJ139496
quadrivittatus	DZTM.220	ZM.11136	35.2092	-107.6274	female	KJ139471
cinereicollis	DZTM.151	ZM.11086	33.8995	-107.5107	female	JN042411
cinereicollis	DZTM.152	ZM.11087	33.8441	-107.561	male	JN042412
cinereicollis	DZTM.223	ZM.11110	33.7269	-108.9771	female	JN042419
cinereicollis	DZTM.226	ZM.11111	33.671	-109.3495	female	JN042418
cinereicollis	DZTM.228	ZM.11113	33.671	-109.3495	male	KJ139539
cinereicollis	DZTM.230	ZM.11115	34.1106	-109.5931	male	JN042414
cinereicollis	DZTM.231	ZM.11116	34.375	-110.9796	female	JN042420
cinereicollis	DZTM.225	ZM.11108	33.7269	-108.9771	male	KJ139547
cinereicollis	DZTM.237	ZM.11378	34.9337	-111.356	female	JN042417

Supplementary Table 3.3: Multilocus McDonald-Kreitman tests. For each comparison, a Mantel-Haenszel test of Homogeneity was performed to assess rate homogeneity among loci. When applicable, the p-value for the multilocus McDonald-Kreitman test is listed for each comparison. N/A refers to the inability to calculate a test statistic due to a lack of nonsynonymous substitutions.

Species 1	Species 2	Chi-Square (Mantel-Haenszel Test of Homogeneity)	p-value	Mantel-Haenszel Estimator Omega	Chi-Square	p-value	alpha
quadrivittatus	*cinereicollis*	1.853	0.999	N/A	N/A	N/A	N/A
quadrivittatus	*rufus*	3.853	0.985	1.425	0.593	0.441	-0.684
quadrivittatus	*canipes*	11.579	0.48	1.525	2.948	0.085	-0.703
quadrivittatus	*dorsalis*	2.635	0.997	N/A	N/A	N/A	N/A
quadrivittatus	*umbrinus*	3.917	0.984	N/A	0.6942	0.405	N/A
cinereicollis	*rufus*	4.959	0.959	1.816	3.243	0.071	-0.537
cinereicollis	*canipes*	8.983	0.704	1.756	5.076	**0.024**	-0.438
cinereicollis	*dorsalis*	1.674	0.999	N/A	N/A	N/A	N/A
cinereicollis	*umbrinus*	7.025	0.855	1.61	0.114	0.735	-0.365
rufus	*canipes*	9.183	0.687	2.074	5.818	**0.015**	-0.366
rufus	*dorsalis*	4.85	0.962	1.074	0	0.991	-0.43
rufus	*umbrinus*	8.125	0.775	1.277	0.281	0.591	-0.15
canipes	*dorsalis*	8.395	0.753	1.262	0.845	0.357	-0.466
canipes	*umbrinus*	19.463	0.077	1.52	2.878	0.089	-0.299
dorsalis	*umbrinus*	2.31	0.998	N/A	N/A	N/A	N/A

Supplementary Table 3.4: Tests for deviations from neutrality. Tajima's D, Fu and Li's D_2*, and Fu and Li's F* were estimated for each locus and all loci combined per species. A: *T. canipes*; B: *T. cinereicollis*; C: *T. dorsalis*; D: *T. quadrivittatus*; E: *T. rufus*; F: *T. umbrinus*.

A:

Locus	Tajima's D	p-value	Fu and Li's D2*	p-value	Fu and Li's F*	p-value
ATP6	0	0	0	0	0	0
ATP8	0	0	0	0	0	0
COI	-0.97762	0.11046	-0.97762	0.11046	-0.97762	0.11046
COII	-1.123974	0	-1.123974	0	-1.123974	0
COIII	-1.184406	0	-1.184406	0	-1.184406	0
CytB	-1.214724	0	-1.214724	0	-1.214724	0
ND1	-0.197401	0.45083	-0.197401	0.45083	-0.197401	0.45083
ND2	-0.596333	0.29572	-0.596333	0.29572	-0.596333	0.29572
ND3	-0.174749	0.36047	-0.174749	0.36047	-0.174749	0.36047
ND4H	-1.174318	0	-1.174318	0	-1.174318	0
ND4L	-0.972558	0	-0.972558	0	-0.972558	0
ND5	-1.078084	0.06615	-1.078084	0.06615	-1.078084	0.06615
ND6	-1.145536	0	-1.145536	0	-1.145536	0
All	-1.010666	0.12632	-1.010666	0.12632	-1.010666	0.12632

B:

Locus	Tajima's D	p-value	Fu and Li's D2*	p-value	Fu and Li's F*	p-value
ATP6	0.150569	0.60121	-0.009942	0.42523	0.029276	0.49705
ATP8	0.986274	0.66681	0.840403	0.4351	0.888288	0.66681
COI	-0.215582	0.44199	-0.417306	0.31208	-0.375904	0.3482
COII	0.157798	0.59741	-0.04974	0.38635	0	0.49567
COIII	0.532164	0.7416	0.472403	0.6342	0.49637	0.68094
CytB	-0.480357	0.33988	-0.347678	0.38689	-0.380836	0.36889
ND1	0.783727	0.82024	0.508173	0.62546	0.585629	0.72418
ND2	0.152053	0.60128	-0.079632	0.41814	-0.024074	0.46864
ND3	-0.075406	0.48149	-0.264179	0.28436	-0.223159	0.43011
ND4H	0.042067	0.55708	0.081307	0.49555	0.073256	0.50883
ND4L	0.195897	0.62361	-0.221036	0.35902	-0.124671	0.52608
ND5	0.258639	0.63755	-0.001495	0.47027	0.060245	0.50634
ND6	-0.328122	0.39761	-0.432274	0.31899	-0.415192	0.33905
All	0.121829	0.59412	-0.030127	0.45367	0.006929	0.47716

C:

Locus	Tajima's D	p-value	Fu and Li's D2*	p-value	Fu and Li's F*	p-value
ATP6	-0.970556	0.16477	-0.875871	0.19556	-0.928407	0.18781
ATP8	-0.912217	0.18571	-0.81667	0.2201	-0.865872	0.19912
COI	-1.375866	0.08034	-1.235604	0.1321	-1.311889	0.12025
COII	-1.043234	0.14914	-0.966527	0.18234	-1.016944	0.17137
COIII	-1.137053	0.12738	-1.075416	0.15535	-1.125469	0.15162
CytB	-0.992033	0.15918	-0.816635	0.21342	-0.888959	0.19433
ND1	-1.019811	0.15266	-0.836391	0.20204	-0.911429	0.19073
ND2	-1.302934	0.09326	-1.256125	0.12814	-1.30829	0.12164
ND3	-1.437091	0.06868	-1.48579	0.08617	-1.519115	0.0874
ND4H	-1.14665	0.12468	-1.074599	0.15866	-1.127773	0.15191
ND4L	-1.596787	0.04376	-1.64779	0.05668	-1.684379	0.05895
ND5	-1.13602	0.13486	-1.169989	0.14807	-1.192028	0.14586
ND6	-0.930486	0.17947	-1.048218	0.16782	-1.045241	0.17008
All	-1.196217	0.11649	-1.105179	0.1557	-1.1647	0.14611

D:

Locus	Tajima's D	p-value	Fu and Li's D2*	p-value	Fu and Li's F*	p-value
ATP6	0.957	0.871	1.382	0.983	1.309	0.973
ATP8	0.453846	0.71523	0.623123	0.65387	0.59544	0.72025
COI	0.792407	0.8339	1.307816	0.97907	1.212334	0.96267
COII	1.338129	0.9414	1.453101	0.99163	1.467701	0.98814
COIII	1.133779	0.9087	1.591583	0.99729	1.519966	0.99114
CytB	0.965507	0.87805	1.36197	0.98696	1.299762	0.97636
ND1	1.015034	0.88766	1.275858	0.97109	1.246753	0.96645
ND2	0.912372	0.86482	1.520159	0.99859	1.406875	0.98694
ND3	0.616865	0.77529	1.135199	0.92432	1.03351	0.91351
ND4H	0.871448	0.85839	1.366145	0.98883	1.278011	0.97391
ND4L	0.607578	0.76509	1.489884	0.97114	1.303574	0.95757
ND5	0.848197	0.84634	1.444637	0.99575	1.32924	0.98044
ND6	1.188288	0.9195	1.252471	0.96241	1.27445	0.96879
All	0.971042	0.88119	1.418727	0.99702	1.344881	0.9849

E:

Locus	Tajima's D	p-value	Fu and Li's D2*	p-value	Fu and Li's F*	p-value
ATP6	1.224745	0.61731	1.224745	0.61731	1.224745	0.61731
ATP8	0	0	0	0	0	0
COI	-0.816497	0	-0.816497	0	-0.816497	0
COII	1.224745	0.61468	1.224745	0.61468	1.224745	0.61468
COIII	0.243139	0.45021	0.243139	0.45021	0.243139	0.45021
CytB	-0.972558	0	-0.972558	0	-0.972558	0
ND1	1.224745	0	1.224745	0	1.224745	0
ND2	0.91278	0.74944	0.91278	0.74944	0.91278	0.74944
ND3	0.698995	0.63421	0.698995	0.63421	0.698995	0.63421
ND4H	1.718304	0.9351	1.718304	0.9351	1.718304	0.9351
ND4L	1.224745	0.61809	1.224745	0.61809	1.224745	0.61809
ND5	0.660554	0.68986	0.660554	0.68986	0.660554	0.68986
ND6	0	0	0	0	0	0
All	0.927095	0.79921	0.927095	0.79921	0.927095	0.79921

F:

Locus	Tajima's D	p-value	Fu and Li's D2*	p-value	Fu and Li's F*	p-value
ATP6	-0.498858	0.32248	0.453536	0.64884	0.221764	0.57223
ATP8	-0.203435	0.44741	0.376717	0.60894	0.239549	0.56491
COI	-0.078593	0.50893	0.945936	0.88387	0.707949	0.79502
COII	-0.282624	0.41472	0.719558	0.74234	0.482224	0.68676
COIII	0.027752	0.55406	1.049676	0.88444	0.815441	0.83225
CytB	-0.058728	0.48854	0.61185	0.7223	0.461507	0.67267
ND1	-0.364479	0.37829	0.424163	0.63968	0.233101	0.57948
ND2	0.050346	0.56816	0.977908	0.90207	0.765358	0.82398
ND3	-0.295347	0.40674	0.775302	0.79037	0.521529	0.7114
ND4H	-0.242612	0.43178	0.490749	0.67366	0.315399	0.61489
ND4L	0.074092	0.54315	1.16252	0.91421	0.918013	0.84738
ND5	-0.192677	0.4549	0.830283	0.83509	0.589258	0.74719
ND6	-0.320602	0.39483	0.572779	0.70145	0.359177	0.63294
All	-0.203233	0.44893	0.762273	0.81969	0.533877	0.72344

Supplementary Table 3.5: Additional tests for selection. For each of the six species in the *quadrivittatus* group (*T. canipes*, *cinereicollis, dorsalis*, *quadrivittatus*, *rufus*, and *umbrinus*), we performed a series of analyses (SLAC, FEL, IFEL, REL, and FUBAR) for each protein-coding gene. + refers to the number of sites identified as being under positive selection. – refers to the number of sites identified as being under negative selection.

Locus	SLAC +	SLAC -	FEL +	FEL -	IFEL +	IFEL -	REL +	REL -	FUBAR Diversifying	FUBAR Purifying
ATP6	0	16	0	68	0	20	0	0	0	121
ATP8	0	2	0	9	0	1	3	9	0	8
COI	0	44	0	171	0	61	0	0	0	444
COII	0	13	0	75	0	28	1	132	1	138
COIII	0	19	0	83	0	30	0	0	0	176
CytB	0	49	1	166	0	73	0	269	1	308
ND1	0	26	0	129	0	56	0	0	0	260
ND2	0	38	0	122	0	54	0	216	0	252
ND3	0	6	0	38	0	14	0	72	0	57
ND4H	0	50	0	153	0	76	0	294	0	349
ND4L	0	8	0	30	0	6	0	58	0	38
ND5	0	57	0	203	0	87	0	386	0	442
ND6	0	18	0	39	0	22	1	97	0	60
Total	0	346	1	1286	0	528	5	1533	2	2653

Chapter 4

Phylogenomic and Population Genomic Characterization of Patterns of Diversification in Central and Southern Rocky Mountains Chipmunks (Sciuridae: *Tamias*)

Abstract

Evidence from natural systems suggests that hybridization between species is more common than traditionally thought. Studies have historically relied on a handful of loci, often mitochondrially-linked, to describe patterns of introgression. This study describes a series of genomic analyses in chipmunks, a system with a documented pattern of widespread mitochondrial introgression among species. Here, we use a targeted exon capture approach to sequence thousands of nuclear loci from chipmunks in the central and southern Rocky Mountains belonging to the *T. quadrivittatus* group. In contrast to a number of studies focused on describing the extent of mitochondrial introgression in this system, relatively little nuclear-genomic analysis has been performed. We used a series of phylogenomic analyses to resolve the systematic relationships among the six species in this group. Additionally, we performed several population genomic analyses to characterize nuclear genomes and infer coancestry among individuals. We found that, even though mitochondrial introgression is rampant among some species pairs, there appears to be little evidence of nuclear introgression. These results suggest that other forces, such as sexual selection, play an important role in preventing nuclear genomic admixture in chipmunks.

Introduction

The characterization of genetic changes underlying speciation is one of the most central topics in evolutionary biology (e.g., Coyne and Orr 2004; Butlin and Ritchie 2009; Nosil and Schluter 2011; Butlin et al. 2012; Seehausen et al. 2014). Speciation is often understood as a continuous process governed by the accumulation of genetic differences between incipient lineages that ultimately reduce or eliminate reproductive success between diverging lineages (Nosil et al. 2009; Strasburg et al. 2012; Seehausen et al. 2014). Consequently, ongoing gene flow can occur throughout the speciation process and may even occur between lineages recognized as distinct species; closely related species may not be completely reproductively isolated and produce viable interspecific hybrids (e.g., Coyne and Orr 2004; Mallet 2005; Rieseberg 2009). Furthermore, hybridization may result in movement of genetic regions across species boundaries via introgression (Anderson 1949), resulting in heterospecific genomes and generating genetic diversity through the novel combination of genotypes or elevated allelic diversity (e.g., Rieseberg et al. 1996; Castric et al. 2008; Kim et al. 2008; Twyford and Ennos 2012).

The notion that the genomes of well-described species are semi-permeable has been discussed for some time (reviewed in Harrison and Larson (2014)). Conceptual characterizations have been formulated that attempt to explain genomic patterns of divergence in the face of ongoing gene flow (Wu 2001; Pinho and Hey 2010; Smadja and Butlin 2011; Nosil and Feder 2012). These models, generally termed “speciation-with-gene flow” models, generate predictions that can be tested using genome-scale data from natural populations.

Recent evidence suggests that the radiation of western North American chipmunks (Sciuridae: *Tamias*) may be described by speciation-with-gene flow models. The genus *Tamias* is distributed throughout Asia (1 species, *Tamias sibiricus*, subgenus *Eutamias*), eastern North America (1 species, *Tamias striatus*, subgenus *Tamias*), and western North America (23 species, subgenus *Neotamias*). Species can be identified by their genital morphology, with the baculum (*os penis*) acting as a diagnostic character.

Sullivan et al. (2014) summarized work from the past 12+ years in this system and, in particular, outlined several well-studied cases of introgression. Extensive mtDNA introgression within two *Neotamias* subgroups is described in a series of studies (Good et al. 2003, 2008; Hird et al. 2010; Reid et al. 2010, 2012; Sullivan et al. 2014). Initially, asymmetric introgression from *T. ruficaudus* into *T. amoenus* was described in the northern Rocky Mountains (Good et al. 2003, 2008; Hird et al. 2010; Reid et al. 2010). Subsequent work documented the extent of introgression in the central and southern Rocky Mountains, specifically six species in the *T. quadrivittatus* group (Sullivan et al. 2014; Sarver et al. in prep.). Among these species, mitochondrial introgression is rampant among four of the six, and one species (*T. dorsalis*) appears to inherit the mitochondrial genome of species it is co-distributed with.

Even though mitochondrial introgression is pervasive and suggests widespread hybridization, relatively less work has focused on the extent of nuclear introgression in chipmunks. Hird and Sullivan (2009) characterized a contact zone between *T. ruficaudus ruficaudus* and *T. ruficaudus simulans* and found that bacular morphotypes and nuclear loci showed stark delimitation at the Lochsa River; however they identified substantial gene flow across the Bitterroot divide along an axis perpendicular to the bacular contact zone.

Mitochondrial data from the same individuals indicated introgression that attenuated with increasing distance from the zone of contact. Furthermore, Reid et al. (2012) developed a series of nuclear markers associated with reproductive protein genes in order to attempt to resolve chipmunk systematics. They used multiple phylogenetic approaches, including concatenation and several methods of species-tree estimation, but failed to resolve all nodes with high statistical support. There are, therefore, several open questions in this system that can be addressed with genomic-scale data.

Here, we characterize the extent of nuclear introgression in one group from the central and southern Rocky Mountains. Specifically, we use a targeted exon capture approach to sequence thousands of loci from six species in the *T. quadrivittatus* group. We first use population genetic approaches to document the extent of nuclear introgression, which has not been assessed in these taxa. We then use phylogenomic approaches to resolve systematic relationships among these six species using a series of techniques for estimating species trees from multilocus data.

Materials and Methods

Sample preparation and sequencing

Fifty-one *T. quadrivittatus*-group chipmunks (five *T. canipes*, nine *T. cinereicollis*, 11 *T. dorsalis*, 11 *T. quadrivittatus*, five *T. rufus*, and ten *T. umbrinus*) were selected from the 231 individuals used by Sullivan et al. (2014) to characterize mtDNA introgression. For each species, we included individuals containing introgressed and non-introgressed mtDNA. We also included data from three *T. striatus* individuals published by Bi et al. (2012). DNA was isolated from heart or liver tissue, and extractions were performed using Qiagen DNEasy DNA extraction kits. Samples were eluted into 50 *u*L of 10 mM Tris-Cl and stored

at -20°C before use. We then performed exon capture using Agilent SureSelect microarrays following Bi et al. (2012). Samples were sequenced on two lanes of an Illumina HiSeq 2000 at the Vincent Coates Genome Sequencing Laboratory at the University of California – Berkeley. Sequences were processed using a comprehensive cleaning pipeline. First, duplicate sequences were removed. Remaining sequences were screened for quality and residual adapters using SeqyClean (Zhbannikov et al. in prep*;* http://bitbucket.org/izhbannikov/seqyclean). Finally, overlapping reads were consolidated using Flash (Magoč and Salzberg 2011). Population genetic analyses used the same pipeline without Flash overlapping

Reads were assembled into contigs using Assembly by Reduced Complexity (ARC; Hunter et al. in prep; https://github.com/ibest/ARC), an approach that uses reference sequences as a starting point for assembly; reads were mapped to the sequences, and each pool of reads corresponding to a single sequence was assembled in a *de novo* fashion. The assembled contigs were used as a new reference for another round of mapping and assembly, and this process was repeated until one of a series of termination conditions was achieved. ARC has the ability to recruit additional reads throughout iterations to extend target sequence length into flanking regions. Since no genomic reference is available for *Tamias*, we used ARC to produce sets of contigs with targets from which capture probes were designed used as reference sequences (see Bi et al. 2012). Because targets were originally designed using a pooled-tissue transcriptome, we removed any targets that included redundant exons (resulting from isotigs) to avoid inclusion of multiple targets with the same sequence. Finally, in order to capture heterozygous sites, sequencing reads were mapped to each ARC contig using GATK (McKenna et al. 2010; DePristo et al. 2011).

High-quality heterozygous sites, as identified through high mapping quality and depth of coverage, were translated into their corresponding IUPAC ambiguity codes and injected back into the ARC contig.

Phylogenetic analyses

The final contig set produced by ARC for each individual was processed using R v3.0.2 (R Core Team 2013) and several other applications. First, all results were trimmed to include only targets where ARC produced a single contig across all libraries. Then, a multiple sequence alignment was performed on each set of sequences using MUSCLE v3.8.31 (Edgar 2004). The resulting matrices were then squared by trimming hanging ends.

We inferred phylogenies with this dataset using several approaches. First, all contigs were concatenated using Phyutility (Smith and Dunn 2008). A phylogeny was inferred using RAxML v8.0.5 (Stamatakis 2014) using a full ML search across 1000 bootstrap replicates under a GTR + G model of nucleotide sequence evolution. Modern assemblers do not incorporate heterozygotes into contigs; instead, assemblers either use the first base encountered as the reference call or use a majority-rule call, removing heterozygous sites from the final assembly. Therefore, in order to assess the impact of heterozygous sites on phylgenomics, we also constructed a dataset consisting of ARC contigs without heterozygous sites (i.e., contigs resulting from assembly with no subsequent modification). This dataset was subject to the same processing treatment as the dataset above, and the same phylogenetic inference was performed using RAxML v8.0.5 (Stamatakis 2014).

We then conducted a series of species-tree inferences using two datasets. First, we approximated the chromosomal location of each ARC contig relative to the GRCm38 reference assembly of *Mus musculus* using BLAT (Kent 2002); despite deep divergence

between murids and sciruids, chromosome painting studies (Li et al. 2004) have demonstrated remarkable conservatism in karyotypes and patterns of synteny. In order for assignment to a (pseudo)chromosome, all individuals in the data set must have been assigned to the same chromosome unambiguously. Contigs were then concatenated based on their chromosome assignment. Model selection was performed on each contig set using DT-ModSel (Minin et al. 2003). Phylogenetic trees were estimated using Garli v2.01 (Zwickl 2006) under the selected model and a termination threshold of 0.01 for 50,000 generations. Ten independent search replicates were performed. The best tree among replicates was used for subsequent analysis. Each tree was made ultrametric using treePL (Smith and O'Meara 2012) and the three *T. striatus* samples as an outgroup with the minimum split time set to 7 MYA. We view this approach as a moderately-informed binning procedure.

We also implemented a naïve binning approach (Bayzid and Warnow 2013). All contigs were randomly assigned without replacement into 10 bins of equal size. Combination of loci, model selection, phylogenetic inference, and ultrametric transformation were performed as described for the chromosome case. Twenty-five random binnings and phylogenetic estimations were performed.

Finally, species trees were inferred for trees inferred under the moderately-informed binning procedure and the naïve binning procedure. We used three separate approaches: MP-EST (Liu et al. 2010), STAR (Liu et al. 2009), and STEAC (Liu et al. 2009). MP-EST estimates the species tree under the coalescent using a pseudo-likelihood approach. In contrast, STAR uses ranks of coalescent times, and STEAC uses average coalescent times in order to produce an estimate of the species tree. STAR and STEAC trees were estimated under both neighbor-joining (NJ) and unweighted pair-group method with arithmetic mean

(UPGMA) clustering approaches. Comparisons between naïve binning replicates and the concatenated RAxML topology were made using Robinson-Foulds distances (Robinson and Foulds 1981) as implemented in the R package ape (Paradis et al. 2004).

Population genomic analyses

For population genomic analyses, one individual (S600; *T. umbrinus*) was arbitrarily selected to serve as a genomic reference. The ARC assembly for this individual was pruned to include targets that produced three or fewer contigs in order to account for mis-assembly. Variants were then called for each individual using this reference set. Post-processed reads were aligned to the ARC assembly using Bowtie 2 v2.1.0 (Langmead and Salzberg 2012). Alignments were improved using GATK (McKenna et al. 2010; DePristo et al. 2011), and variants were phased and missing genotypes were imputed with BEAGLE v4.0 (Browning and Browning 2007, 2009; Browning and Yu 2009). Finally, using VCFtools v0.1.12a (Danecek et al. 2011), variants were screened to remove sites that violated Hardy-Weinberg Equilibrium (HWE) per species group at a *p*-value of 0.05. Remaining variants were filtered using a minor allele frequency cutoff of 0.05. *T. striatus* was removed for these analyses.

Individual coancestry was estimated using ADMIXTURE v1.23 (Alexander et al. 2009) with 10 rounds of cross-validation and 10 values of K (number of genotypic clusters) after selecting a single variant from each ARC contig (due to a lack of information about linkage). The same single-variant dataset was used for a multidimensional scaling analysis visualized in two dimensions to provide an overall assessment of population genomic structure; this analysis was performed using PLINK v.1.07 (Purcell et al. 2007). Additionally, Weir and Cockerham's F_{ST} (Weir and Cockerham 1984) and average allele frequencies were estimated for all SNPs across all species pairs using the Genotype-

Phenotype Association Toolkit ++ (available at https://github.com/jewmanchue/vcflib). F_{ST} values less than zero were converted to zero. Number of heterozygotes, observed heterozygosity, and F_{IS} were also calculated fore each species. SNPs that lacked variability were removed before calculating statistics. For comparison, F_{ST} was also calculated using DnaSP (Librado and Rozas 2009) from a concatenated set of mitochondrial protein-coding genes assembled as part of another study (Chapter 3).

Results

Assembly and processing

The pruned ARC targets file included a total length of 4,003,445 bp of sequence data consisting of 7,627 genes and 11,976 exons or targeted loci. ARC performed 1,300,188 assemblies total across all targets and individuals; 5,640 assemblies were terminated due to the incorporation of a repetitive element (as diagnosed through the incorporation of a huge number of reads relative to the previous ARC iteration) or assembly timeout (indicating assemblies took longer than 20 minutes to complete), most likely due to the incorporation of difficult-to-resolve repetitive sequences.

For molecular phylogenetics, sequences were only included in the final analysis if they were present across all libraries. Furthermore, in order to avoid possible errors due to improper resolution of sequence order (i.e., two contigs from a single gene may or may not be called in a consistent order across assemblies due to the stochastic nature of the assembly process), only genes that produced a single contig were included in downstream analyses. This resulted in 1,106 sequences. After alignment, this set was truncated to remove sequences that were 100% identical among all libraries and assemblies or contained too many divergent sites (indicative of assembly errors). This resulted in 1,060 sequences with

between 0.33% and 15.2% variable sites (calculated by counting the number of non-identical columns in the alignment and dividing by the length of the alignment). The final alignment consists of 221,556 bp per individual with no missing data.

The S600 *T. umbrinus* reference was generated by selecting targets for which ARC produced three or fewer contigs per gene. Of the 7,627 capture genes, 6,827 (89.5%) met this criterion;10,088 of 11,976 (84.2%) loci were included. 4,326 (56.7%) genes were resolved as a single contig for this library. The cutoff of three or fewer contigs per gene is conservative but recovers approximately 85% of targeted sequences.

Phylogenetic inference

Concatenated sequence data analyzed using RAxML produced a tree that recovers *T. umbrinus* as sister to the rest of the *T. quadrivittatus* group (Figure 1). Strong support was recovered for the relationships among *T. umbrinus*, *T. dorsalis*, *T. quadrivittatus*, and *T. cinereicollis*. Moderate support (bipartition frequencies between 70% and 80%) was recovered for bipartitions including *T. rufus* and *T. canipes*. When heterozygous sites were not explicitly accounted for, the same topology was recovered. However, all bootstrap support values were greater than 94% (Figure 2). These results were surprising, especially considering that many phylogenomic methods do not assess whether sites are truly homozygous post-assembly.

Species-tree estimation was performed using loci putatively identified to *Mus* chromosomal locations. 802 of the 1060 ARC contigs were unambiguously assigned to a chromosome and were used for this analysis. The number of contigs assigned to each chromosome ranges from 26 to 66, with the exception of a single contig assigned to the X and zero assigned to the Y. Each of the 19 *Mus* autosomes was represented.

Several of the species-tree estimation approaches (MP-EST, STAR-UPGMA and STAR-NJ) trees recover the same relationships as the concatenated RAxML tree from gene trees generated from datatsets of genes assigned to chromosomes, (Fig. 3). The estimate from STEAC differs, albeit with short internal branches; the STEAC-UPGMA tree suggests that *T. canipes* and *T. rufus* form a clade, whereas the STEAC-NJ tree flips the relationship between *T. canipes* and *T. dorsalis*.

Furthermore, the 1060 ARC contigs were subject to 25 rounds of naïve binning (Table 1). For each type of analysis (MP-EST, STAR, STEAC), Robinson-Foulds (RF) distances were computed between the trees relative to the concatenated tree; no comparison had a RF-distance greater than 2 from the concatenated tree. Across all replicate analyses, trees were in agreement with the concatenated tree 84% of the time or greater, lending strong support for the relationships suggested by the concatenated tree. These conclusions differ from phylogenies estimated in previously published analyses (Reid et al. 2012; Sullivan et al. 2014). Bootstrap values and posterior probabilities resolving the placement of *T. canipes* and *T. rufus* in Reid et al. (2012) were low and their placement on the tree was swapped relative to this study, but other relationships among the *T. quadrivittatus* group species are consistent with our findings.

Population genomics

There were, on average, 4,667,279 genotyped sites per library. A total of 218,792 variant sites consisting of 214,149 SNVs and 4,643 indels were identified. A small fraction of positions were multi-allelic (6,277). Two filters were applied to the raw variants. The first removed sites where less than 75% of individuals were genotyped, and the second removed sites with a minor allele frequency greater less than 1%. After the filtering steps, 180,879

variant sites remained (176,554 SNVs and 4,325 indels). Filtering on HWE and minor allele frequency (a final filtering at 5%) resulted in a final count of 111,441 variants. Selecting one variant per contig resulted in a thinned dataset of 7,530 SNPs for population assignment analyses.

ADMIXTURE's cross-validation approach suggests the optimum number of populations is 7 (Supplementary Figure 1), though it is important to interpret other values of *K* in a biological context. ADMIXTURE coancestry plots (Fig. 4) revealed a progression of resolution across K values, with *T. umbrinus* being resolved as its own population first (*K*=2), then *T. dorsalis* (*K*=3), followed by *T. rufus/T. canipes* and *T. cinereicollis/T. quadrivittatus* (*K*=5) and a resolution of *T. cinereicollis* and *T. quadrivittatus* (*K*=6). At *K*=7, populations, each species was identified and substructure suggested within *T. dorsalis*. There was little indication of interspecific admixture.

Multidimensional scaling revealed very similar clustering of individuals into the six species (Fig. 5). *T. umbrinus* and *T. dorsalis* were separated from the other four species in multivariate space. *T. rufus* and *T. canipes* cluster cleanly but were separated by much less distance. *T. cinereicollis* and *T. quadrivittatus* showed clear clustering but nearly overlap.

Average observed heterozygosity ranged from 0.11% to 0.14% with *T. umbrinus* having the lowest. F_{IS} estimates range from 0.0374 to 0.0755 (Table 2). Pairwise F_{ST} values showed clear differences among species (Table 3). Any comparison that contained *T. umbrinus* had a higher F_{ST} relative to other comparisons (> 0.38). F_{ST} values including *T. canipes*, *T. dorsalis,* or *T. rufus* showed intermediate estimates. The *T. quadrivittatus* and *T. cinereicollis* estimate was the lowest (0.10). There was a stark contrast between values calculated from mitochondrial and nuclear data. In some comparisons, such as those

involving *T. rufus* and *T. canipes* that lack mitochondrial introgression, F_{ST} values were large and approach one in some cases. Other comparisons, especially those with *T. dorsalis*, have lower values with some approaching 0.05.

Discussion

Advances in sequencing now allow for genomic characterization in non-model systems (Seehausen et al. 2014). Indeed, natural systems and hybrid zones often provide insights into the evolutionary process that cannot be obtained in model systems (Hewitt 1988). Until recently, the lack of molecular biological and bioinformatics tools in non-model systems impeded analysis. While not as sophisticated as the tools developed for *Homo* or *Mus*, techniques currently exist that can be used to manipulate non-model data and drawing biological inferences.

The present study uses some of these techniques to investigate genomic patterns of divergence in chipmunks. Previous studies in this system have shown widespread mitochondrial introgression (Good et al. 2003, 2008; Hird and Sullivan 2009; Ried et al. 2012; Sarver et al. in prep.; Sullivan et al. 2014), whereas few studies (Hird and Sullivan 2009; Reid et al. 2012) focus on the nuclear introgression that may also result from hybridization. Our results show that, in this system, concatenating nuclear data produces an estimate of the species tree that is generally consistent with other methods. Since concatenation assumes a single evolutionary process governs all loci, we employ species tree estimation methods in order to account for phylogenetic discordance among loci (Edwards 2009). Trees estimated from loci assigned to *Mus* chromosomes are consistent across methods (with the exception of STEAC) with the concatenated tree. We attempt to increase confidence in our estimate by randomly binning subsets of loci and inferring

species trees. A majority of methods agree with the concatenated tree (84% or greater among replicates), providing strong support for these relationships reflecting the accurate evolutionary history of this group.

We also find that the failure to account for heterozygous sites in phylogenomic inference may not influence results greatly. Here, trees were estimated with identical species relationships regardless of whether heterozygous sites were included in the analysis or not. However, bootstrap support values differed substantially between analyses; in particular, the node between *T. cinereicollis/T. quadrivittatus* and *T. rufus* shifts from being highly supported to weakly supported, and the same holds for the node uniting *T. rufus* and *T. canipes*.

We also estimate population genetic statistics by calling variants relative to a set of ARC contigs from a single individual. This provides a useful set of variants in the absence of an established chipmunk reference. Since previous studies indicate rampant mitochondrial introgression, we expected to see some evidence of admixture. Mean estimates of F_{ST} suggest that gene flow may present between species pairs, with the greatest amount between *T. cinereicollis* and *T. quadrivittatus* and the least between *T. umbrinus* and any other species. This differs from the conclusions drawn from mitochondrial phylogenomic studies; such analyses indicated an absence of mitochondrial introgression between *T. rufus*, *T. canipes* but rampant mitochondrial introgression in *T. dorsalis* and other species. Furthermore, F_{ST} values calculated from mitochondrial protein-coding genes are consistent with the widespread mitochondrial introgression observed in this system. In contrast, *T. rufus* and *T. canipes* F_{ST} estimates obtained from nuclear data are approximately equivalent

to other comparisons, producing a stark contrast to the low values estimated from mitochondrial data.

Furthermore, ADMIXTURE results indicate little coancestry among species; across all species, individuals almost exclusively share coancestry exclusively with individuals of their same species. Multidimensional scaling indicates that each species group shows that individuals cluster only with their assigned species. These results suggest that even though there has been rampant mitochondrial introgression within *Tamias*, relatively little nuclear introgression has taken place.

It may be that some mechanism, such as selection against hybrids, prevents nuclear introgression in the face of ongoing gene flow. In *T. sibiricus*, male reproductive success is correlated with range size, and hybrids of intermediate size may be inferior competitors for large territories (Marmet et al. 2012). However, it is unclear whether this holds for western North American chipmunks, though similar findings in the Eastern chipmunk (*T. striatus*) indicate that this may be the case (Yahner 1978). Recent work in *Mus* has shown that divergent bacular morphologies are likely the result of sexual selection (Stockley et al. 2013; Simmons and Firman 2014). Mechanical stimulation may play an important role in reproductive success, and hybrid bacula may perform suboptimally in this context. It is possible, then, that sexual selection plays a central role in explaining genetic patterns in this system. Breeding work will be required to address these hypotheses.

This study represents the first genomic-scale study in central and southern Rocky Mountains chipmunks and one of the first in chipmunks as a species (see Bi et al. 2012, 2013). Here, however, we only consider species that comprise 24% of the diversity of the genus. Future work will focus on analyzing genus-wide capture data to arrive at more

general conclusions about the genomic nature of divergence in this system. Specifically, the assembly and phylogenomic approaches implemented here can be used across species to provide resolution across the genus and build on approaches using few loci (e.g., Reid et al. 2012). A resolved phylogeny, in concert with population genomic estimates of divergence, gene flow, and population structure, will result in a comprehensive characterization of this natural system.

Conclusion

Here, we use targeted exon capture to sequence thousands of nuclear loci from chipmunks in the *T. quadrivittatus*-group. Using phylogenomic approaches, we are able to produce a phylogeny using a variety of techniques that resolves the systematics of this group. Furthermore, we document a lack of nuclear introgression in the face of substantial mitochondrial introgression. Future work will characterize this system further using additional analyses and increased, genus-wide sampling.

References

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–64.

Anderson, E. 1949. Introgressive Hybridization. Wiley & Sons, New York.

Bayzid, M. S., and T. Warnow. 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29:2277–84.

Bi, K., T. Linderoth, D. Vanderpool, J. M. Good, R. Nielsen, and C. Moritz. 2013. Unlocking the vault: next-generation museum population genomics. Mol. Ecol. 22:6018–32.

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. 84:210–23.

Browning, B. L., and Z. Yu. 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. 85:847–61.

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–97.

Butlin, R., A. Debelle, C. Kerth, R. R. Snook, L. W. Beukeboom, R. F. Castillo Cajas, W. Diao, M. E. Maan, S. Paolucci, F. J. Weissing, L. van de Zande, A. Hoikkala, E. Geuverink, J. Jennings, M. Kankare, K. E. Knott, V. I. Tyukmaeva, C. Zoumadakis, M. G. Ritchie, D. Barker, E. Immonen, M. Kirkpatrick, M. Noor, C. Macias Garcia, T. Schmitt, and M. Schilthuizen. 2012. What do we need to know about speciation? Trends Ecol. Evol. 27:27–39.

Butlin, R. K., and M. G. Ritchie. 2009. Genetics of speciation. Heredity. 102:1–3.

Castric, V., J. Bechsgaard, M. H. Schierup, and X. Vekemans. 2008. Repeated adaptive introgression at a gene under multiallelic balancing selection. PLoS Genet. 4:e1000168.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Sunderland, Massachusetts.

Danecek, P., A. Auton, G. Abecasis, C. a Albers, E. Banks, M. a DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–8.

DePristo, M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–8.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–7.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Good, J. M., J. R. Demboski, D. W. Nagorsen, and J. Sullivan. 2003. Phylogeography and introgressive hybridization: chipmunks (genus Tamias) in the northern Rocky Mountains. Evolution 57:1900–16.

Good, J. M., S. Hird, N. Reid, J. R. Demboski, S. J. Steppan, T. R. Martin-Nims, and J. Sullivan. 2008. Ancient hybridization and mitochondrial capture between two species of chipmunks. Mol. Ecol. 17:1313–27.

Harrison, R. G. and E. L. Larson. 2014. Hybridization, Introgression, and the Nature of Species Boundaries. Journal of Heredity. In press.

Hewitt, G. M. 1988. Hybrid zones-natural laboratories for evolutionary studies. Trends Ecol. Evol. 3:158–67.

Hird, S., N. Reid, J. Demboski, and J. Sullivan. 2010. Introgression at differentially aged hybrid zones in red-tailed chipmunks. Genetica 138:869–83.

Hird, S., and J. Sullivan. 2009. Assessment of gene flow across a hybrid zone in red-tailed chipmunks (Tamias ruficaudus). Mol. Ecol. 18:3097–109.

Kent, W. J. 2002. BLAT--The BLAST-Like Alignment Tool. Genome Res. 12:656–664.

Kim, M., M.-L. Cui, P. Cubas, A. Gillies, K. Lee, M. A. Chapman, R. J. Abbott, and E. Coen. 2008. Regulatory genes control a key morphological and ecological trait transferred between species. Science 322:1116–9.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–9.

Li, T., P. C. M. O'Brien, L. Biltueva, B. Fu, J. Wang, W. Nie, M. a Ferguson-Smith, A. S. Graphodatsky, and F. Yang. 2004. Evolution of genome organizations of squirrels (Sciuridae) revealed by cross-species chromosome painting. Chromosome Res. 12:317–35.

Librado, P., and J. Rozas. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–2.

Liu, L., L. Yu, and S. V Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302. BioMed Central Ltd.

Liu, L., L. Yu, D. K. Pearl, and S. V Edwards. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–77.

Magoč, T., and S. L. Salzberg. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–63.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–37.

Marmet, J., B. Pisanu, J.-L. Chapuis, G. Jacob, and E. Baudry. 2012. Factors affecting male and female reproductive success in a chipmunk (Tamias sibiricus) with a scramble competition mating system. Behav. Ecol. Sociobiol. 66:1449–1457.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. a DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–303.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. Syst. Biol. 52:674–683.

Nosil, P., and J. L. Feder. 2012. Genomic divergence during speciation: causes and consequences. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:332–42.

Nosil, P., D. J. Funk, and D. Ortiz-Barrientos. 2009. Divergent selection and heterogeneous genomic divergence. Mol. Ecol. 18:375–402.

Nosil, P., and D. Schluter. 2011. The genes underlying the process of speciation. Trends Ecol. Evol. 26:160–7.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289–290.

Pinho, C., and J. Hey. 2010. Divergence with Gene Flow: Models and Data. Annu. Rev. Ecol. Evol. Syst. 41:215–230.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–75.

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

Reid, N., J. R. Demboski, and J. Sullivan. 2012. Phylogeny estimation of the radiation of western North American chipmunks (Tamias) in the face of introgression using reproductive protein genes. Syst. Biol. 61:44–62.

Reid, N., S. Hird, A. Schulte-Hostedde, and J. Sullivan. 2010. Examination of nuclear loci across a zone of mitochondrial introgression between Tamias ruficaudus and T. amoenus. J. Mammal. 91:1389–1400.

Rieseberg, L. H. 2009. Evolution: replacing genes and traits through hybridization. Curr. Biol. 19:R119–22.

Rieseberg, L., B. Sinervo, C. Linder, M. Ungerer, and D. Arias. 1996. Role of Gene Interactions in Hybrid Speciation: Evidence from Ancient and Experimental Hybrids. Science (80-.). 272:741–5.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. a Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, A. Brännström, A. Brelsford, C. S. Clarkson, F. Eroukhmanoff, J. L. Feder, M. C. Fischer, A. D. Foote, P. Franchini, C. D. Jiggins, F. C. Jones, A. K. Lindholm, K. Lucek, M. E. Maan, D. a Marques, S. H. Martin, B. Matthews, J. I. Meier, M. Möst, M. W. Nachman, E. Nonaka, D. J. Rennison, J. Schwarzer, E. T. Watson, A. M. Westram, and A. Widmer. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–92. Nature Publishing Group.

Simmons, L. W., and R. C. Firman. 2014. Experimental evidence for the evolution of the Mammalian baculum by sexual selection. Evolution 68:276–83.

Smadja, C. M., and R. K. Butlin. 2011. A framework for comparing processes of speciation in the presence of gene flow. Mol. Ecol. 20:5123–40.

Smith, S. a, and C. W. Dunn. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–6.

Smith, S. A., and B. C. O'Meara. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics 28:2689–90.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2010–2011.

Stockley, P., S. a Ramm, A. L. Sherborne, M. D. F. Thom, S. Paterson, and J. L. Hurst. 2013. Baculum morphology predicts reproductive success of male house mice under sexual selection. BMC Biol. 11:66.

Strasburg, J. L., N. A. Sherman, K. M. Wright, L. C. Moyle, J. H. Willis, and L. H. Rieseberg. 2012. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:364–73.

Sullivan, J., J. R. Demboski, K. C. Bell, S. Hird, B. A. J. Sarver, N. Reid, and J. M. Good. 2014. Divergence-with-gene-flow within the recent chipmunk radiation (Tamias). Heredity.

Twyford, A. D., and R. A. Ennos. 2012. Next-generation hybridization and introgression. Heredity. 108:179–89.

Weir, B., and C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 38:1358–1370.

Wu, C.-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Yahner, R. H. 1978. The adaptive nature of the social system and behavior in the eastern chipmunk, Tamias striatus. Behav. Ecol. Sociobiol. 3:397–427.

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.

Figures

Figure 4.1: Best-scoring RAxML tree estimated including heterozygous sites. Bootstrap values are listed above branches. Individuals have been collapsed into species groups for ease of viewing. Tree is rooted on the branch leading to *T. striatus*.

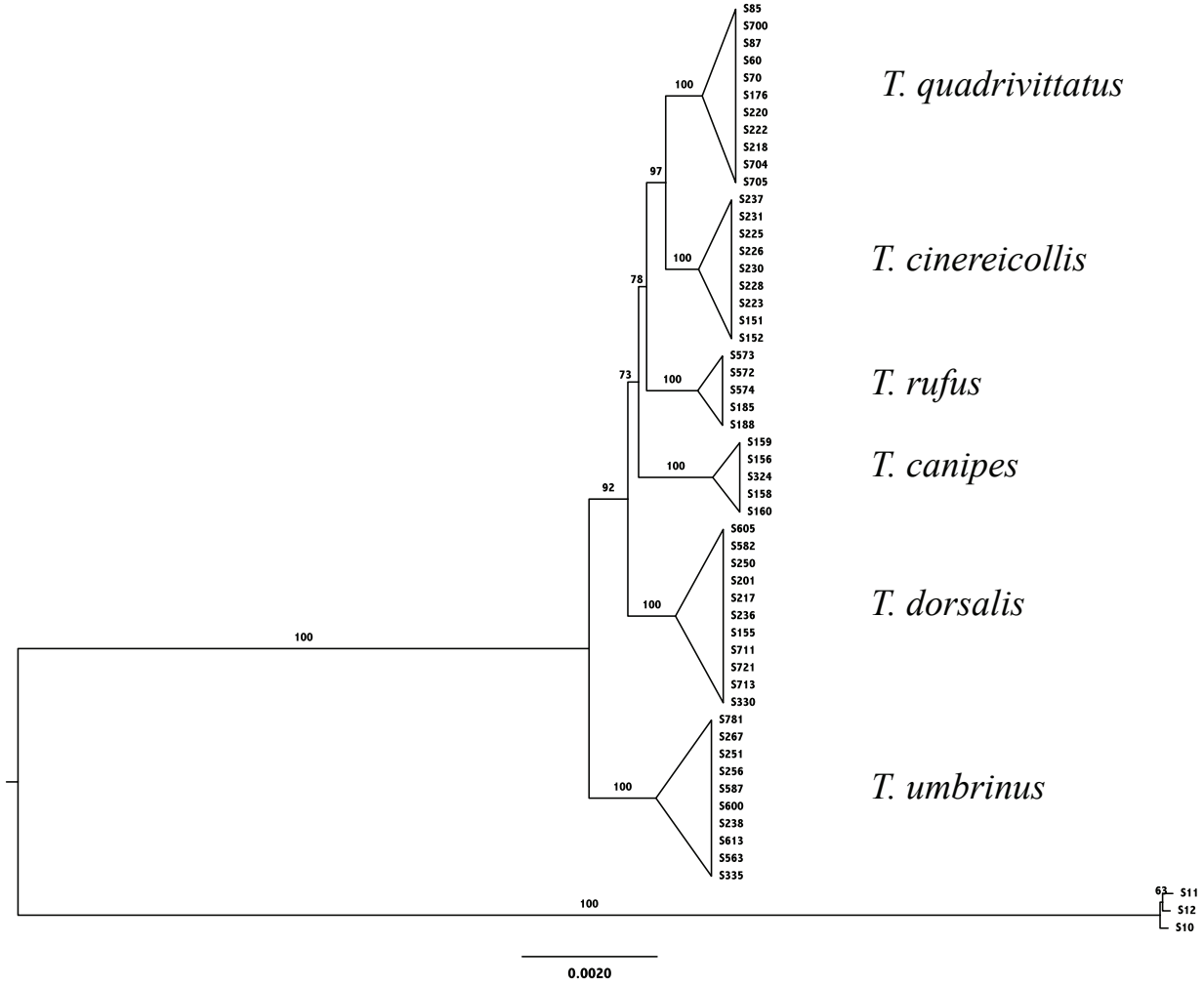


Figure 4.2: Best-scoring RAxML tree estimated excluding heterozygous sites. Bootstrap values are listed above branches. Individuals have been collapsed into species groups for ease of viewing. Tree is rooted on the branch leading to *T. striatus*. Bootstraps are higher for splits with *T. rufus* and *T. canipes* than in Figure 1.

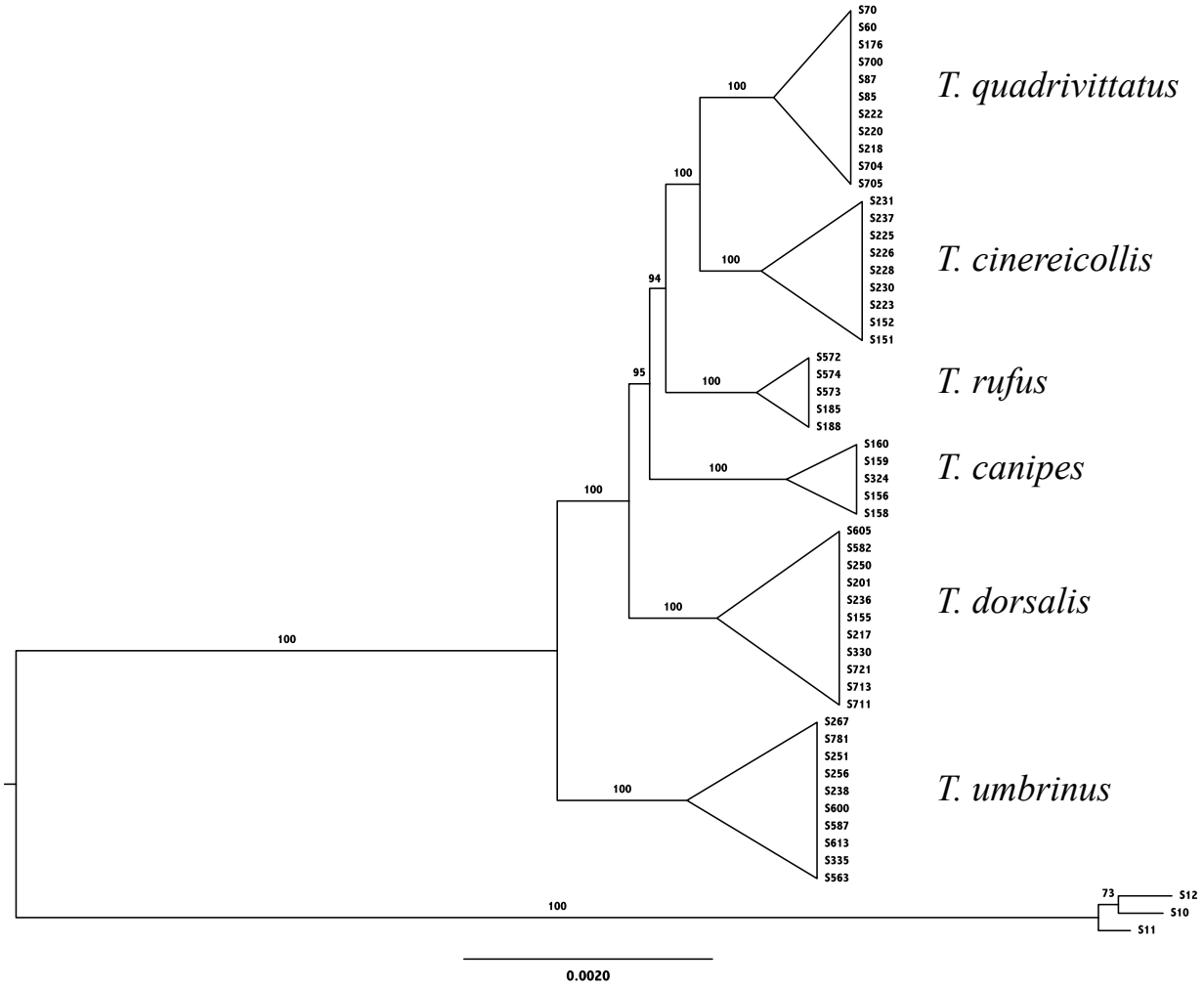
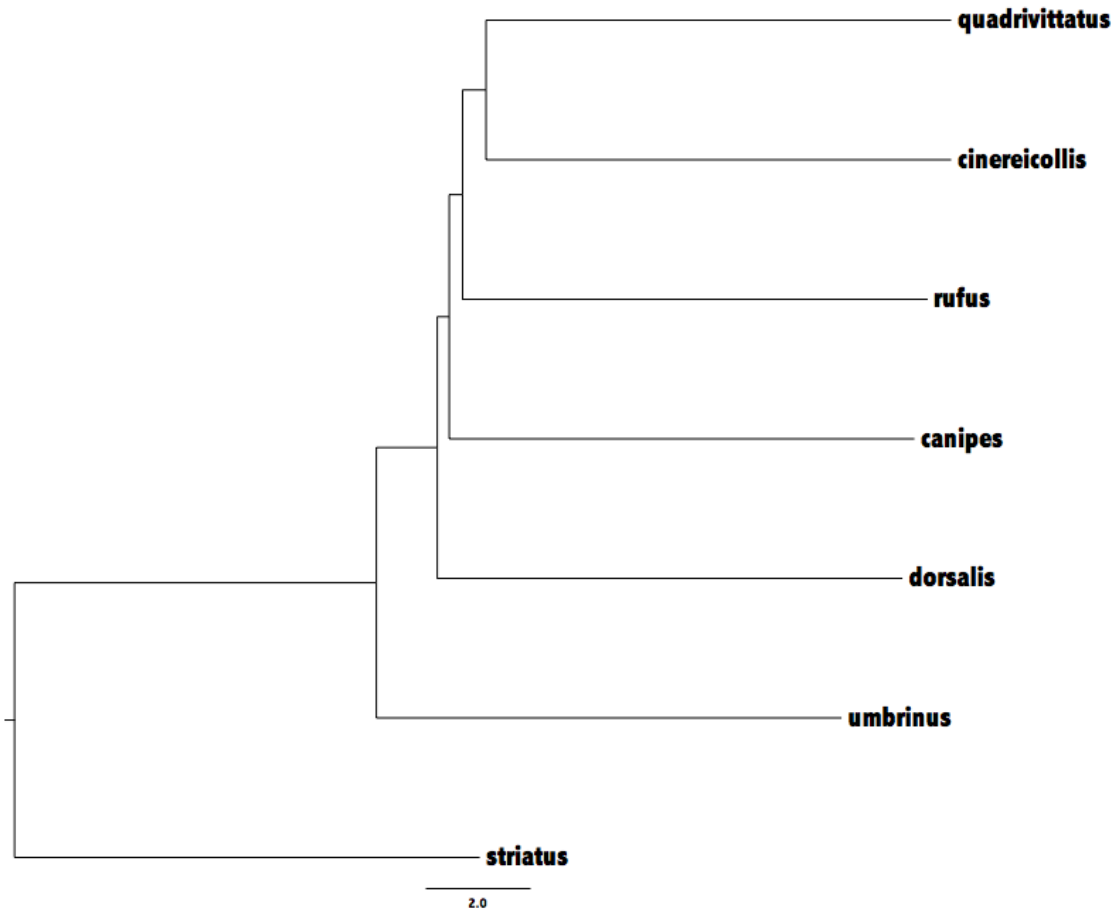
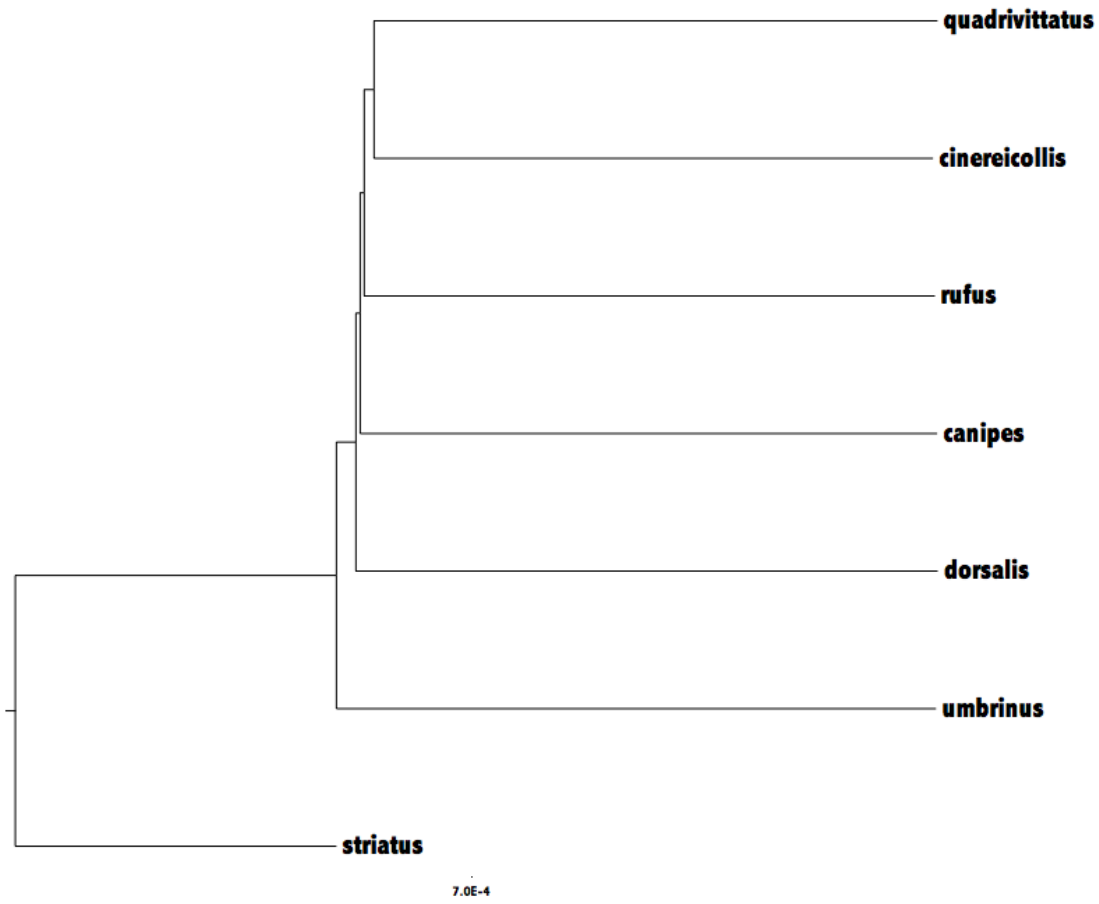


Figure 4.3: Species trees estimated from MP-EST, STAR, and STEAC for ARC contigs assigned to chromosomes. All trees are in agreement with the exception of a single STEAC tree. A: MP-EST; B: STAR-NJ; C: STAR-UPGMA; D: STEAC-NJ; E: STEAC-UPGMA.

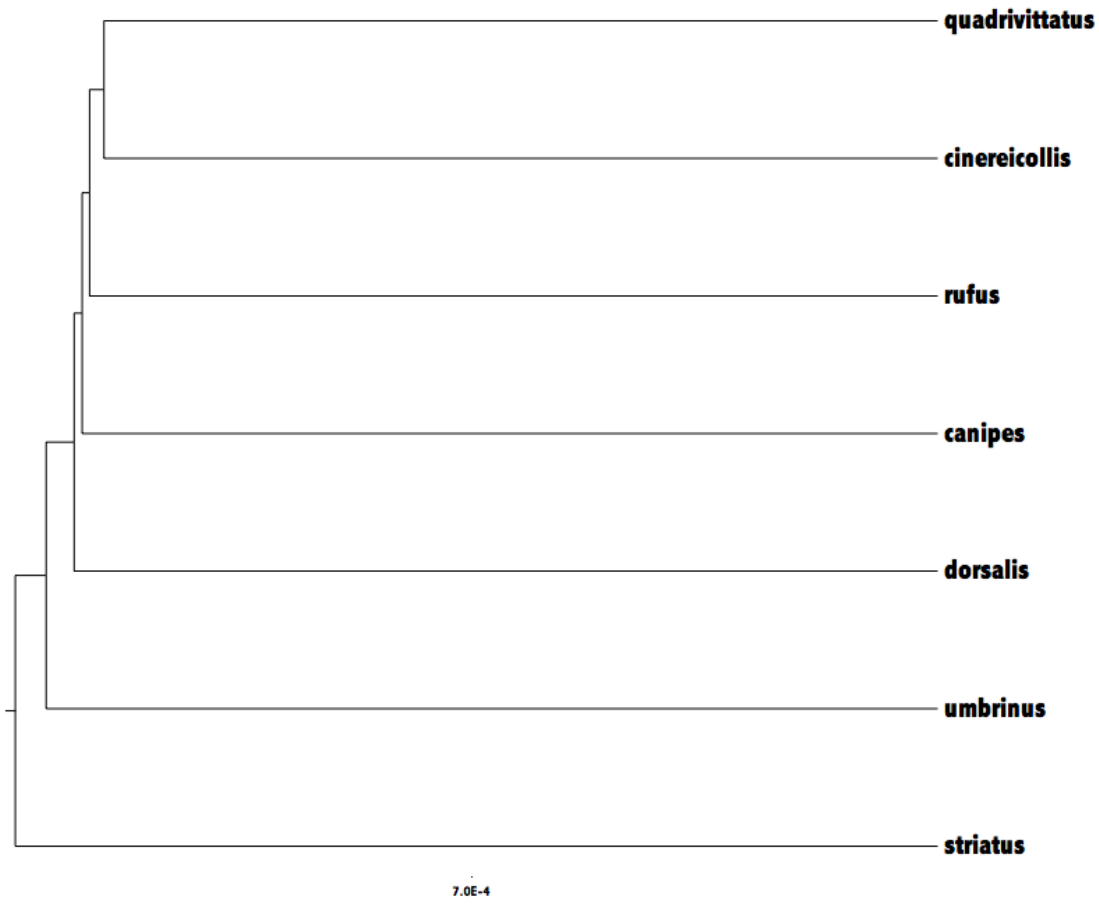
A:



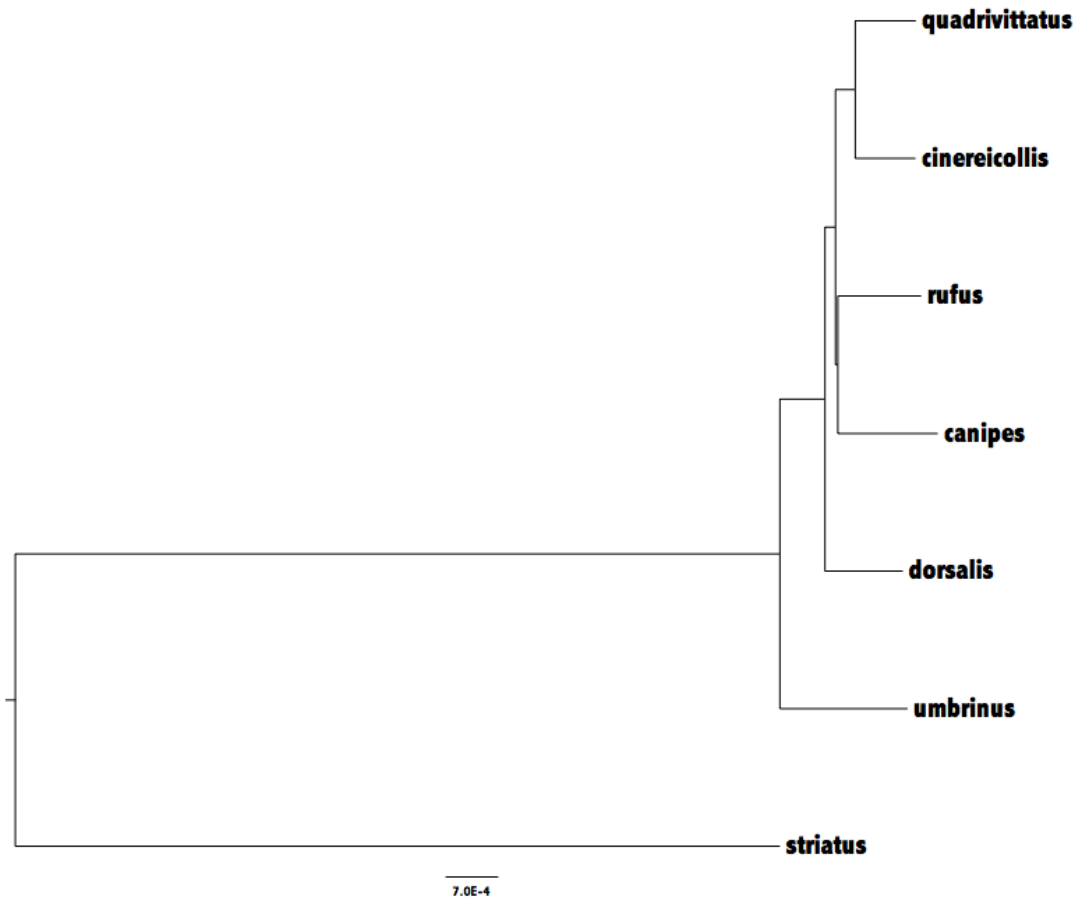
B:


quadrivittatus
cinereicollis
rufus
canipes
dorsalis
umbrinus
striatus
7.0E-4

C:


quadrivittatus
cinereicollis
rufus
canipes
dorsalis
umbrinus
striatus
7.0E-4

D:


quadrivittatus
cinereicollis
rufus
canipes
dorsalis
umbrinus
striatus
7.0E-4

E:

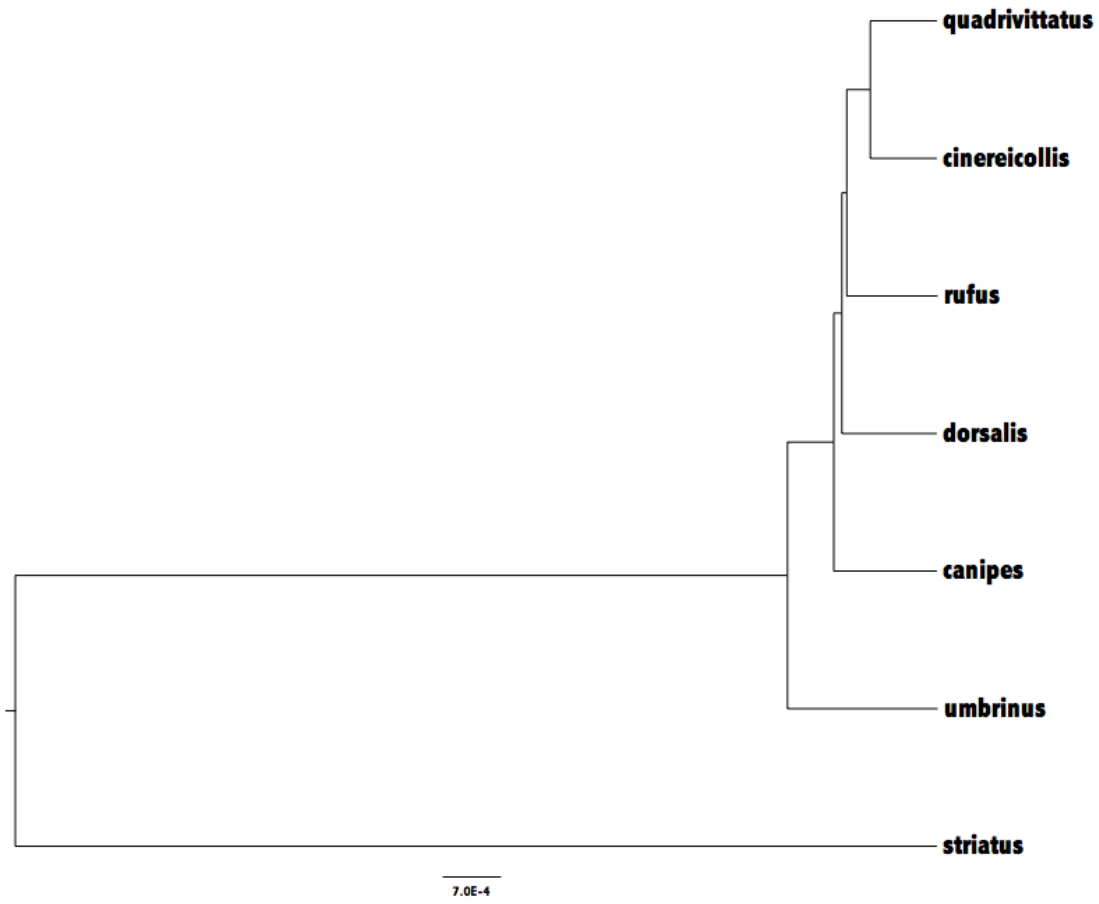

quadrivittatus
cinereicollis
rufus
dorsalis
canipes
umbrinus
striatus
7.0E-4

Figure 4.4: ADMIXTURE coancestry plot for 10 values of K. Colors are indicative of population assignment. The best-selected value of K based on cross-validation error is 7. Little shared coancestry is indicated among species.

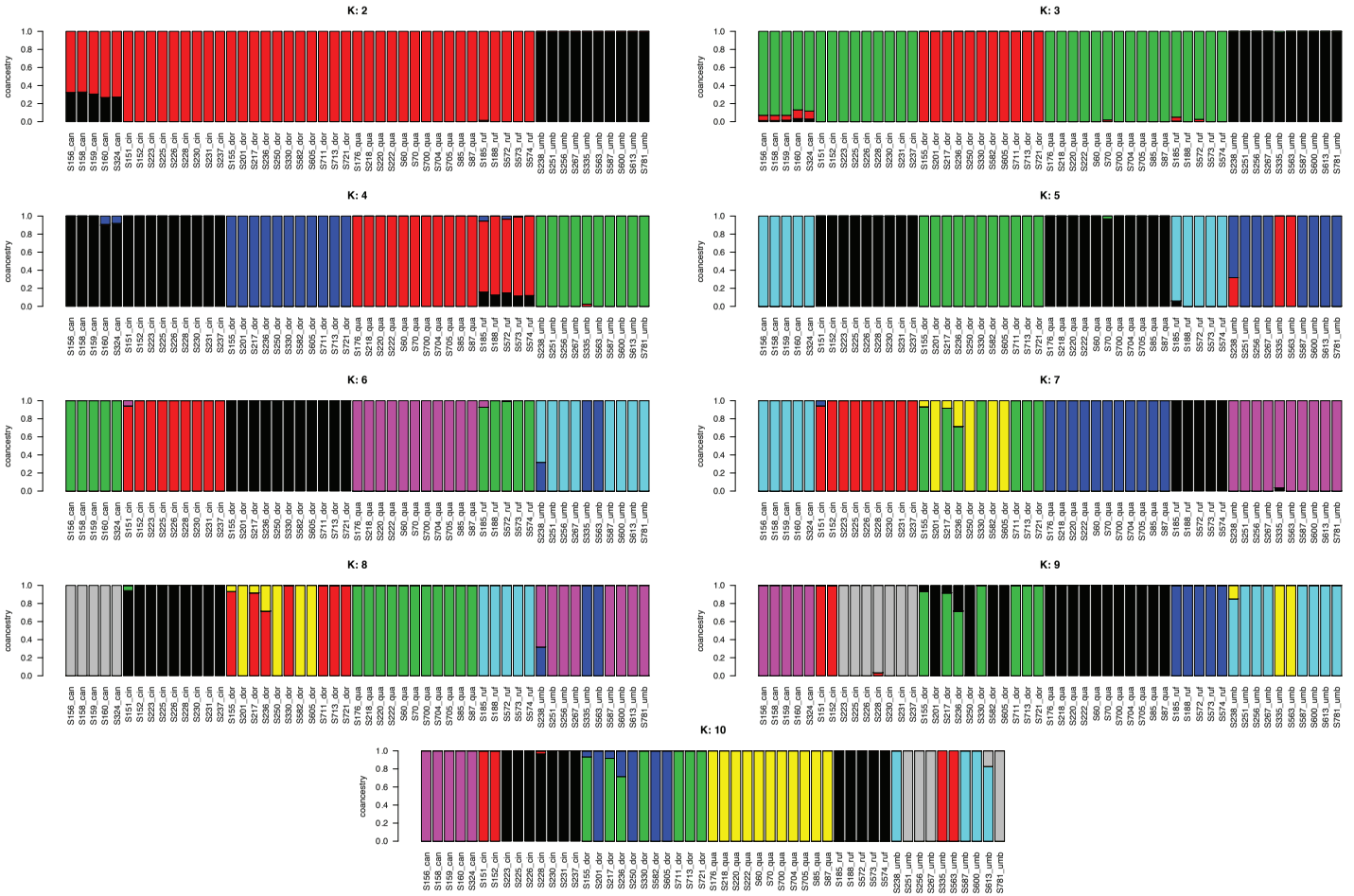
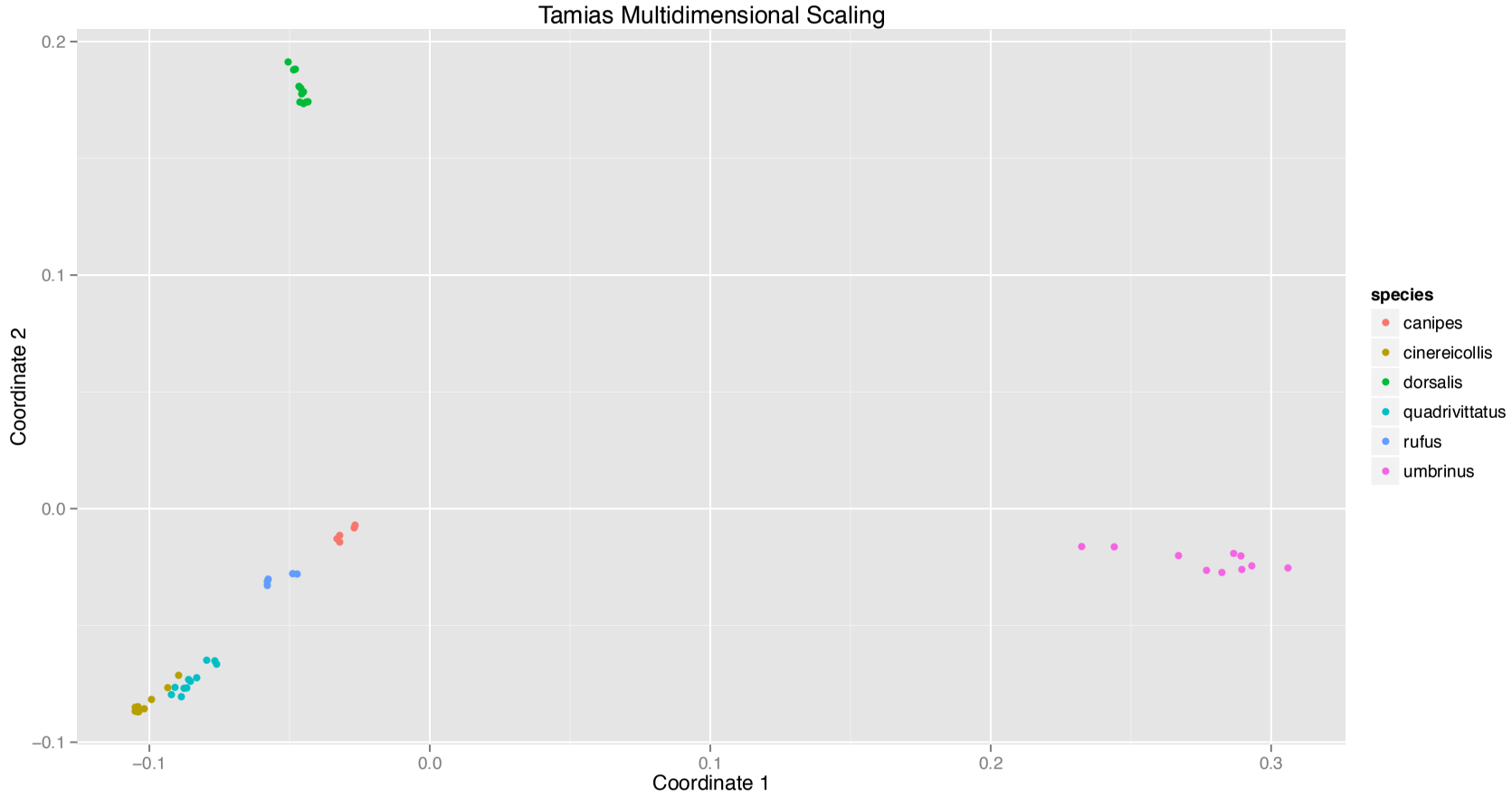


Figure 4.5: Two-dimensional multidimensional scaling plot generated from genome-wide SNPs. Individuals cluster according to their species assignments. Distances coincide with the phylogeny estimated using a subset contigs.



Tables

Table 4.1: Robinson-Foulds distances relative to the concatenated RAxML phylogeny. Distances are grouped based on the method of species tree estimation. The final row lists the number of replicate trees that are in agreement with the RAxML tree. There is a high percentage of concordance across all replicates and all approaches.

Replicate	MP-EST	STAR (NJ)	STAR (UPGMA)	STEAC (NJ)	STEAC (UPGMA)
1	2	0	2	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	2	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	2	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
11	0	0	0	2	2
12	0	0	0	0	0
13	0	0	0	0	0
14	0	0	0	0	0
15	0	0	0	0	0
16	0	0	0	0	0
17	2	0	0	0	0
18	0	0	0	0	0
19	2	2	2	0	0
20	0	0	0	0	0
21	0	2	2	0	0
22	0	0	0	0	0
23	0	0	0	0	0
24	0	0	0	0	0
25	0	0	0	0	0
% in agreement:	**0.84**	**0.92**	**0.84**	**0.96**	**0.96**

Table 4.2: Genomic characterization. Mean number of sites refers to the total length of sequence data that has a sequencing depth of at least one averaged across all individuals in each species pool. F_{IS} is calculated per SNP for each species pool and then averaged over all sites. Non-informative sites are removed before calculation.

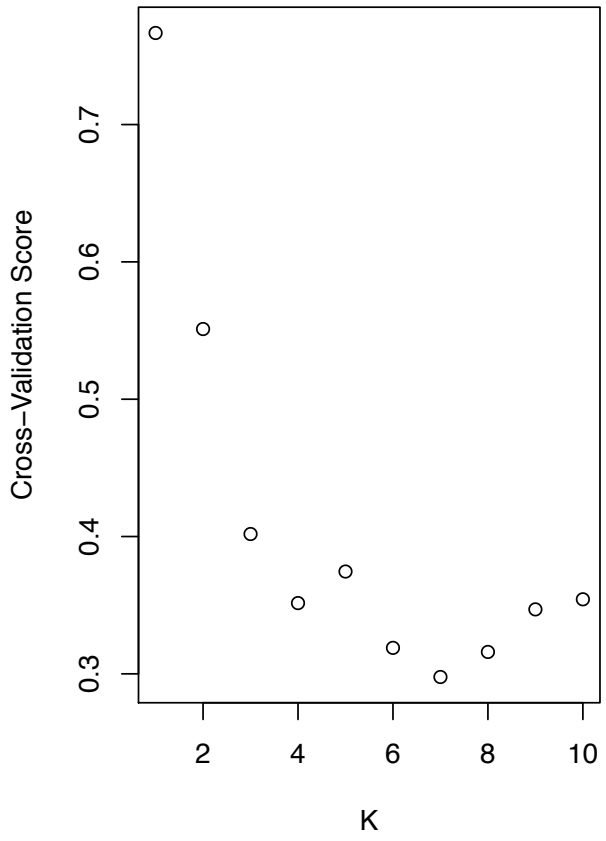
Species	Mean Number of Sites	Total Number of Heterozygous Genotypes	Mean Observed Heterozygosity	F_{IS}
T. canipes	4298388	30658	0.0014	0.0473
T. cinereicollis	4793771	61382	0.0014	0.0465
T. dorsalis	4826024	73022	0.0014	0.0683
T. quadrivittatus	4564067	67181	0.0013	0.0479
T. rufus	4409452	29748	0.0013	0.0374
T. umbrinus	4781474	54540	0.0011	0.0755

Table 4.3: Pairwise F_{ST} estimates. Estimates from this study are on the lower diagonal. Sites with F_{ST} values less than zero are set equal to zero. Non-informative sites are removed prior to calculating the mean. The upper diagonal contains estimates from mitochondrial exons. Mitochondrial estimates suggest introgression between some pairs (i.e., estimates involving *T. dorsalis*) and a lack of introgression between others (i.e., estimates involving *T. canipes* and *T. rufus*), in agreement with previous studies. Nuclear and mitochondrial estimates often differ substantially.

	T. canipes	*T. cinereicollis*	*T. dorsalis*	*T. quadrivittatus*	*T. rufus*	*T. umbrinus*
T. canipes	-	0.90039	0.77282	0.7423	0.94093	0.83481
T. cinereicollis	0.3039	-	0.13329	0.34286	0.86096	0.39919
T. dorsalis	0.3185	0.3005	-	0.0903	0.63409	0.05138
T. quadrivittatus	0.2939	0.1032	0.2899	-	0.59239	0.22374
T. rufus	0.281	0.2563	0.305	0.2348	-	0.73139
T. umbrinus	0.4012	0.3998	0.3762	0.3905	0.4017	-

Supplementary Material

Supplementary Figure 4.1: ADMIXTURE cross-validation plot for 10 values of K.



Chapter 5

Conclusions and Future Directions

This dissertation contains three separate studies. Chapter 2 describes series of phylogenetic simulations used to address inferences of diversification rates from molecular phylogenies. Chapter 3 is a mitogenomic study that investigates patterns and causes of mitochondrial introgression in central and southern Rocky Mountains chipmunks. Chapter 4 is a nuclear phylogenomic/population genomic study that resolves phylogenetic relationships and describes patterns of nuclear introgression in the same system.

Chapter 2

Phylogenetic simulations reveal that, unless there is extreme rate heterogeneity, the choice of tree prior and choice of molecular clock does not strongly affect estimates of diversification rates. As a result, an uncorrelated lognormal relaxed molecular clock should be used to avoid error associated with a failure to accurately model rate heterogeneity among lineages. Since trees are now made from large, phylogenomic datasets spanning multiple families, it is reassuring that diversification rate estimates are not significantly affected by prior misspecification. However, Bayesian phylogenetic techniques are often computationally intractable with many individuals or a large amount of sequence data. Current methods for analyzing large datasets consist of using approximate-likelihood approaches, such as RAxML (Stamatakis et al. 2005; Stamatakis 2014), under a single model of nucleotide sequence evolution, and trees are transformed using a molecular clock afterwards. It remains to be seen whether computational and methodological advances in likelihood-based phylogenetics can accommodate the amount of data generated using high-

throughput sequencing techniques. Additionally, this study simulates trees under combinations of tree priors and molecular clocks by sampling from fixed priors, an approach not seen in other studies. Future studies can use trees simulated this way to assess the impact of phylogenetic error on the estimation of the magnitude and number of shifts in the rate of morphological diversification (e.g., Eastman et al. 2011; Slater et al. 2012).

Chapter 3

Results from selection analyses suggest that selection is not playing a role in governing mitochondrial introgression in central and southern Rocky Mountains chipmunks. Demographic factors, such as population expansion, provide a possible explanation of introgression patterns. An interesting evolutionary case involving tRNA-lysine reveals that iterative assembly using ARC is an appropriate approach to recover mitochondrial genomes. Furthermore, I develop a Bayesian model-averaging approach to estimate model-averaged parameters from a series of selection models implemented in codeml in PAML (Yang 2007). This approach builds on work on decision-theoretic approaches in phylogenetic model selection (Minin et al. 2003) and can be used to incorporate model uncertainty into selection analyses. Future work will include sequencing mitochondrial genomes from across *Tamias*, allowing the characterization of genus-wide patterns of introgression and increase power to detect selection. Demographic scenarios can be tested using coalescent simulations; increased sampling will produce parameter estimates that can be used to erect a series of simulations and test demographic hypotheses, as well as selection, explicitly (Hudson 2002; Ewing and Hermisson 2010).

Chapter 4

In contrast to patterns of mitochondrial introgression, there appears to be little nuclear introgression in central and southern Rocky Mountains chipmunks. Estimates of individual coancestry, combined with clustering resulting from multidimensional scaling, indicate that each species is recovered as a distinct group using nuclear data. F_{ST} estimates generally follow the phylogeny, with *T. quadrivittatus* and *T. cinereicollis* having the smallest estimate and any comparison with *T. umbrinus* having the greatest estimate. All other estimates fall within this range. F_{ST} estimates from mitochondrial genomic data support the phylogenetic notion of rampant introgression among some species, and near isolation among others. Furthermore, phylogenies estimated from concatenation and a variety of species-tree approaches, whether loci are binned based on chromosome assignment or at random, recover a single tree in the majority of cases across all methods. As a result, this study produces the first conclusive phylogeny for central and southern Rocky Mountains chipmunks. Future work will focus on detecting fine-scale patterns of introgression using ABBA-BABA tests (Green et al. 2010; Durand et al. 2011) and identifying loci that act as divergence and/or F_{ST} outliers. In addition, exome capture datasets are being generated for individuals spanning the genus. This data will, for the first time, allow for the characterization of genomic patterns at this scale in chipmunks and, potentially, resolve phylogenetic relationships among all species. The phylogenomic approaches implemented for this chapter will be combined into an R library for use in other systems. This will increase the usability of ARC among biologists by expediting analyses in non-model systems.

References

(The following references cover the "Introduction" and "Conclusions and Future Directions" sections only)

Bi, K., D. Vanderpool, S. Singhal, T. Linderoth, C. Moritz, and J. M. Good. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13:403.

Brown, J. H. 1971. Mechanisms of Competitive Exclusion Between Two Species of Chipmunks. Ecology 52:305.

Charlesworth, B. 2010. Molecular population genomics: a short history. Genet. Res. 92:397–411.

Coyne, J. A., and H. A. Orr. 2004. Speciation.

Currat, M., M. Ruedi, R. J. Petit, and L. Excoffier. 2008. The hidden side of invasions: massive introgression by local genes. Evolution 62:1908–20.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–75.

Drummond, A. J., M. a Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–73.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–52.

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–89.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Eisen, J. A. 1998. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Res. 8:163–167.

Ewing, G., and J. Hermisson. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064–5.

Excoffier, L., M. Foll, and R. J. Petit. 2009. Genetic Consequences of Range Expansions. Annu. Rev. Ecol. Evol. Syst. 40:481–501.

Funk, D. J., and K. E. Omland. 2003. Species-level paraphyly and polyphyly : Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.

Gernhard, T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–78.

Good, J. M., J. R. Demboski, D. W. Nagorsen, and J. Sullivan. 2003. Phylogeography and introgressive hybridization: chipmunks (genus Tamias) in the northern Rocky Mountains. Evolution 57:1900–16.

Good, J. M., S. Hird, N. Reid, J. R. Demboski, S. J. Steppan, T. R. Martin-Nims, and J. Sullivan. 2008. Ancient hybridization and mitochondrial capture between two species of chipmunks. Mol. Ecol. 17:1313–27.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. a Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo. 2010. A draft sequence of the Neandertal genome. Science 328:710–22.

Guttmacher, A. E., and F. S. Collins. 2003. Welcome to the genomic era. N. Engl. J. Med. 349:996–8.

Heller, H. C. 1971. Altitudinal Zonation of Chipmunks (Eutamias): Interspecific Aggression. Ecology 52:312.

Heller, H. C., and D. M. Gates. 1971. Altitudinal Zonation of Chipmunks (Eutamias): Energy Budgets. Ecology 52:424.

Hird, S., N. Reid, J. Demboski, and J. Sullivan. 2010. Introgression at differentially aged hybrid zones in red-tailed chipmunks. Genetica 138:869–83.

Hird, S., and J. Sullivan. 2009. Assessment of gene flow across a hybrid zone in red-tailed chipmunks (Tamias ruficaudus). Mol. Ecol. 18:3097–109.

Hodges, E., M. Rooks, Z. Xuan, A. Bhattacharjee, D. Benjamin Gordon, L. Brizuela, W. Richard McCombie, and G. J. Hannon. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protoc. 4:960–74.

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M. N. Molla, S. W. Smith, C. M. Middle, M. J. Rodesch, T. J. Albert, G. J. Hannon, and W. R. McCombie. 2007. Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39:1522–7.

Höhna, S. 2014. Likelihood inference of non-constant diversification rates with incomplete taxon sampling. PLoS One 9:e84184.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Jorde, L. B., W. S. Watkins, and M. J. Bamshad. 2001. Population genomics : a bridge from evolutionary history to genetic medicine. Hum. Mol. Genet. 10:2199–2208.

Kendall, D. G. 1948. On the Generalized "Birth-and-Death" Process. Ann. Math. Stat. 19:1–15.

Klopfstein, S., M. Currat, and L. Excoffier. 2006. The fate of mutations surfing on the wave of a range expansion. Mol. Biol. Evol. 23:482–90.

Maddison, W. P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Mallet, J. 2007. Hybrid speciation. Nature 446:279–83.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–37.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. Syst. Biol. 52:674–683.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994a. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 344:77–82.

Nee, S., R. M. May, and P. H. Harvey. 1994b. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 344:305–11.

Pinho, C., and J. Hey. 2010. Divergence with Gene Flow: Models and Data. Annu. Rev. Ecol. Evol. Syst. 41:215–230.

Reid, N., J. R. Demboski, and J. Sullivan. 2012. Phylogeny estimation of the radiation of western North American chipmunks (Tamias) in the face of introgression using reproductive protein genes. Syst. Biol. 61:44–62.

Reid, N., S. Hird, A. Schulte-Hostedde, and J. Sullivan. 2010. Examination of nuclear loci across a zone of mitochondrial introgression between Tamias ruficaudus and T. amoenus. J. Mammal. 91:1389–1400.

Revell, L., L. Harmon, and R. Glor. 2005. Under-parameterized Model of Sequence Evolution Leads to Bias in the Estimation of Diversification Rates from Molecular Phylogenies. Syst. Biol. 54:973–983.

Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. a Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, A. Brännström, A. Brelsford, C. S. Clarkson, F. Eroukhmanoff, J. L. Feder, M. C. Fischer, A. D. Foote, P. Franchini, C. D. Jiggins, F. C. Jones, A. K. Lindholm, K. Lucek, M. E. Maan, D. a Marques, S. H. Martin, B. Matthews, J. I. Meier, M. Möst, M. W. Nachman, E. Nonaka, D. J. Rennison, J. Schwarzer, E. T. Watson, A. M. Westram, and A. Widmer. 2014. Genomics and the origin of species. Nat. Rev. Genet. 15:176–92.

Slater, G. J., L. J. Harmon, D. Wegmann, P. Joyce, L. J. Revell, and M. E. Alfaro. 2012. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. Evolution 66:752–62.

Stadler, T. 2013. How can we improve accuracy of macroevolutionary rate estimates? Syst. Biol. 62:321–9.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–63.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2010–2011.

Sullivan, J., J. R. Demboski, K. C. Bell, S. Hird, B. A. J. Sarver, N. Reid, and J. M. Good. 2014. Divergence-with-gene-flow within the recent chipmunk radiation (Tamias). Heredity.

Wertheim, J. O., and M. J. Sanderson. 2011. Estimating diversification rates: how useful are divergence times? Evolution 65:309–20.

Wu, C.-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851–865.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–91.

Yule, G. U. 1925. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. Philos. Trans. R. Soc. B Biol. Sci. 213:21–87.