Title
Using the Birth-Death Process to Infer Changes in the Pattern of Lineage Gain and Loss

Author
Hallinan, Nathaniel Malachi

Publication Date
2011

Peer reviewed|Thesis/dissertation

Using the Birth-Death Process to Infer Changes in the Pattern of Lineage Gain and Loss

by

Nathaniel Malachi Hallinan

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Lindberg, Chair
Professor John Huelsenbeck
Professor David Aldous

Fall 2011

Using the Birth-Death Process to Infer Changes in the Pattern of Lineage Gain and Loss

Abstract

Using the Birth-Death Process to Infer Changes in the Pattern of Lineage Gain and Loss

by

Nathaniel Malachi Hallinan

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor David Lindberg, Chair

The birth-death process has been used to study the evolution of a wide variety of biological entities from genes to species. Much recent work has turned to detecting changes in the patterns of lineage splitting by comparing data to birth-death models in which the parameters vary between lineages or over time. Here, I develop methods to investigate how the birth-death process varies under three very different circumstances: changes in the pattern of taxon diversification through time; the effect of whole genome duplications on the pattern of chromosome gain and loss; and changes in the pattern of gene gain and loss on branches of a taxon tree. For all three cases I apply my methods to some real data.

For the last fifteen years researchers have studied the distribution of branching times of a phylogeny of extant taxa in order to detect temporal changes in the process of diversification. Theoretical work on this subject has been based on different implementations of the birth-death process and has proceeded along three basic lines: the comparison of actual branching times to a birth-death process; the inference of the effects of different birth-death processes on the distribution of branching times; and the derivation of analytical results that describe various aspects of different birth-death processes. In chapter 2 I make contributions to all three lines of research for the reconstructed time variable birth-death process.

Previous work had shown how to calculate the distributions of number of lineages and branching times for a reconstructed constant rate birth-death process that started with one or two reconstructed lineages at some time or ended with some number of lineages in the present. In chapter 2 I expand that work to include any time variable birth-death process that starts with any number of reconstructed lineages and/or ends with any number of reconstructed lineages at any time. I also introduce the discrete time birth-death process which operates as an efficient and accurate numerical solution to any time-variable birth death process and allows for the analytical incorporation of sampling and mass extinctions. Furthermore, I show how to simulate random trees under any of these models.

In order to compare phylogenetic trees to these models, I use these methods to calculate two statistics that describe the fit of a set of branching times to any time variable birth-death model: the maximum likelihood, which can be compared to the distribution of the

maximum likelihood for a random sample of trees or to that the maximum likelihood of other birth-death models using the Akaike Information Criterion; and the Komolgorov-Smirnov test, which is based on the fact that the branching times should be independently and identically distributed under many time variable birth-death models. I also demonstrate two new methods for visualizing the distribution of branching times: the lineage through time null plot uses a heat map to show the distribution of the number of lineages at different times; and the waiting time null plot does the same for waiting times between branching times. These plots can be used either to see how different time variable birth-death processes affect these distributions or to compare a data set to any time variable birth-death process. I use all these methods to analyze two data sets of reconstructed taxon branching times.

The study of paleopolyploidies requires the comparison of multiple whole genome sequences. If researchers could identify the branch of a phylogeny on which a whole genome duplication occurred, before sequencing the genomes of multiple taxa, then they could select taxa that would give them a better picture of that whole genome duplication. In chapter 3 I describe a likelihood model in which the number of chromosomes in a genome evolves according to a Markov process with three stochastic rates: a rate of chromosome duplication and a rate of chromosome loss that are proportional to the number of chromosomes in the genome; and a rate of whole genome duplication that is constant. I implemented software that calculates the maximum likelihood under this model for a phylogeny of taxa in which the chromosome counts are known. I compared the maximum likelihoods of a model in which the genome duplication rate varies to one in which it is fixed at zero using the Akaike information criterion, in order to determine if a model with whole genome duplications is a good fit for the data. Once it has been determined that the data does fit the model, we infer the phylogenetic position of paleopolyploidies by using this model to calculate the posterior probability that a whole genome duplication occurred on each branch of the taxon tree. I applied this model to a phylogeny of 125 molluscan taxa and inferred three places on that phylogeny where it is very likely that a whole genome duplication occurred: a single branch within the Hypsogastropoda; one of two branches at the base of the Stylommatophora; and one or two branches near the base of Cephalopoda.

Thanks to the wealth of readily available comparative genomic data, it has become apparent that gene family expansion and contraction is critical for the evolution of organisms. Several researchers have developed likelihood methods that use counts of genes in gene families from a number of taxa to deduce on which branches of the phylogenetic tree there has been an unusual amount of gene duplication or gene loss in that gene family. Gene family counts are readily available, but there is a great deal of information in the gene family tree that is unavailable when using gene counts alone. In chapter 4, I develop a method that uses the gene family tree to infer changes in the process of gene gain and loss on a taxonomic tree. This method relies on calculating the probability of a gene tree given a taxon tree and a set of birth-death parameters by which that gene tree evolves on the taxon tree. I use a reversible-jump MCMC to sample from the joint posterior distribution of a set of birth-death parameters and assignments of those parameters to the branches of a taxon tree given

a gene tree and a taxon tree. Different assignments are compared using Bayes factors. I use simulations to show that this method has much more power than a method which relies only on counts of gene family members to determine if a gene family evolved by a different process on a pair of taxon branches, and whether that difference is a consequence of differences in the birth rate or the death rate.

In section 4.5 I expand my method to include uncertainty in the gene tree topology, by using a set of gene alignments as my data rather than the fully resolved gene tree. Under this implementation I calculate the probability of those sequences given the gene tree, in addition to the probability of the gene tree given the taxon tree. I modify the reversible-jump MCMC so that it now samples from the posterior distribution of the nucleotide evolution parameters and the gene trees, in addition to the birth-death parameters and their assignments to the branches of the taxon tree. I demonstrate the use of this method on two real gene families found in the Bilateria. I found that a clade of 46 protein tyrosine kinase genes from three taxa is characterized by an increase in the gene duplication rate on the branch leading to *Caenorhabditis elegans*. Furthermore, a separate analysis of all the posterior hox genes from nine taxa implies that their evolution has been characterized by massive gene loss throughout the Bilateria with a lower rate of turn over in the chordates and at the base of the deuterostomes than is found in the protostomes or in the echinoderms.

Contents

Chapter 1

Variation in the Process of Lineage Gain and Loss

1.1 The Birth-Death Process

My obsession with the birth-death process began in the mid-90s. My roommate was in New York at the time, and so I spent a night alone drinking beer and watching the Highlander Director's Cut. I started wondering how the Highlander, who had only been around for a few hundred years, could possibly beat the bad ass Kurgan, who had been around for thousands. It struck me that actually the Kurgan's age was a severe disadvantage. If each of them had a constant stochastic rate of getting their head chopped off, then the Kurgan had a much higher probability of losing his head before 1985 than the Highlander did even if the Kurgan's stochastic death rate was much smaller, because the Kurgan had to survive for so much longer. I took this insight added a birth rate and spent that night and the following weekend deriving a number of results for the constant rate birth-death process.

The birth-death process is a continuous time Markov process with two stochastic rates birth and death that describe the size dynamics of a population. Each individual in a population has an instantaneous probability of producing another individual, λ, and an instantaneous probability of dying, μ. Once an individual dies it is lost forever. Therefore, the stochastic rate for a population to gain or lose an individual is directly proportional to the number of individuals in the population, and once a population reaches zero, it is extinct forever. It is not difficult to derive many basic results from this process, such as: the probability of a single individual producing N individuals after time t; the probability of a population going extinct after time t, even as t approaches infinity; the expected number of individuals after time t in a population starting with N individuals; the probability of an individual reproducing M times before they die; and many more.

Of course most of this work had already been done by the time my parents were born. Yule (1925) first introduced a continuous rate pure birth model, in which μ is zero and λ

is constant, and calculated the probabilities of a single lineage producing n lineages at time t and the expected number of lineages at time t among several other results. Feller (1939) added lineage loss at a constant rate to the process, and calculated the expectation under the birth-death process and the probability of n lineages leaving N descendants at time t under the pure birth process. Finally, Kendall (1948) derived basic results for any birth-death process in which the values of λ and μ varied with time, including the distribution of the number of lineages descended from a single lineage at time t, the probability of a lineage being lost by time t, the distribution of the amount of time before a lineage is lost, and the distribution of the total number of births that have taken place.

The birth-death process makes sense as a model for many different types of biological lineages. The fundamental principal of the birth-death process is that at each instant each lineage has a probability of reproducing and a probability of dying that is independent of every other lineage. That is intuitively appealing for the literal birth and death of individuals in populations, for the speciation and extinction of species, for the duplication and loss of genes and for a plethora of other biological systems. Although it is certainly untrue that the probability of birth or death is the same for all lineages at all times, the model is still useful. The constant rate birth-death process can be used as a null hypothethis against which to test for differences between lineages in their propensity to diverge and disappear. Furthermore, it is possible to construct an unending number of birth-death processes in which the birth and death rates are not constant in order to determine what possible changes in the process of diversification could account for a set of observations.

1.1.1 The Birth-Death Process and Macroevolution

From its beginning the birth-death process was used to model the diversity of taxa. Yule (1925) first introduced the pure birth process in order to explain the distribution of numbers of species among genera. He found that the data did in fact fit the model very well. Raup et al. (1973) reinvigorated the use of the birth-death process in biology, when they used it as a null model for the diversity of taxa through time, so that it could be contrasted with theories that sought specific mechanisms to explain fluctuations in diversity. They found that changes in the number of fossil lineages within clades did not appear to differ from what one would expect under this null model.

Thompson (1975) expanded the probabilities that could be calculated under the birth-death process from those simply involving counts of lineages to the probabilities of full phylogenies. Among other results, he showed how to calculate the density of a phylogenetic tree with branching times. This was shown to be a reasonable approach by Nee et al. (1992), when they compared multiple distributions to the branching times of a real bird phylogeny and concluded the birth-death process was the best. Nee et al. (1994b) made a major break through with the introduction of the reconstructed birth-death process. The reconstructed birth-death process assumes that all the lineages observed survived to the present, and thus allows for the analysis of molecular phylogenies, which include only extant

taxa. Nee et al. (1994a,b) showed how to analyze real phylogenies under the reconstructed birth-death process using lineage through time plots and estimate the birth-death parameters by maximum likelihood. In addition to being used to analyze already derived phylogenies, the birth-death process has also been used as a prior in the inference of ultrametric tree topologies and branching times (Rannala and Yang 1996; Huelsenbeck et al. 2002).

1.1.2 The Birth-Death Process and Gene Family Evolution

The birth-death process has been employed not only to explain taxonomic diversity but also to explain the diversity of gene sequences. Early in the study of gene evolution it was believed that genes evolved through a process of duplication and slow divergence, but starting in the 1970s the paradigm of concerted evolution came to dominate the field of molecular evolution (see Nei and Rooney 2005). However, a slew of phylogenetic studies since the 90s indicated that in fact most genes evolve via a branching process similar to taxon lineages (Ota and Nei 1994; Nei et al. 1997; Annilo et al. 2006). Several loci that were previously thought to evolve by concerted evolution turned out instead to just be evolving very slowly (Nei et al. 2000; Piontkivska et al. 2002). This implied that the birth-death process would be an appropriate model for gene diversification, and Karev et al. (2002) showed that it did a good job of explaining the power law distribution of domain family size. Furthermore, Reed and Hughes (2004) showed that a birth-death process with exponentially distributed time of origin could also explain the power law distribution of gene family size.

Since these initial suggestions that the birth-death process might be useful in analyzing gene family evolution, there has been an explosion of studies and published methods that do just that. Lynch and Conery (2003) used the birth-death process to estimates rates of gene duplication and turn over from gene counts and concluded that gene family size is based on a steady state process. Gu and Zhang (2004) devised a birth-death model of gene family size to calculate the distance between pairs of genomes based on gene family content and used it to reconstruct a tree of life by neighbor joining. Cotton and Page (2005) estimated birth-death parameters from trees of human gene families using methods similar to those found in Nee et al. (1994b). Hahn et al. (2005) showed how to use the birth-death process to infer branches of a taxon tree on which an exceptionally large number of gene duplications occurred by analyzing the number of gene family members in different taxa. Csűrös and Miklós (2006) described and implemented a model that included stochastic rates for birth, death and horizontal gene transfer and could reconstruct the gene content at the nodes of a taxon tree based on the gene content of the taxa represented by the tips of that tree. Cohen and Pupko (2010) also used a model that included birth, death and horizontal transfer to infer the posterior probability of the number of gains and losses along a branch of the taxon tree from the number of genes at the tips. Iwasaki and Takagi (2007) used a similar model to reconstruct ancestral gene content but did not include horizontal transfer and allowed the birth-death rates to vary between the branches of the taxon tree. Arvestad et al. (2003, 2009) showed how to calculate the probability of a fully resolved gene tree or a particular

reconciliation of a gene tree given a taxon tree and a set of birth-death parameters, and used that probability as a prior in order to infer a gene tree (Åkerborg et al. 2009).

1.2 Variation in Evolutionary Rates

I came to be obsessed with varying evolutionary rates through more standard academic pathways. As an undergraduate, I simply could not believe that evolutionary processes operated the same during adaptive radiations as they do in the descendants of those radiations. As my intellectual world expanded I learned about a growing literature focused on inferring changes in the rate of character evolution. Most of these methods work by comparing a model in which a set of stochastic rates describing the evolution of a character are the same throughout a taxonomic tree to a model in which those rates differ between one or more clade and the rest of the tree (e.g. Pagel 1994; O'Meara et al. 2006). In retrospect it seems inevitable that these two obsessions should merge in my mind, as it is apparent that these types of methods could be extended to incorporate variation in the birth-death process. Indeed many researchers are already working on doing just that.

1.2.1 Variable Rates of Taxon Diversification

There is a great deal of evidence that the process of diversification varies among taxon clades. Many but not all are based on comparing the balance of real phylogenies to a random branching null model, of which the constant rate birth-death process is a special case. In the initial study to demonstrate this fact Guyer and Slowinski (1991) looked at the frequency of each of the three possible unlabeled topologies for a five species clade in three large cladograms and found that the most imbalanced topology occurred much more than we would expect under a random branching process. Since then several more complete studies have been done. Savolainen et al. (2002) found that for a large number of molecular phylogenetic trees the branches immediately descended from the shorter of two sister branches were shorter than the branches descended from the longer of those two sister branches, implying that the time between speciations was in fact heritable. Blum and Francois (2006) showed that the values for measures of imbalance found in trees from Treebase were significantly higher than the values found under a random branching process. Heard (1996) ran simulations in which the speciation rate varied in a heritable manor and found that simulations with a higher rate of speciation rate evolution did in fact produce more imbalanced trees.

A number of methods have now been developed that rely on the birth-death process to detect clades with abnormal diversity. Magallon and Sanderson (2001) calculated a value of λ for a whole clade assuming that μ was zero, and then attempted to identify especially large or small monophyletic groups within that clade by calculating the two tailed probability that each subclade would be as large or small as it is given that value for λ. Sims and McConway

(2003) and McConway and Sims (2004) contrasted the diversity in pairs of sister clades by comparing the χ^2 distribution to the maximum likelihood ratio between a model in which both clades had different birth-death parameters and a model in which those parameters were the same. Moore et al. (2004) recognized that a clade could appear more diverse than its sister clade because of a shift of diversification processes within that clade rather than at its base, so for any clade they calculated the same likelihood ratio as Sims and McConway (2003) for it and its sister clade and for the two basal clades within it. They then calculated the difference between these ratios for every clade in a phylogeny and compared it to the distribution of that statistic on a set of randomly generated trees under a pure birth process to determine if any clades were diversifying at an abnormal rate.

Another set of methods use modifications of the birth-death process to infer correlations between the state of a biological character and the process of diversification. Ree (2005) calculated the correlation between a binary trait and the branching process by simulating multiple random assignments of states to the branches of the tree, averaging the maximum likelihood estimates of λ under a pure birth process over those simulations, and then comparing the difference between these rates to the same statistic averaged over the same tree topology with branch lengths randomly assigned from a pure birth process. Maddison et al. (2007) showed how to calculate the probability of a phylogeny and the distribution of a binary character on the tips of that phylogeny given that λ and μ are dependent on the state of that character and compared the maximum likelihood calculated under this model to a maximum likelihood in which the birth-death rates are constant throughout the tree in order to determine if the character affected the branching pattern. Paradis (2005) calculated the correlation between the branching pattern and a continuous character by first determining the maximum likelihood reconstruction of the character on the phylogeny under Brownian motion and then calculating the probability of the phylogeny under a pure birth process in which $\lambda/(1-\lambda)$ was linearly related to the value of the character. FitzJohn (2010) generalized the calculation of the probability of any number of continuous characters and a phylogeny given that the birth-death parameters are any function of the character state and time and that the character evolves by any diffusion process. He used this to detect correlations between a character and the diversification process by comparing the likelihood ratio of any pair of nested models to the χ^2 distribution.

The process of taxon diversification varies not just along the branches of a phylogeny, but also varies as a function of time. Changes in diversification with time may be a result of underlying changes in a subclade of the studied taxa the effects of which can be seen when studying the whole group, interactions among lineages such as competition that may limit diversification as the number of species increases, environmental changes that can affect every member of a clade at once, or many other processes that I have yet to think of. This has been shown to be true for fossil Bilateria by Foote (1993), who looked at the fossil history of Blastoidea and several trilobite clades and found that diversification rates tended to be high early and then decrease with time. Using only extant species counts, Strathmann and Slatkin (1983) showed that the distribution of the number of species among phyla can only

be explained by models in which the birth-death parameters vary through time or with the number of taxa.

Two complementary methods compare the distribution of branching-times to a constant rate birth-death process in order to infer temporal variation in diversification. Nee et al. (1994b) introduced the lineage through time plot in which the log of the number of lineages is plotted against time. The shape of this curve can be compared to what we would expect to see under a constant rate birth-death process, to see if the data fits the assumptions. Pybus and Harvey (2000) described a statistic, γ, which, assuming that there is no extinction, should be greater than zero if λ increases towards the present and less than zero if it decreases. The significance of a particular value of γ can be inferred by determining the distribution of that statistic on a large number of simulated trees. Many researchers have used these methods to demonstrate that diversification has varied with time for a number of diverse lineages (e.g. Purvis et al. 1995; Harmon et al. 2003; Shaw et al. 2003; Kadereit et al. 2004; Rüber and Zardoya 2005; Turgeon et al. 2005).

Several methods have been developed in which the maximum likelihood of a set of data is calculated under a constant rate birth death process and under a birth death process that varies with time, and these likelihoods are then compared by the Akaike Information Criterion (Akaike 1974). Paradis (1997) calculated the likelihoods of a set of branching times under a constant rate pure birth model and two different time variable pure birth models. He provided analytical solutions for the maximum likelihoods, but he made a mathematical error which was corrected by Nee (2001). Rabosky (2006a) introduced a group of methods in which the probability of a set of branching times is calculated under a constant rate birth death model and under a model in which the birth-death parameters vary at specific times. He also made an error in calculating this second probability which is corrected in chapter 2. Rabosky (2006b) and Rabosky and Lovette (2008a,b) have calculated the likelihoods for a number of other time variable birth-death models, which allows for the testing of specific hypothethes about the origins of temporal changes in the diversification process. Several studies using these methods have shown that diversification rates do vary with time (Dolman and Hugall 2008; Steeman et al. 2009; Burbrink and Pyron 2010; Valente et al. 2010).

1.2.2 Variable Rates of Gene Family Diversification

There is also substantial evidence that the diversification process for gene families is not uniform. Karev et al. (2003, 2004) showed that the distribution of domain family and gene family sizes were best described by a model in which the rate of gene family growth was not directly proportional to the number of genes, but instead varied with the square of the number of family members. This implies large gene families expand at a higher per lineage rate than small gene families, which the authors suggested may be a consequence of differential selection pressures acting on gene family size. Furthermore, Cotton and Page (2006) demonstrated that human gene family trees are more imbalanced than taxon trees or what we would expect under a constant rate birth-death process, implying that the rates of

gene family diversification evolve on the gene tree. Cotton and Page (2005) also concluded that the diversification process for gene families has not been constant through time.

The diversification of gene families, unlike taxa, can also differ between the branches of the taxon tree on which they evolve, and there is much evidence that they do so. Lynch and Conery (2003) use a model based on the birth-death process to estimate diversification rates from gene counts in a number of taxa, and identified significant interspecies differences in rates. Iwasaki and Takagi (2007) used a birth-death model of gene counts with rate variation between branches of the taxon tree to show that rates differ between lineages throughout the tree of life. Using the method described by Hahn et al. (2005) multiple studies have shown that there is a great deal of variation in the rate of gene turn over between taxon lineages and in rates of expansion and loss between different gene families, that are taxon lineage specific (see Demuth and Hahn 2009).

1.3 Summary of the Chapters

In this paper I develop three different methods to detect changes in the rate of the gain and loss of biological elements using the birth death process. These methods focus on different types of systems - taxa, chromosomes and gene families - different causes of variation - time, genome duplication, branches of the taxon tree - and different ways of interpreting the results - visualizations, frequentist and Bayesian. However, all revolve around calculating the probability of some data under different implementations of a birth-death process to compare models, and all overlap to some degree in the equations that are used to make those calculations.

In chapter 2, I generalize the reconstructed birth-death process to account for any time variable set of birth-death parameters and any assumptions about the number of taxa at any time. I show how to calculate the distributions of numbers of taxa and waiting times under this process and the density of a set of branching times. I also introduce a simple numerical solution to any time variable process that allows one to incorporate sampling and mass extinctions. It is easy to calculate the inverse of these probabilities using this numerical solution, and thus find quantiles and take random samples.

In the latter part of chapter 2 I introduce several methods to determine how well a set of branching times fits any time variable birth-death model. The distribution of lineages through time or waiting times through time can be plotted against their expected quantiles through time for any time variable distribution, which allows one to readily recognize violations of a model. I also implement the calculation of the maximum likelihood under any time variable process, and echo the method of Rabosky (2006a), in which models are compared using the AIC. Finally, I compare a real data set to a number of models using all these methods and conclude that this sort of model fitting should only be done in the context of hypothethis testing, as multiple crazy models can all fit the data well.

In chapter 3 I model the evolution of chromosome numbers on the branches of a taxonomic

tree using the birth-death process in order to infer the phylogenetic location of whole genome duplications. Rather than considering cases in which the birth-death parameters take on different values at different times, I introduce an additional parameter, the stochastic rate of genome doubling. Under this model I assume that there is an instantaneous probability that every chromosome in a genome would duplicate at once. I use a maximum likelihood implementation of this model to analyze the evolution of chromosome numbers on a tree of 125 molluscan taxa, and conclude that it is a much better fit for the data than a model in which the stochastic rate of genome doubling is zero. I then calculate the posterior probability of genome doubling on each branch and infer that there are three branches where it is very likely that a paleopolyploidy occurred.

In chapter 4 I use a reversible-jump Markov Chain Monte Carlo (MCMC) method to determine if the birth-death process for gene family evolution differs between the branches of the taxon tree. Unlike all previous methods I consider not just the counts of gene family members in the extant taxa, but instead consider the whole gene tree. I calculate the probability of a gene tree based on a taxon tree using the method of Arvestad et al. (2003, 2009) and sum over all possible sets of birth-death parameters and estimate Bayes factors with the MCMC. In section 4.2 I rederive the probability of the gene tree using equations from chapter 2, and in section 4.3 I describe the implementation of the reversible jump MCMC. In section 4.4 I use simulations to compare my method to one which only considers gene counts and conclude that mine has much greater power both to detect differences and distinguish between changes in the rate of gene duplication and changes in the rate of gene loss, at least when the true gene tree topology is known.

In section 4.5 I use my method to analyze the evolution of two real gene families, a clade of protein tyrosine kinase genes (PTK) and the bilaterian posterior hox. Since for any real gene family the true gene tree is unknown, I sum over the uncertainty in the gene tree topology by having the MCMC search among the birth-death parameters, the gene tree and the gene sequence evolution parameters based on the probability of the gene sequence alignments given those parameters and the taxon tree. So that my target distribution is the joint posterior density of all those parameters given the gene sequence and the taxon tree. I conclude that the structure of the PTK gene family tree is a consequence of an increase in the duplication rate on the lineage leading to *Caenorhabditis elegans*. On the other hand, the posterior hox gene tree is characterized by gene loss throughout the Bilateria, although the rate of turn over is lower in the chordates and at the base of the deuterostomes than it is in the protostomes or the echinoderms.

Enjoy!

Chapter 2

The Reconstructed Time Variable Birth-Death Process

2.1 Introduction

In recent years, interest in using phylogenies of extant taxa to infer macroevolutionary patterns has grown to fill a niche created by an explosion in the number of molecular phylogenies. In the past such questions were the strict purview of paleontologists, but the development of statistical methods that rely on the shapes of phylogenies of extant taxa has allowed neontologists to investigate macroevolution as well. There are currently methods available that allow researchers to ask whether diversification rates differ between clades (e.g. Agapow and Purvis 2002; McConway and Sims 2004; Moore and Donoghue 2007), are correlated with biological characters (e.g. Maddison et al. 2007; Paradis 2005), or vary over time (e.g. Nee et al. 1994b; Rabosky 2006a). This paper is primarily concerned with the last of these question.

Most of these methods for investigating tree shape rely on the birth-death process as a null model of lineage diversification. The birth-death process is a venerable stochastic process in which each lineage splits in two at some probabilistic rate and dies at some probabilistic rate (Kendall 1948). If one assumes that these rates do not vary, then it serves as a reasonable null model to test for changes in the rate of diversification. This model was first applied to macroevolutionary data by Yule (1925), who showed that a process with no extinction was a good fit for the distribution of species among genera. Raup et al. (1973) used a constant rate birth-death model in order to show that the fluctuations over time in the number of fossil lineages within clades did not depart from a null expectation. Nee et al. (1994b) applied the birth-death process to phylogenies of extant taxa by introducing the reconstructed birth-death process and the lineage through time plot. They demonstrated how to compare branching time data to a birth-death model using both maximum likelihood and visual inspection.

Since this original work, methods that test the fit of a set of branching times to a birth-death process using a single statistic have proliferated. Paradis (1997) gave a method based on survival models for detecting changes in the diversification rate, assuming there is no extinction; he made a mathematical error that was corrected by Nee (2001). Pybus and Harvey (2000) originated a test for the fit of branching times to a pure birth process based on the closeness of the nodes to the root. This method can not only determine if a data set fits a pure birth process but can also distinguish whether the branching times occur unexpectedly early or late. Rabosky (2006a,b) and Rabosky and Lovette (2008b) provided methods for calculating the likelihood of a set of branching times under a number of different time variable birth-death models, and suggest that the fit of these models to the data should be compared using the Akaike Information Criterion (Akaike 1974).

Lineage through time plots in which the log of the number of reconstructed lineages is plotted against time have also become quite popular. These plots are usually just compared to a plot of the expected number of lineages over time under some birth-death process. Deviance from the expectations are described without any reference to the quantiles of the distribution. Rabosky and Lovette (2008a) and Crisp and Cook (2009) have visualized the distribution of the number of lineages through time for several time variable birth-death processes by simulating many trees and plotting all their lineage through time plots together.

Mathematical work to describe the distributions of numbers of lineages and branching times under both the regular and the reconstructed birth-death process has also proliferated. Aldous and Popovic (2005) investigated the critical branching process, in which the origin of a clade is evenly distributed between the infinite past and the present, for a birth-death process in which the splitting rate and the loss rate are equal and constant. They calculated a number of different probabilities including the distributions of the number of lineages and the number of extinct lineages over time. Gernhard (2008a) calculated the distribution of branching times under the constant rate birth-death process. Stadler (2008) calculated the expectation of the nth reconstructed branching time given random taxon sampling and the distribution of the number of lineages over time for the constant rate birth-death process. All of these models assume that there were one or two reconstructed lineages at some time in the past or that there were some number of lineages in the present and that the rates of lineage gain and loss did not vary with time.

Here I make the reconstructed birth-death process general for any time variable birth-death process and any set of assumptions about how many reconstructed lineages there are. That is to say I characterize reconstructed processes for which one assumes that there are n reconstructed lineages at time t no matter what value n and t have. I show how to calculate the distribution of the number of reconstructed lineages over time, the distribution and density for the waiting times between branching events and the density of a set of branching times for all these processes. I also show how to simulate trees. Furthermore, I introduce the discrete time birth-death process which can serve as a practical and flexible numerical solution for the calculation of these probabilities and their inverses, and allows us to easily incorporate random sampling and mass extinctions.

I also introduce a pair of graphical methods for investigating the temporal distribution of branching times in a phylogeny for any birth-death process, which rely on the analytical solutions described in this paper. In the null lineage through time plot the two-tailed distribution of the number of lineages is plotted over time. This plot can be compared to an actual lineage through time plot or observed alone in order to see what effect different time variable birth-death processes have on the distribution of lineages through time. The second method, the null waiting times plot, involves the distribution of the time between branching times, which can be plotted against real data or investigated on its own.

I have developed the telos software package for the R statistical programming language (R Development Core Team 2010) based on the ape package (Paradis et al. 2004) in order to implement these calculations and plotting tools.

2.2 Time Variable Birth-Death Process

2.2.1 Definitions

The birth-death process is a continuous time Markov process that describes the growth of a group of lineages. At any given time under this process each lineage has a probability of splitting into two lineages, λ, and a probability of of being lost, μ. I will define the constant rate birth-death process (CRBD) as one in which λ and μ remain constant for the duration of the process. The Yule process is a special case of the CRBD in which μ is zero. This paper will concern itself mainly with the time variable birth-death process (TVBD) for which λ and μ can vary with time, but at any time they are the same for all lineages. This is equivalent to the generalized birth-death process introduced by Kendall (1948). The CRBD is a special case of the TVBD. For most of this paper I will explore the reconstructed time variable birth-death process (RTVBD), a modification of the TVBD in which we only concern ourselves with those lineages that survive to some time when they are observed - usually the present (Nee et al. 1994b). I will also demonstrate the utility of the discrete time birth-death process (DTBD), a special case of the TVBD in which time is divided into several intervals within which λ and μ are constant but between which λ and μ may vary. This paper does not deal with birth-death processes for which λ and μ differ between lineages alive at the same time.

We will begin by describing the TVBD in such a way that time is divided into several intervals. Let t_j be the amount of time before the present for any integer j, such that $t_0 = 0$, and $t_k > t_j > t_i > 0$ (Figure 2.1). We will define $\lambda(t_j)$ as the per lineage rate of lineage splitting at time t_j and $\mu(t_j)$ as the per lineage rate of lineage loss at time t_j . Furthermore, let N_i be the number of lineages at time t_i, n_j^i be the number of reconstructed lineages at time t_j that survive to time t_i, and n_j be the number of reconstructed lineages at time t_j that survive to the present, so that $n_i^i = N_i$ and $n_j^0 = n_j$ (Figure 2.1). It is apparent that $n_j^i \geq n_k^i$ and $n_k^j \geq n_k^i$.

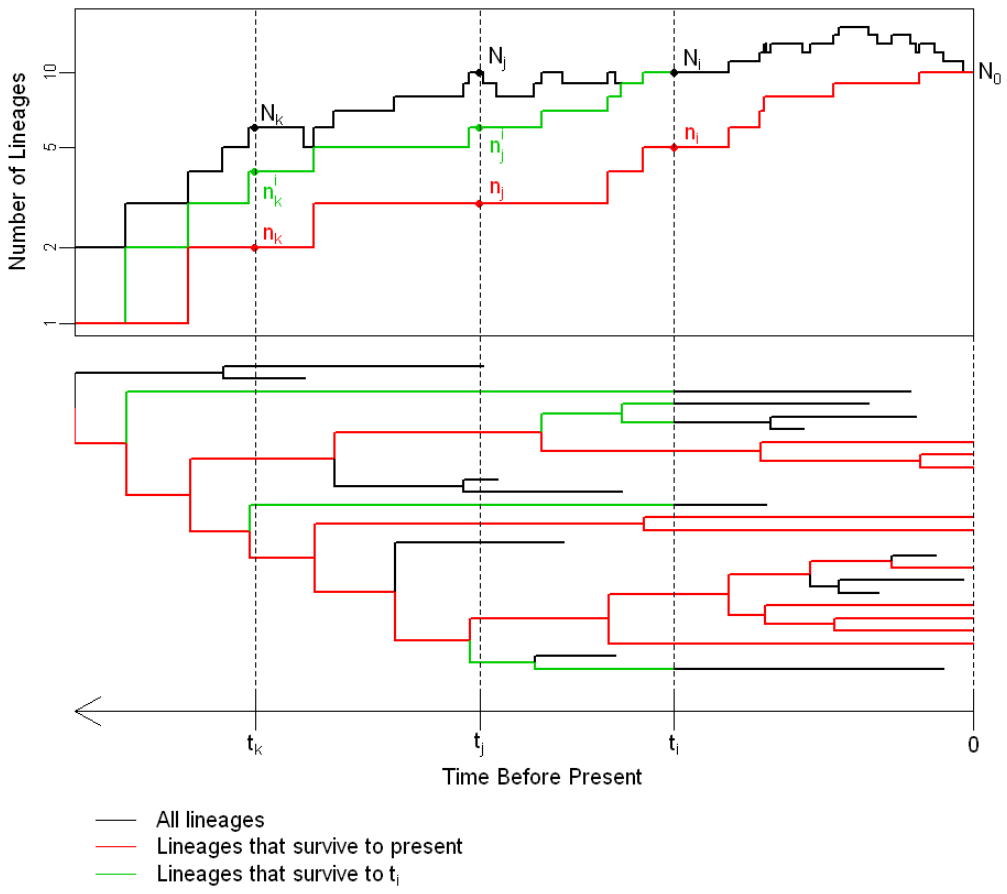

Figure 2.1: The definitions of variables for the time variable birth-death process. A random phylogenetic tree with ten terminal lineages is shown plotted on a time axis. Time is proceeding backwards such that larger times precede older times, and $t_k > t_j > t_i > 0$. Lineages that survive to the present are marked in red, those that survive to t_i but die before the present are marked in green and the rest are marked in black. The top plot shows the number of lineages in this clade that are alive at any time. The red line is the number of reconstructed lineages that survive to the present; the green line is the number of reconstructed lineages that survive to time t_i; and the black line is the total number of lineages. Values of N, the total number of lineages, n, the total number of lineages that survive to the present, and n^i, the number of lineages that survive to time t_i are marked at times t_k, t_j, t_i and 0.

Let $E_i(t_j)$ be the probability that one taxon at time t_j does not leave any descendants at time t_i. So that:

$$E_i(t_j) \equiv P(N_i = 0 | N_j = 1)$$

We can now establish the fundamental relationship between N_j and n_j^i by recognizing that the only way for there to be n_j^i lineages alive at time t_j that survived to time t_i if there were N_j lineages at time t_j, is if $N_j - n_j^i$ of them were lost by time t_i and n_j^i survived.

$$P(n_j^i | N_j) = \binom{N_j}{n_j^i} (1 - E_i(t_j))^{n_j^i} (E_i(t_j))^{N_j - n_j^i} \tag{2.1}$$

Let $B_i(t_j)$ be the probability that a single lineage at time t_j which survives to time t_i leaves multiple descendant lineages at time t_i.

$$B_i(t_j) \equiv P(N_i > 1 | n_j^i = 1)$$

We will also define $B_i(t_k, t_j)$ as the probability that a single lineage at time t_k that survives to time t_i will leave more than one lineage at time t_j that survive to time t_i.

$$B_i(t_k, t_j) \equiv P(n_j^i > 1 | n_k^i = 1)$$

We can clearly see that $B_i(t_k, t_k) = 0$, $B_i(t_i) = 0$, $B_i(t_k, t_i) = B_i(t_k)$ and that:

$$\begin{aligned} 1 - B_i(t_k, t_j) &= P(n_j^i = 1 | n_k^i = 1) \\ &= \frac{P(N_i = 1 | n_k^i = 1)}{P(N_i = 1 | n_j^i = 1)} \\ &= \frac{1 - B_i(t_k)}{1 - B_i(t_j)} \end{aligned} \tag{2.2}$$

We know from Kendall (1948, eq. 8) that, when N_i is greater than zero:

$$P(N_i | N_j = 1) = (1 - E_i(t_j))(1 - \eta_{t_j})(\eta_{t_j})^{N_i - 1}$$

for some function η_{t_j}. It is easy to see that η_{t_j} is equivalent to $B_i(t_j)$ by comparing the probability of one lineage at time t_j leaving a single lineage at time t_i to the probability of a single reconstructed lineage at time t_j leaving a single lineage at time t_i.

$$\begin{aligned} B_i(t_j) &= 1 - P(N_i = 1 | n_j^i = 1) \\ &= 1 - \frac{P(N_i = 1 | N_j = 1)}{P(n_j^i = 1 | N_j = 1)} \\ &= 1 - \frac{(1 - E_i(t_j))(1 - \eta_{t_j})}{1 - E_i(t_j)} \\ &= \eta_{t_j} \end{aligned} \tag{2.3}$$

Therefore within our context the probability mass of N_i lineages is best described as:

$$P(N_i|N_j=1) = (1-E_i(t_j))(1-B_i(t_j))(B_i(t_j))^{N_i-1} \tag{2.4}$$

Kendall (1948) derived a series of equations for the birth death process when λ and μ are constant that will apply under the CRBD.

$$E_i(t_j) = au(t_j - t_i) \tag{2.5}$$

$$B_i(t_j) = u(t_j - t_i) \tag{2.6}$$

where

$$u(t) \equiv \frac{\exp(rt)-1}{\exp(rt) - a} \tag{2.7}$$

when λ does not equal μ, and

$$u(t) \equiv \frac{\lambda t}{\lambda t + 1} \tag{2.8}$$

when it does, $r \equiv \lambda - \mu$ and $a \equiv \mu/\lambda$. This is a common reparameterization in which a is unitless and amounts to a shape parameter, while the absolute value of r is essentially a scaling parameter.

2.2.2 The Birth-Death Process Divided into Time Intervals

Kendall (1948) gives general solutions for $B_i(t_k)$ and $E_i(t_k)$ for the TVBD that are dependent on the values of λ and μ and involve integrating a complicated equation. Here we will derive a set of equations for $B_i(t_k)$ and $E_i(t_k)$ for any TVBD by dividing time into two intervals at t_j. These equations depend only on the values of B and E within those intervals, $B_i(t_j)$, $B_j(t_k)$, $E_i(t_j)$ and $E_j(t_k)$, without regard to what the values of λ and μ are within those intervals.

We can calculate $E_i(t_k)$ for this process by recognizing that a lineage alive at t_k will leave no descendant lineages at time t_i so long as none of its descendants at time t_j have any descendants of their own at time t_i.

$$\begin{aligned} E_i(t_k) &= \sum_{N_j=0}^{\infty} P(N_j|N_k=1)P(N_i=0|N_j) \\ &= E_j(t_k) + \sum_{N_j=1}^{\infty} (1-E_j(t_k))(1-B_j(t_k))(B_j(t_k))^{N_j-1}(E_i(t_j))^{N_j} \\ &= E_j(t_k) + \frac{E_i(t_j)(1-E_j(t_k))(1-B_j(t_k))}{1-B_j(t_k)E_i(t_j)} \end{aligned} \tag{2.9}$$

With some rearranging this yields the generally useful relationship:

$$1-E_i(t_k) = \frac{(1-E_j(t_k))(1-E_i(t_j))}{1-B_j(t_k)E_i(t_j)} \tag{2.10}$$

In order to derive a general equation for $B_i(t_k)$ that is divided into two periods we must first derive a function for $B_i(t_k, t_j)$. We can calculate this probability by summing over all possible values of N_j and substituting in (2.4) and (2.1), and then using (2.10) to simplify the equation.

$$\begin{aligned}
1 - B_i(t_k, t_j) &= P(n_j^i = 1 | n_k^i = 1) \\
&= \frac{\sum_{N_j=1}^{\infty} P(n_j^i = 1 | N_j) P(N_j | N_k = 1)}{P(n_k^i = 1 | N_k = 1)} \\
&= (1 - B_j(t_k)) \frac{(1 - E_i(t_j))(1 - E_j(t_k))}{1 - E_i(t_k)} \sum_{N_j=1}^{\infty} N_j (E_i(t_j) B_j(t_k))^{N_j - 1} \\
&= \frac{(1 - B_j(t_k))}{1 - B_j(t_k) E_i(t_j)} \qquad (2.11)
\end{aligned}$$

We can now calculate $B_i(t_k)$ by rearranging (2.2) and then substituting in (2.11).

$$1 - B_i(t_k) = \frac{(1 - B_j(t_k))(1 - B_i(t_j))}{1 - B_j(t_k) E_i(t_j)} \qquad (2.12)$$

2.2.3 Sampling and Mass Extinctions

Sampling can have a large effect on the distribution of branching times in a phylogeny (Slatkin and Hudson 1991; Cusimano and Renner 2010). Furthermore, it is likely that almost all real phylogenies represent a sample of the extant diversity for a clade. Therefore we should account for the effects of random sampling in the estimation of the distribution of branching times. We can easily calculate this using the methods we have already established by imagining that there is a very brief discrete period starting at time $t_0^<$ immediately before t_0 during which the chance of lineage splitting is zero, so that $B_0(t_0^<) = 0$ and $E_0(t_0^<) = 1 - p$, where p is the probability of an extant lineage being sampled. In this way we model sampling as if the unsampled lineages died immediately before the present (Nee et al. 1994b). Under these circumstances we can use (2.12) and (2.10) to find the values of $B_0(t_j)$ and $E_0(t_j)$ respectively.

$$1 - B_0(t_j) = \frac{1 - B_0^<(t_j)}{1 - (1 - p) B_0^<(t_j)} \qquad (2.13)$$

$$1 - E_0(t_j) = \frac{p(1 - E_0^<(t_j))}{1 - (1 - p) B_0^<(t_j)} \qquad (2.14)$$

where $B_0^<(t_j)$ is the probability of one reconstructed lineage at time t_j leaving more that one lineage in the present before sampling, and $E_0^<(t_j)$ is the probability of one lineage at time t_j leaving no lineages in the present before sampling. These values can be calculated as we

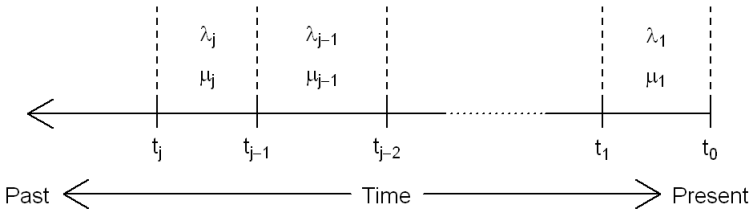

Figure 2.2: The definitions of variables for the discrete time birth-death process. For any two times t_j and t_{j-1}, $t_{j-1} < t_j$ and the values of the lineage splitting rate and the lineage loss rate between them are constant and referred to as λ_j and μ_j respectively.

would calculate $B_0(t_j)$ and $E_0(t_j)$ if there were no sampling. Special cases of these same equations were derived by Yang and Rannala (1997) and Stadler (2010) for the CRBD.

A similar method can be used to study the effects of mass extinctions on the distribution of reconstructed lineages and waiting times. Essentially, mass extinctions are sampling events that happened at some time in the past. If a mass extinction happened at time t_j, then we can imagine that it happened over a very brief period between $t_j^<$ and $t_j^>$, such that $B_j^>(t_j^<) = 0$ and $E_j^>(t_j^<) = 1-p$, where p is the probability of a lineage surviving the mass extinction. Therefore we can calculate $B_i(t_k)$ and $E_i(t_k)$ using (2.12) and (2.10) respectively.

$$1 - B_i(t_k) = \frac{(1 - B_j^<(t_k))(1 - B_i(t_j^>))}{1 - B_j^<(t_k)(1 - p(1 - E_i(t_j^>)))} \tag{2.15}$$

$$1 - E_i(t_k) = \frac{p(1 - E_j^<(t_k))(1 - E_i(t_j^>))}{1 - B_j^<(t_k)(1 - p(1 - E_i(t_j^>)))} \tag{2.16}$$

Once again $B_j^<(t_k)$, $B_i(t_j^>)$, $E_j^<(t_k)$ and $E_i(t_j^>)$ can each be calculated as you would calculate $B_j(t_k)$, $B_i(t_j)$, $E_j(t_k)$ and $E_i(t_j)$ if there were no mass extinction. In that case under the CRBD we get these relationships.

$$1 - B_i(t_k) = \frac{1 - u(t_k - t_i)}{p + (1-p)\frac{\exp(r(t_j - t_i)) + a}{\exp(r(t_k - t_i)) + a}} \tag{2.17}$$

$$1 - E_i(t_k) = \frac{p(1 - au(t_k - t_i))}{p + (1-p)\frac{\exp(r(t_j - t_i)) + a}{\exp(r(t_k - t_i)) + a}} \tag{2.18}$$

2.2.4 Discrete Time Birth-Death Process

I will now describe the DTBD, for which time will be divided into a series of time intervals, as in subsection 2.2.2, but now λ and μ will be constant within each interval. We will derive readily useful formulas for $B_i(t_j)$ and $E_i(t_j)$ that will apply under this process. Between any

times t_i and t_{i-1}, the birth rate and the death rate will not change and they will be referred to as λ_i and μ_i respectively (Figure 2.2). As λ and μ are constant within these intervals we can use (2.5) and (2.6) to derive functions that apply within this interval.

$$E_{i-1}(t_i) = a_i u_i \tag{2.19}$$

$$B_{i-1}(t_i) = u_i \tag{2.20}$$

where

$$u_i \equiv \frac{\exp(r_i(t_i - t_{i-1}))-1}{\exp(r_i(t_i - t_{i-1})) - a_i} \tag{2.21}$$

when λ_i does not equal μ_i, and

$$u_i \equiv \frac{\lambda_i(t_i - t_{i-1})}{\lambda_i(t_i - t_{i-1})+1} \tag{2.22}$$

when it does, $r_i \equiv \lambda_i - \mu_i$ and $a_i \equiv \mu_i/\lambda_i$.

We can use these equations and our previous results to derive a general equation for $E_i(t_k)$ under the DTBD that is dependent only on the values of λ and μ during the periods between t_k and t_i by separating the calculation of $E_i(t_k)$ into the period between t_{k-1} and t_k and the interval between t_{k-1} and t_i by using (2.10) and then substituting in (2.19) and (2.20).

$$\begin{aligned} 1-E_i(t_k) &= \frac{(1-E_i(t_{k-1}))(1-E_{k-1}(t_k))}{1-B_{k-1}(t_k)E_i(t_{k-1})} \\ &= (1-E_i(t_{k-1}))\frac{1-a_k u_k}{1-u_k E_i(t_{k-1})} \end{aligned} \tag{2.23}$$

We can then proceed by separating $E_i(t_{k-1})$ into two intervals in the same manor and continue with each subsequent interval until we have reached the last time period between t_i and t_{i+1}.

$$1-E_i(t_k) = \prod_{j=i+1}^{k} \frac{1-a_j u_j}{1-u_j E_i(t_{j-1})} \tag{2.24}$$

It is also possible to derive a general equation for $B_i(t_k)$ that is dependent only on the values of λ and μ during the periods between t_k and t_i by using (2.12) and substituting in (2.20).

$$\begin{aligned} 1-B_i(t_k) &= \frac{(1-B_i(t_{k-1}))(1-B_{k-1}(t_k))}{1-B_{k-1}(t_k)E_i(t_{k-1})} \\ &= (1-B_i(t_{k-1}))\frac{1-u_k}{1-u_k E_i(t_{k-1})} \end{aligned} \tag{2.25}$$

We can then proceed to subdivide each subsequent period in the same way as we did for $E_i(t_k)$.

$$1-B_i(t_k) = \prod_{j=i+1}^{k} \frac{1-u_j}{1-u_j E_i(t_{j-1})} \tag{2.26}$$

These simple equations can be used to examine models of lineage diversification in which the parameters of the birth death process changed at a specific time, or as a numerical solution to any TVBD. Kendall (1948) provided equations for $B_0(t_j)$ and $E_0(t_j)$ that would work for continuously varying values of λ and μ. Making these calculations requires solving two integrals. These integrals can be solved analytically for some TVBDs with continuously varying parameters, but for many others one must use numerical integration. The DTBD laid out here suggests an obvious alternative numerical solution that is not computationally burdensome. The period of time over which the parameters vary can be broken down into many discrete intervals. Appropriate values for λ and μ can be calculated for each of those intervals given the equations we have for those parameters. We can then treat each period as if λ and μ are constant for its duration and use (2.26) and (2.24) to calculate $B_0(t_j)$ and $E_0(t_j)$ respectively. This method has two advantages over Kendall's. First it fits into the framework that we have already established and thus does not require a great deal of additional analysis. Second it is easy to solve for t_j when given $B_0(t_j)$, a property which we put to use in calculating random branching times and quantiles for waiting times.

2.2.5 The Inverse of B_0 under the Discrete Time Birth-Death Process

Once we know $B_0(t_j)$ it is easy to calculate t_j under the DTBD. The first step is to calculate $B_0(t_i)$ for every time t_i at the beginning of each constant parameter time period. Then we should identify t_{j-1}, such that t_{j-1} is the beginning of the earliest constant parameter time period for which $B_0(t_{j-1}) < B_0(t_j)$. We now know that t_j must have occurred during the constant parameter time period before t_{j-1}, so that λ_j and μ_j are equal to those constants during that period. We can now solve (2.25) for u_j:

$$u_j = \frac{B_0(t_j, t_{j-1})}{1 - E_0(t_{j-1})(1 - B_0(t_j, t_{j-1}))} \tag{2.27}$$

and finally we can calculate t_j by solving (2.21) when λ_j does not equal μ_j:

$$t_j = \log\left(\frac{1 - a_j u_j}{1 - u_j}\right)/r_j + t_{j-1} \tag{2.28}$$

or by solving (2.22) when it does:

$$t_j = \left(\frac{u_j}{1 - u_j}\right)/\lambda_j + t_{j-1} \tag{2.29}$$

2.2.6 A general relationship between B_0 and E_0

We can use the DTBD to derive a general relationship between $B_k(t_i)$ and $E_k(t_i)$. Let us imagine that a TVBD operates between t_k and t_i, in that case we can divide that period

into q intervals of length $(t_k - t_i)/q$ and treat the process as a DTBD, so that $t_k = t_{i+q}$. In that case we find from (2.24) and (2.26) that:

$$\begin{aligned}\frac{1-E_k(t_i)}{1-B_k(t_i)} &= \prod_{j=i+1}^{i+q} \frac{1-a_j u_j}{1-u_j} \\ &= \exp\left(\sum_{j=i+1}^{i+q} r_j(t_j - t_{j-1})\right) \end{aligned} \tag{2.30}$$

We can see that as q approaches infinity.

$$\frac{1-E_k(t_i)}{1-B_k(t_i)} = \exp\left(\int_{t_i}^{t_k} r(t_j)dt_j\right) \tag{2.31}$$

Kendall (1948, eq. 13) provided this same equation and the agreement confirms our analysis.

2.2.7 The Derivatives of B_0 and E_0

We can also calculate the derivative of $B_0(t_j)$ by t_j for any TVBD by choosing some time t_{j-1} such that λ and μ do not change between t_j and t_{j-1}. Then we can take the derivative of (2.25).

$$\begin{aligned}\frac{\partial B_0(t_j)}{\partial t_j} &= \frac{(1-B_0(t_{j-1}))(1-E_0(t_{j-1}))}{(1-u_j E_0(t_{j-1}))^2}\frac{\partial u_j}{\partial t_j} \\ &= \frac{(1-B_0(t_{j-1}))(1-E_0(t_{j-1}))}{(1-u_j E_0(t_{j-1}))^2}\lambda_j(1-u_j)(1-a_j u_j) \\ &= \lambda(t_j)(1-E_0(t_j))(1-B_0(t_j)) \end{aligned} \tag{2.32}$$

We can also calculate the partial derivative of $B_0(t_k, t_j)$ by taking the derivative of (2.2).

$$\begin{aligned}\frac{\partial B_0(t_k,t_j)}{\partial t_k} &= \frac{1}{1-B_0(t_j)}\frac{\partial B_0(t_k)}{\partial t_k} \\ &= \lambda(t_k)(1-E_0(t_k))(1-B_0(t_k,t_j)) \end{aligned} \tag{2.33}$$

To calculate the derivative of $E_0(t_j)$ by t_j we can now simply take the derivative of (2.23).

$$\begin{aligned}\frac{\partial E_0(t_j)}{\partial t_j} &= \frac{(1-E_0(t_{j-1}))(a-E_0(t_{j-1}))}{(1-u_j E_0(t_{j-1}))^2}\frac{\partial u_j}{\partial t_j} \\ &= \lambda(t_j)(1-E_0(t_j))(a-E_0(t_j)) \end{aligned} \tag{2.34}$$

It is apparent that all these equations will hold true even as t_{j-1} approaches t_j, and thus they are appropriate for any TVBD. It should be noted that these derivatives are different from those provided by Kendall (1948) for $\partial B_0(t_j)/\partial t_0$ and $\partial E_0(t_j)/\partial t_0$, although they will be identical when λ and μ are constant. One interesting thing to note is that $E_0(t_j)$ will actually decrease as t_j increases if $E_0(t_j)$ is greater than a.

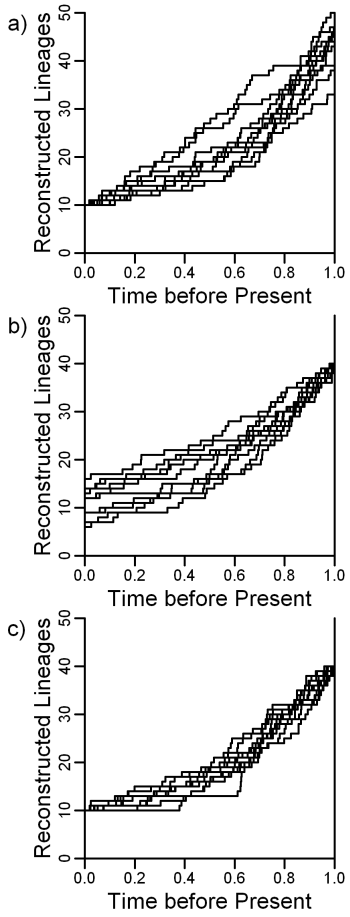

Figure 2.3: Random lineage through time plots under three different assumptions about the number of reconstructed lineages. a) Random lineage through time plots generated under Assumption 1 in which we assume that there are ten reconstructed lineages one time unit before the present. b) Random lineage through time plots generated under Assumption 2 in which we assume that there are forty lineages in the present. c) Random lineage through time plots generated under Assumption 3 in which we assume that there are ten reconstructed lineages one time unit before the present and forty lineages in the present.

2.3 Distribution of Reconstructed Lineages

In contrast to the regular birth-death process, the reconstructed birth-death process is based on the use of molecular phylogenies. Molecular phylogenies only contain lineages that have extant members. Thus for calculating the probabilities of the reconstructed process, we must assume that all the lineages we infer existed at any time, survived until the present. Nee

et al. (1994b) calculated the basic results for that process, formulas that allow us to calculate the probability mass of a number of lineages at some time in the past, given that there was one lineage at some time before then. Here I will produce some more general results that will allow one to calculate the probability mass and expectation for the number of reconstructed lineages at some time, given that there were some known number of lineages at any other time during the process.

I will calculate the probability of a given number of reconstructed lineages under three different sets of assumptions. Under Assumption 1 the number of reconstructed lineages is known at some time before the time we are concerned with (Figure 2.3a); under strict Assumption 1 we know the timing of the last common ancestor for a clade. Under Assumption 2 the number of reconstructed lineages is known at some time after the time we are concerned with (Figure 2.3b); under strict Assumption 2 we know the number of lineages alive in the present. Under Assumption 3 the number of reconstructed lineages is known both before and after the time we are concerned with (Figure 2.3c); under strict Assumption 3 we know the timing of the last common ancestor for a clade and the number of lineages alive in the present. Our results from the previous section in which the birth-death process was separated into different time intervals, make these calculations trivial.

2.3.1 Reconstructing Birth-Death from the Past

The simplest way to approach the birth-death process is to start at some time when you know the state of the process, and investigate how the process unfolds as we proceed forward in time. This was the approach taken by Kendall (1948) and Nee et al. (1994b), and it is the approach that I will begin with. Nee et al. (1994b) calculated the probability mass of the number of reconstructed lineages at t_j given that there was one reconstructed lineage at t_k. Here I will rederive that function and extend that result so that we can calculate the probability mass for a process that starts with any number of reconstructed lineages at t_k. To do this, I must first calculate $P(n_j|N_k=1)$, the probability that one lineage at time t_k will leave exactly n_j lineages at time t_j that survive to the present. I will sum over all the possible values of N_j, the number of lineages at time t_j that are descended from our single lineage at time t_k, substitute in (2.1) and (2.4) and use (2.11) to simplify the notation.

$$\begin{aligned}
P(n_j|N_k=1) &= \sum_{N_j=n_j}^{\infty} P(N_j|N_k=1)P(n_j|N_j) \\
&= (1-E_j(t_k))(1-B_j(t_k))(1-E_0(t_j))^{n_j} \sum_{N_j=0}^{\infty} \binom{N_j}{n_j} (B_j(t_k))^{N_j-1} (E_0(t_j))^{N_j-n_j} \\
&= \frac{(1-E_0(t_k))(1-B_j(t_k))}{1-B_j(t_k)E_0(t_j)} \left(1 - \frac{1-B_j(t_k)}{1-B_j(t_k)E_0(t_j)}\right)^{n_j-1} \\
&= (1-E_0(t_k))(1-B_0(t_k,t_j))(B_0(t_k,t_j))^{n_j-1} \qquad (2.35)
\end{aligned}$$

Now it is trivial to calculate $P(n_j|n_k=1)$, as it is the same as $P(n_j|N_k=1)$, except that it assumes that the lineage alive at t_k survives to t_0. I can then substitute in (2.1) and (2.35) to complete the derivation.

$$\begin{aligned} P(n_j|n_k=1) &= \frac{P(n_j|N_k=1)}{P(n_k=1|N_k=1)} \\ &= (1-B_0(t_k,t_j))(B_0(t_k,t_j))^{n_j-1} \end{aligned} \tag{2.36}$$

This is the same as equation 9 in Nee et al. (1994b), although I derived it in a different way.

We can make (2.36) more general by considering the case in which n_k is greater than 1. First we must introduce another term, S_{kj}, the set of all possible s, where s is an arrangement of the n_j reconstructed lineages among the n_k reconstructed lineages that they are descended from, such that s is an n_k-tuple, $(s_1, s_2, s_3, ...s_{n_k})$, and s_a is the number of reconstructed lineages at time t_j that descended from the ath of the initial n_k lineages. In that case $S_{kj} \equiv \{s:|s|=n_k, s_a \in \mathbb{N}^* \text{ and } \sum_{a=1}^{n_k} s_a = n_j\}$. In order to determine the size of S_{kj}, we must recognize that each of the initial n_k lineages has at least one descendant at t_j; therefore we need to ask how many ways the other n_j–n_k descendant lineages can be distributed among n_k ancestral lineages and $|S_{kj}| = \binom{n_j-1}{n_k-1}$. We can calculate the probability of any particular s, an arrangement of numbers of descendant lineages among the initial n_k, as the product of the probabilities that each of the initial n_k left s_a, the appropriate number of descendant lineages at t_j. We can then calculate $P(n_j|n_k)$ by summing these probabilities over all s in S_{kj} to account for all the possible arrangements.

$$\begin{aligned} P(n_j|n_k) &= \sum_{s\in S_{kj}} \prod_{a=1}^{n_k} P(n_j=s_a|n_k=1) \\ &= \sum_{s\in S_{kj}} \prod_{a=1}^{n_k} (1-B_0(t_k,t_j))(B_0(t_k,t_j))^{s_a-1} \\ &= \sum_{s\in S_{kj}} (1-B_0(t_k,t_j))^{n_k} (B_0(t_k,t_j))^{\sum_{a=1}^{n_k} s_a - n_k} \\ &= \binom{n_j-1}{n_k-1} (1-B_0(t_k,t_j))^{n_k} (B_0(t_k,t_j))^{n_j-n_k} \end{aligned} \tag{2.37}$$

This is a negative binomial distribution with n_j targeted successes and the probability of success being $B_0(t_k,t_j)$, so it is trivial to calculate the expectation and variance of n_j given n_k. Feller (1939) derived this same equation for the pure birth process.

2.3.2 Reconstructing Birth-Death from the Past and the Present

Next we will examine the case in which the number of lineages alive at t_i that survived to the present is known, and the number of lineages alive at time t_k that survived to the present is known, and we want to look at the distribution of lineages between those two times by

investigating the probability of having n_j reconstructed lineages at time t_j. We can calculate this value as the probability that n_k lineages leave n_j lineages and that n_j lineages leave n_i lineages given that n_k lineages leave n_i lineages. We can then substitute in (2.37) and use (2.2) to simplify the notation.

$$\begin{aligned} P(n_j|n_k,n_i) &= \frac{P(n_j|n_k)P(n_i|n_j)}{P(n_i|n_k)} \\ &= \binom{n_i-n_k}{n_j-n_k}\left(\frac{B_0(t_k)-B_0(t_j)}{B_0(t_k)-B_0(t_i)}\right)^{n_j-n_k}\left(\frac{B_0(t_j)-B_0(t_i)}{B_0(t_k)-B_0(t_i)}\right)^{n_i-n_j} \\ &= \binom{n_i-n_k}{n_j-n_k}\left(1-\frac{B_0(t_j,t_i)}{B_0(t_k,t_i)}\right)^{n_j-n_k}\left(\frac{B_0(t_j,t_i)}{B_0(t_k,t_i)}\right)^{n_i-n_j} \end{aligned} \tag{2.38}$$

It is interesting that this probability does not depend on the actual number of lineages, but instead on the change in the number of lineages since time t_k. The probability mass of this process is a binomial distribution with n_i–n_k trials, n_i–n_j successes and probability of success $B_0(t_j,t_i)/B_0(t_k,t_i)$ so it is easy to calculate the expectation and the variance. Stadler (2008) and Rannala (1997) both provided versions of this formula, but only for the CRBD and the specific cases in which $t_i=0$ and n_k is either one or two. Their results were the same as mine despite using very different derivations.

The probability of n_i, given n_k and n_j is dependent only on n_j, and thus the probability mass of this process can be calculated using (2.37). Furthermore the probability of n_k, given n_j and n_i, will depend only on n_j. I will now demonstrate how we can solve for this probability.

2.3.3 Reconstructing Birth-Death from the Present

Once we have observed some lineages at some time in the past that survived to the present it would be useful to know the probability that a given number of ancestral lineages were alive at some time before our observation (Figure 2.3b). We can calculate the probability of n_k given n_j, based on the probability of n_j given n_k.

$$P(n_k|n_j) = \frac{P(n_j|n_k)P(n_k)}{P(n_j)}$$

The problem is figuring out how to calculate $P(n_k)$, the prior probability of their being n_k reconstructed lineages at time t_k. Two approaches have commonly been used in the past. Under the first approach, the number of reconstructed lineages is considered to be one at some specific time in the past (Stadler 2008). This is simply a special case of Assumption 3 and the probability densities can be calculated using (2.38) and substituting 1 for n_k. However, this approach requires you to assume that there is one lineage at some time and that assumption leads to many problems. The second approach assumes that there is one

reconstructed lineage at some time in the past that is uniformly distributed between zero and infinity (Aldous and Popovic 2005; Gernhard 2008a; Stadler 2008). This relieves us of the assumption that there is one lineage at a specific time. However, although it is true that any extant clade being investigated has a single reconstructed ancestral lineage stretching back to infinity, or at least to the origin of life, this same lineage also has countless other lineages that have branched off it before the origin of the clade in question. Therefore we can not use the birth-death process to predict the number of lineages within our clade that are descended from this truly ancient lineage, because we have lost those earlier diverging clades by deciding to investigate only our clade, not through a random death process.

Another approach is to recognize that once our clade has diverged from its extant sister clade it will follow the birth-death process. Therefore, if we can calculate the probability of the sister clade splitting off at a given time we can use that to reconstruct the birth-death process from the present into the past. I will define v_1 as the time at which the extant sister clade diverged from the clade we are examining, after which our clade can be considered a single reconstructed lineage. The distribution of v_1 will vary with time, because in order for our lineage to split from its sister lineage two things must happen. First the lineages must split with probability $\lambda(v_1)$, which will vary with time. Second the sister lineage must survive to the present, which will happen with probability $1–E_0(v_1)$, assuming that the same parameters that govern our clade also govern its sister clade. If we assume that clades have split off from our lineage at this rate going back to the beginning of time, we can just take our sister clade as a random draw from all these clades. Therefore we can calculate the prior probability density v_1 at any particular time as:

$$f(v_1) = \frac{\lambda(v_1)(1-E_0(v_1))}{\int_0^\infty \lambda(t_j)(1-E_0(t_j))dt_j} \tag{2.39}$$

This is an improper prior unless $B_0(\infty)<1$, as it will not integrate to 1 (Berger 1980).

Once we have made this calculation we can calculate the probability density for v_1 assuming that there are a given number of taxa at some time t_j after v_1.

$$\begin{aligned} f(v_1|n_j) &= \frac{P(n_j|n^>_{v_1})f(v_1)}{\int_{t_j}^\infty P(n_j|n^>_{v_1})f(v_1)dv_1} \\ &= \frac{\lambda(v_1)(1-E_0(v_1))(1-B_0(v_1,t_j))(B_0(v_1,t_j))^{n_j-1}dv_1}{\int_{t_j}^\infty \lambda(v_1)(1-E_0(v_1))(1-B_0(v_1,t_j))(B_0(v_1,t_j))^{n_j-1}} \\ &= n_j\frac{(B_0(v_1,t_j))^{n_j-1}}{(B_0(\infty,t_j))^{n_j}}\frac{\partial B_0(v_1,t_j)}{\partial v_1} \end{aligned} \tag{2.40}$$

where $n^>_{v_1}$ is the number of reconstructed lineages immediately after v_1, which must by definition be one. The equation derived for the density of the time of origin by Gernhard

(2008a,b) is a special case of this equation in which λ and μ are constant, $t_j = 0$ and $B_0(\infty) = 1$, and equation 3 in Aldous and Popovic (2005) is a special case of this formula in which $\lambda = \mu = 1$ and $t_j = 0$. These authors assumed that the time at which there was one lineage, which may or may not have survived to the present, was uniformly distributed between the present and infinity, while I assumed that the time at which there was one reconstructed lineage was distributed according to the probability its sister lineage arising and surviving to the present. It is apparent from the calculations why these two assumptions should produce the same result.

We can also use this equation to calculate the cumulative distribution of v_1 given some number of reconstructed lineages at some time.

$$\begin{aligned} F(v_1|n_i) &= \int_{t_i}^{v_1} f(v_1 = t_j|n_i)\partial t_j \\ &= \int_{t_i}^{v_1} n_i \frac{(B_0(t_j,t_i))^{n_i-1}}{(B_0(\infty,t_i))^{n_i}} \partial B_0(t_j,t_i) \\ &= \left(\frac{B_0(v_1,t_i)}{B_0(\infty,t_i)}\right)^{n_i} \end{aligned} \tag{2.41}$$

Gernhard (2008a) also provided an alternative version of this equation for the special case in which λ and μ are constant, $t_j = 0$ and $B_0(\infty) = 1$.

Once we know v_1 it is easy to calculated the probability of having n_j reconstructed taxa, as $P(n_j|n_i, v_1)$ is the same as $P(n_j|n_i, n^{>}_{v_1})$. We can use this to calculate $P(n_j|n_i)$, the probability that you will have n_j reconstructed lineages given that you have n_i at some time after that by integrating over all the possible values of v_1.

$$\begin{aligned} P(n_j|n_i) &= \int_{t_j}^{\infty} P(n_j|n_i, v_1) f(v_1|n_i) dv_1 \\ &= n_i \binom{n_i - 1}{n_j - 1} \int_{t_j}^{\infty} \left(1 - \frac{B_0(t_j,t_i)}{B_0(v_1,t_i)}\right)^{n_j-1} \left(\frac{B_0(t_j,t_i)}{B_0(v_1,t_i)}\right)^{n_i-n_j} \frac{(B_0(v_1,t_i))^{n_i-1}}{(B_0(\infty,t_i))^{n_i}} \partial B_0(v_1,t_i) \\ &= \binom{n_i}{n_j} \left(\frac{B_0(t_j,t_i)}{B_0(\infty,t_i))}\right)^{n_i-n_j} \left(1 - \frac{B_0(t_j,t_i)}{B_0(\infty,t_i)}\right)^{n_j} \end{aligned} \tag{2.42}$$

Now we have calculated a probability density for predicting the number of reconstructed lineages going backwards in time that is based on a reasonable prior, and produces pretty equations. This is a binomial distribution with n_i trials, $n_i - n_j$ successes and probability of success $B_0(t_j,t_i)/B_0(\infty,t_i)$.

One interesting thing to note is that under this distribution unlike all the others, zero reconstructed lineages means something intelligible. There will be zero reconstructed lineages before the studied lineages split from their sister clade. At the time that the two clades diverge, the process will go from zero to one reconstructed lineages. In fact when we compare (2.38) and (2.42) we see that $P(n_j|n_i) = P(n_j|n_\infty = 0, n_i)$.

To evaluate (2.42) we must determine the value of $B_0(\infty)$. The limit of u_i as t_i approaches infinity is 1 if a_i is less than 1, and $1/a_i$ if it is greater. Therefore so long as the limit of a_i is less than 1 as t_i approaches ∞, $B_0(\infty)$ will equal 1. Furthermore if a is constant and greater than 1, then $B_0(\infty)$ will equal $1/a$, and we can see from (2.25) that if a is constant and greater than 1 at all times before some time t_j, then:

$$1-B_0(\infty) = \frac{(a-1)(1-B_0(t_j))}{a-E_0(t_j)}$$

However, if $\lim_{t_i\to\infty} a(t_i) > 1$ and $a(t_i)$ varies with time, then we must use numerical integration of the equations found in Kendall (1948) to solve for $B_0(\infty)$. From an abstract position we can assume that $\lim_{t_j\to\infty} a(t_j)$ is in fact less than 1, because if it is not, then $\lim_{t_j\to\infty} E_0(t_j)=1$, and it makes no sense to talk about reconstructed lineages that we know have to be extinct. In other words, because we know that life is very old and very diverse, the speciation rate must have exceeded the extinction rate for most of that time. However, from a practical perspective, if we want our solutions to the equations that involve $B_0(\infty)$ to be correct, we must either solve for $B_0(\infty)$ or explicitly define some time before which a is less than 1.

2.4 Distribution of Branching Times

Comparing the distribution of lineage counts over time under different implementations of the birth-death process can be informative. However, to explore the relationship between individual trees and a particular model of lineage diversification we must look at how the number of lineages at all times relate to each other and not just to the number of lineages at the start or the end of the process. To do so we must investigate the waiting times between reconstructed lineage splitting events.

2.4.1 Cumulative Distribution of Waiting Times

I will define v_n as the time before the present that the clade in question went from $n-1$ to n reconstructed lineages. I will define n_{v_n} as the number of lineages at time v_n, therefore $n_i = n_{v_n}$ if $t_i = v_n$. However, v_n is the instant that the number of lineages changes, so that there are a different number of lineages if you approach v_n from the future or the past. I will define $n_i^<$ as the number of reconstructed lineages immediately before t_i, and $n_i^>$ as the number of lineages immediately after t_i, therefore $n_i^< = n_i^> = n_i$, if no lineages are added at time t_i, but $n_i^< + 1 = n_i^> = n$ if $t_i = v_n$. Following Nee et al. (1994b) we see that if we have n reconstructed lineages at v_n, the probability that the next speciation event occurs after v_{n+1} is the probability that there have been no more reconstructed lineages left by time v_{n+1}. We will use this to calculate the distribution of the waiting time between v_n and v_{n+1}.

If we proceed forward from some starting time at which we know the number of reconstructed lineages we can calculate the distribution of each subsequent waiting time based only on the branching time at the start of that wait, because once we know the number of lineages at that time, then the distribution of lineages at subsequent times will be independent of the number of lineages before that time. Under this circumstance we should calculate the distribution of waiting times between v_n and v_{n+1} using (2.37).

$$F(v_{n+1}|v_n) = F(v_{n+1}|n^{>}_{v_n}) = P(n^{<}_{v_{n+1}}|n^{>}_{v_n}) = (1 - B_0(v_n, v_{n+1}))^n \tag{2.43}$$

This is the same as equation 16 in Nee et al. (1994b).

We can also proceed backwards in time from some time when we know the number of lineages, calculating the cumulative distribution for each successive waiting time. In that case the distribution of waiting times between v_n and v_{n+1} will depend only on v_{n+1} and we should use (2.42).

$$1 - F(v_n|v_{n+1}) = 1 - F(v_n|n^{<}_{v_{n+1}}) = P(n^{>}_{v_n}|n^{<}_{v_{n+1}}) = \left(1 - \frac{B_0(v_n, v_{n+1})}{B_0(\infty, v_{n+1})}\right)^n \tag{2.44}$$

When $B_0(\infty)$ is one, the probability that the waiting time between v_n and v_{n+1} is as long as it is will be the same whether one knows v_n or v_{n+1}

If we want to look at the distribution of waiting times between two times when we know the number of reconstructed lineages, then there are two options available to us. We can either proceed forward from the earlier of those two times or backwards from the later one. Here I will calculate the distribution of waiting times between v_n and v_{n+1} for the former case, so that I will assume that we know v_n and the number of lineages at some later time, t_i. I will use (2.38) to make this calculation.

$$F(v_{n+1}|v_n, n_i) = F(v_{n+1}|n^{>}_{v_n}, n_i) = P(n^{<}_{v_{n+1}}|n^{>}_{v_n}, n_i) = \left(\frac{B_0(v_{n+1}, t_i)}{B_0(v_n, t_i)}\right)^{n_i - n} \tag{2.45}$$

2.4.2 Density of Waiting Times

We can now calculate the probability densities for these waiting times by taking the derivatives of the cumulative distributions for each case. This is easy to do when we recognize that $B_0(t_k)$ and $B_0(t_i)$ are not affected by the value of t_j, as changes in the timing of t_j do not affect the values of λ or μ at any time. This allows us to calculate the probability density for the waiting time after a lineage splitting event that occurred at some time. First I will calculate this density using Assumption 1 by assuming that we know v_n and taking

the derivative of (2.43).

$$\begin{aligned} f(v_{n+1}|v_n) &= \frac{\partial F(v_{n+1}|v_n)}{\partial v_{n+1}} \\ &= n\frac{(1-B_0(v_n))^n}{(1-B_0(v_{n+1}))^{n+1}}\frac{\partial B_0(v_{n+1})}{\partial v_{n+1}} \\ &= n\lambda(v_{n+1})(1-E_0(v_{n+1}))(1-B_0(v_n,v_{n+1}))^n \end{aligned} \tag{2.46}$$

If we use Assumption 2 and assume that we know v_{n+1} then we take the derivative of (2.44).

$$\begin{aligned} f(v_n|v_{n+1}) &= \frac{\partial F(v_n|v_{n+1})}{\partial v_n} \\ &= n\frac{1}{B_0(\infty,v_{n+1})}\left(1-\frac{B_0(v_n,v_{n+1})}{B_0(\infty,v_{n+1})}\right)^{n-1}\frac{\partial B_0(v_n,v_{n+1})}{\partial v_n} \\ &= n\lambda(v_n)(1-E_0(v_n))\frac{1-B_0(v_n,v_{n+1})}{B_0(\infty,v_{n+1})}\left(1-\frac{B_0(v_n,v_{n+1})}{B_0(\infty,v_{n+1})}\right)^{n-1} \end{aligned} \tag{2.47}$$

We can of course also calculate the probability density for the waiting time when we know the timing of the lineage splitting event at the start of that period and we know the number of lineages at some later time by taking the derivative of (2.45).

$$\begin{aligned} f(v_{n+1}|v_n,n_i) &= \frac{\partial F(v_{n+1}|v_n,n_i)}{\partial v_{n+1}} \\ &= (n_i-n)\frac{(B_0(v_{n+1},t_i))^{n_i-n-1}}{(B_0(v_n,t_i))^{n_i-n}}\frac{\partial B_0(v_{n+1},t_i)}{\partial v_{n+1}} \\ &= (n_i-n)\lambda(v_{n+1})(1-E_0(v_{n+1}))\frac{1-B_0(v_{n+1},t_i)}{B_0(v_{n+1},t_i)}\left(\frac{B_0(v_{n+1},t_i)}{B_0(v_n,t_i)}\right)^{n_i-n} \end{aligned} \tag{2.48}$$

2.4.3 Density of a Set of Branching Times

Once we have calculated the probability densities for the timing of each lineage splitting event we can calculate the probability of the set of branching times for our entire clade by multiplying the densities of each event. We will define $V(t_k,t_j)$ as the set of branching times for our clade between t_k and t_j, therefor $V(\infty,t_0)$ is the set of all branching times for our clade. Furthermore we will define $\beta(V(t_k,t_j))$ as the product of the slope of $B_0(v,t_j)$ for all v in $V(t_k,t_j)$.

$$\begin{aligned} \beta(V(t_k,t_j)) &\equiv \prod_{v\in V(t_k,t_j)}\frac{\partial B_0(v,t_j)}{\partial v} \\ &= \prod_{v\in V(t_k,t_j)}\lambda(v)(1-B_0(v,t_j))(1-E_0(v)) \end{aligned} \tag{2.49}$$

As a first step we will use Assumption 1 to calculate the probability of a set of branching times after t_k, given the number of reconstructed lineages at t_k. Here we must include the density of the first branching time after t_k given the number of reconstructed lineages at that time, the density of every other branching time given the branching time before it and the cumulative probability that there are no more lineage splits between the last branching time and time t_0.

$$\begin{aligned} f(V(t_k,t_0)|n_k) &= P(n_0|n^{>}_{v_{n_0}}) f(v_{n_k+1}|n_k) \prod_{m=n_k+1}^{n_0-1} f(v_{m+1}|v_m) \\ &= \frac{(n_0-1)!}{(n_k-1)!}(1-B_0(t_k))^{n_k}\beta(V(t_k,t_0)) \end{aligned} \tag{2.50}$$

Equation 20 in Nee et al. (1994b) is the special case of this equation in which $n_k=2$. Rabosky (2006a) attempted to derive this equation for the DTBD, but failed to account for the effect that changes in μ would have on the loss of unobserved lineages.

We can also calculate the likelihood of a set of branching times before some time t_i given that we know the number of reconstructed lineages at t_i using Assumption 2. This calculation will include all the branching times before t_i back to the most recent common ancestor. The product must include the density of the last branching time given the number of taxa at t_i, and the densities of all the other branching times, given the branching time after them.

$$\begin{aligned} f(V(\infty,t_i)|n_i) &= f(v_{n_i}|n_i) \prod_{m=2}^{n_i-1} f(v_m|v_{m+1}) \\ &= n_i!\frac{(1-B_0(v_2,t_i))B_0(\infty,v_2)}{(B_0(\infty,t_i))^{n_i}}\beta(V(\infty,t_i)) \end{aligned} \tag{2.51}$$

Gernhard (2008a) calculated this equation for the special case in which λ and μ are constant, $t_i=0$ and $B_0(\infty)=1$.

Finally we can calculate the probability of a set of branching times between two times when we assume that we know the number of lineages at both those times using Assumption 3 in the same way as Assumption 1, except we do not have to include the probability that another lineage split did not occur after the last lineage split, as that is an assumption of the model.

$$\begin{aligned} f(V(t_k,t_i)|n_k,n_i) &= f(v_{n_k+1}|n_k,n_i) \prod_{m=n_k+1}^{n_i-1} f(v_{m+1}|v_m,n_i) \\ &= (n_i-n_k)!\frac{\beta(V(t_k,t_i))}{(B_0(t_k,t_i))^{n_i-n_k}} \end{aligned} \tag{2.52}$$

Gernhard (2008a) provides a version of this equation for the special case in which λ and μ are constant, $t_i=0$ and $n_k=1$. We can see from this equation that under Assumption

3, the branching times will be independent and identically distributed by the definition of independence.

It is highly informative to consider how our assumptions about the number of taxa affect our calculation of the probability of a set of branching times. To do so we will compare the probability of the branching times when we assume we know the number of lineages at the beginning and the end of the process to those when we know only one of those. For example the relationship between the probability density under Assumption 1 and Assumption 3 can be described as follows.

$$f(V(t_k, t_0)|n_k) = P(n_0|n_k) f(V(t_k, t_0)|n_k, n_0)$$

The probability density under Assumption 3 is the same as that under Assumption 1, except that it does not include the probability of going from n_k reconstructed lineages to n_0 lineages in the present, as it assumes that this is true. It is critical for a researcher using this assumption to be aware of this fact. If one tried to fit values of λ and μ to a data set using this assumption, those values would be based solely on the distribution of taxa between endpoints, and under those values it may be highly unlikely that one would go from n_k reconstructed lineages at the beginning of the process to n_0 lineages in the present. As a consequence of these key differences, if you integrate the probability densities over all the possible branching times for Assumption 3, you get one, but for Assumption 1 the integral is the probability that reconstructed lineages at time t_k would produce n_0 lineages.

We can also examine the relationship between the probability density of a set of branching times under Assumption 2 and Assumption 3.

$$f(V(\infty, t_i)|n_i) = f(v_2|n_i) f(V(v_2, t_i)|v_2, n_i)$$

Where:

$$f(v_2|n_i) = \frac{\partial F(v_2|n_i)}{\partial v_2} = -\frac{\partial P(n_{v_2} < 2|n_i)}{\partial v_2}$$

Assuming that we know the timing of the first lineage split, changes the dimensionality of the probability density. If we do not assume that we know when the last common ancestor occurred, the probability density will be divided by a unit of time to account for this additional uncertainty. This change in the probability is the probability that the first lineage split would occur when it did, given the number of reconstructed taxa at t_i. Integrating this probability density over all the possible values for the branching times and the origin under Assumption 1 will also produce a probability of 1, as it did under assumption 3.

We can see from (2.51) that v_2 is not distributed the same as v for the other lineage splits under Assumption 2, and thus the values of v will not be independently distributed. However, if we also consider the distribution of v_1 we see that:

$$\begin{aligned} f(V(\infty, t_i), v_1|n_i) &= f(v_1|v_2) f(V(\infty, t_i)|n_i) \\ &= n_i! \frac{\beta(V(\infty, t_i) \cup v_1)}{(B_0(\infty, t_i))^{n_i}} \end{aligned} \tag{2.53}$$

Thus v for all the branching times and v_1 are together all independently and identically distributed. Thus it would be informative if one could include v_1, the time that the clade in question diverged from its extant sister clade, in a calculation of the probability of a set of branching times under Assumption 2.

I implemented these calculations in two R functions, ml.bd and calc.like.bd. calc.like.bd can calculate the probability of a set of branching times given any set of parameter values and assumptions about the number of reconstructed lineages. ml.bd determines the maximum likelihood and the maximum likelihood parameter values for any set of free parameters for a set of branching times. When t_j is a free parameter and it separates two discrete periods with different sets of diversification rates, then each branching time will form a local maximum for t_j. Therefore, when one of the free parameters is t_j, the maximum likelihood value for that parameter is found through an exhaustive search of the branching times. All other parameters are fit using the nlminb function from the R base package (R Development Core Team 2010).

2.4.4 Waiting Times Independent of Number of Lineages

It is difficult to investigate the distribution of waiting times for a set of parameters without a tree to which we can compare them. The waiting times I have calculated so far depend on knowing the number of lineages and the timing of the branching time that precedes the wait; therefore one must have a tree with a set of known branching times that can constrain both the time and the number of lineages. It would be useful not only to know the probability of the wait after a duplication to produce a known number of reconstructed lineages at a known time, but also of the waiting period from some time until the next lineage split independent of how many lineages there are at that time.

We can calculate the probability that a lineage split did not occur over some period of time by summing the probability that the number of reconstructed lineages is the same at the beginning and end of that period over every possible number of reconstructed lineages. Below I show these formulas under all assumptions.

Assumption 1:

$$\begin{aligned} P(V(t_j,t_i)=\varnothing|n_k) &= \sum_{n_j=n_k}^{\infty} P(n_i=n_j|n_j)P(n_j|n_k) \\ &= \left(1+\frac{B_0(t_j)-B_0(t_i)}{1-B_0(t_k)}\right)^{-n_k} \end{aligned} \tag{2.54}$$

Assumption 2:

$$P(V(t_k,t_j)=\varnothing|n_i) = \sum_{n_j=0}^{n_i} P(n_k=n_j|n_j)P(n_j|n_i)$$
$$= \left(1-\frac{B_0(t_k)-B_0(t_j)}{B_0(\infty)-B_0(t_i)}\right)^{n_i} \tag{2.55}$$

Assumption 3:

$$P(V(t_k,t_j)=\varnothing|n_l,n_i) = \sum_{n_k=n_l}^{n_i} P(n_j=n_k|n_k,n_i)P(n_k|n_l,n_i)$$
$$= \left(1-\frac{B_0(t_k)-B_0(t_j)}{B_0(t_l)-B_0(t_i)}\right)^{n_i-n_l} \tag{2.56}$$

All these functions are exponentially distributed with respect to the number of reconstructed lineages that we assume we know. Furthermore they all depend on the difference between B_0 at the beginning and B_0 at the end of our time period and have a great deal in common in general. They will all prove very useful when we investigate the effects of different parameter values without regard to an actual tree.

2.4.5 Simulating Trees

Now that we have calculated the cumulative probability for waiting times under a TVBD with any set of assumptions about the number of reconstructed lineages at various times in the process it is trivial to generate a random tree under such a process. Varying the parameters with time will affect the branching times of a tree under a birth-death process, but it will have no effect on the topology (Thompson 1975; Sanderson and Bharathan 1993). Therefore we can simulate a tree by generating a random set of branching times using the method described below, then randomly choosing pairs of clades to combine into new clades and assigning the branching times to those clades in the reverse of the order that the clades are combined.

The first step in generating a random tree is to establish a set of assumptions about the number of taxa. One could choose from the three strict assumptions or a relaxed version of those three with any set of assumptions about the number of taxa at any time. We have already established in subsection 2.4.3 that under Assumption 3 the branching times are independent and identically distributed. Furthermore we can see from (2.52) that the density of each branching time is:

$$f(v|n_k,n_i) = \lambda(v)\frac{(1-E_0(v))(1-B_0(v,t_i))}{B_0(t_k,t_i)} \tag{2.57}$$

We can now obtain the cumulative distribution of any branching time by integrating this equation.

$$\begin{aligned} F(v|n_k, n_i) &= \int_{t_i}^{v} \lambda(t_j) \frac{(1-E_0(t_j))(1-B_0(t_j,t_i))}{B_0(t_k,t_i)} \partial t_j \\ &= \frac{B_0(v,t_i)}{B_0(t_k,t_i)} \end{aligned} \tag{2.58}$$

We see that under Assumption 3, $B_0(v)$ for any reconstructed lineage split should be uniformly distributed between $B_0(t_i)$ and $B_0(t_k)$. Modifying Felsenstein (2004, pg. 570) and Hartmann et al. (2010) we generate $n_i - n_k$ random branching times between times t_i and t_k by generating n_i–n_k uniform random numbers between $B_0(t_k)$ and $B_0(t_i)$. We should then order those numbers and treat them as $B_0(v)$ for our random branching times, and then use those to calculate the values of all v. For the DTBD we can calculate the inverse of $B_0(v)$ using the methods described in subsection 2.2.5. For some other TVBDs the inverse of $B_0(v)$ can be calculated by solving the integrals in Kendall (1948) and calculating their inverse.

In this way we can calculate random branching times between two known numbers of reconstructed lineages for any time variable birth-death process in a very straight forward way. However, we established in subsection 2.4.3 that under Assumption 2 the branching times will not be independently distributed, but all the branching times together with v_1 will be independently and identically distributed, such that all the values of $B_0(v)$ will be uniformly distributed between $B_0(t_i)$ and $B_0(\infty)$. So, in order to simulate a set of branching times for a process that ends with n_i reconstructed lineages, we should generate n_i values from the uniform distribution between $B_0(t_i)$ and $B_0(\infty)$, and treat the highest value as $B_0(v_1)$ and the other $n_i - 1$ values as $B_0(v)$ for the branching times of our tree.

In order to simulate branching times for a tree starting with n_k reconstructed lineages, we must first establish the number of lineages at the end of the process. Given (2.37), we can do this by taking Δn, a random draw from the negative binomial distribution with n_k targeted successes and probability of failure $B_0(t_k)$. We can then assume that $n_0 = \Delta n + n_k$, and generate the Δn branching times, as we did for Assumption 3.

I have implemented all these procedures in the simulate.tree function in the R package telos. Figure 2.3 shows lineage through time plots for several random trees generated under each assumption.

2.5 Visualizing Distributions for the Time Variable Birth-Death Process

The equations presented in sections 2.2, 2.3 and 2.4 allow us to easily calculate the probability mass for any number of reconstructed lineages and the cumulative probability for the waiting

times at any time given a set of birth-death parameters and any set of assumptions about the number of reconstructed lineages at any other time. I have implemented two plotting tools in the R package telos, that will allow us to create a visual representation of these distributions. The plot.LT.null function plots lineage through time null plots that show several quantiles for the number of reconstructed lineages as a function of time. The plot.WT.notree.null function plots several quantiles for the waiting time until the next reconstructed lineage split as a function of the time at which that waiting period starts. These tools have two obvious uses: to compare different sets of parameter values in order to see what effect those parameters have on the expected distribution; or to compare lineage through time plots for real phylogenies to the actual distribution, in order to see if there are significantly more or less lineages than we would expect at any time, or if the waiting times are excessively long or short.

2.5.1 How Varying Parameters Affects Distribution of Reconstructed Lineages

In order to demonstrate the utility of this method let us investigate how varying the parameters of the CRBD affects the distribution of lineage numbers for all three assumptions about the number of lineages. The number of taxa at the beginning and end of a process has a large effect on the appearance of the distribution. Therefore, we will constrain our parameter values so that for a process starting with ten reconstructed lineages we will expect to have forty lineages after one unit of time. This leaves us with one free parameter, we will vary the shape parameter, a, and use it to choose the scaling parameter r. The maximum likelihood values of λ and μ for a reconstructed process starting at time t_k with n_k lineages and ending at time t_j with n_j are found when $n_j/n_k = 1 - B_0(t_k, t_j)$. Furthermore the expectation of a process starting with n_k reconstructed lineages and using these maximum likelihood parameters for λ and μ will be n_j . For a CRBD that ends at the present, we can use (2.6) and (2.7) to solve for r for a given a, n_k and n_0, when a does not equal one.

$$r = \log\left((1-a)\frac{n_0}{n_k} + a\right)/t_k$$

When a does equal one, we can use (2.8) to solve for λ directly.

$$\lambda = \left(\frac{n_0}{n_k} - 1\right)/t_k$$

This will allow us to calculate the maximum likelihood value for λ and μ given a for a process starting with ten reconstructed lineages one time unit before the present and expected to reach forty lineages in the present.

These parameter values can in turn be used to calculate the probability masses and expectations of the number of reconstructed lineages at any time for this process. Figure 2.4

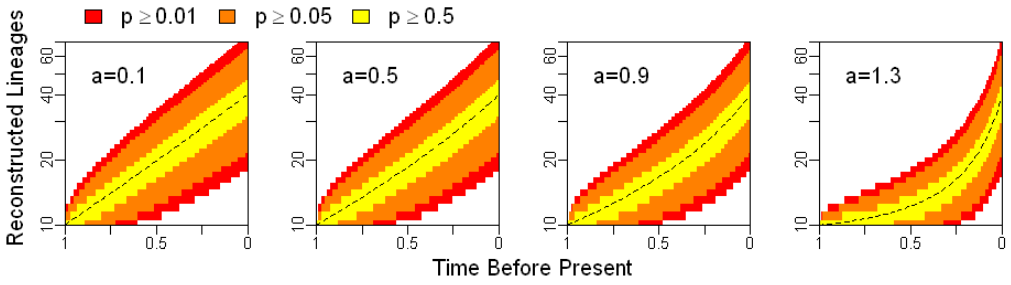

Figure 2.4: Effects of varying the shape parameter, a, on the distribution of reconstructed lineages over time, if we assume that we know the number of reconstructed lineages in the past. We assume that there were ten reconstructed lineages one time unit before the present. We varied r between plots, so that the expected number of lineages in the present would stay 40 as we varied a. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

shows plots of these probabilities, called lineage through time null plots, for values of a ranging from 0.1 to 1.3, assuming that we have ten reconstructed lineages one time unit before the present. It should be noted that the expectation of this process at any time can only be 40 for values of a less than 4/3. All of these plots have an expectation of 40 at the end of the process as they should. Furthermore, for all the plots the variance increases as we move towards the present and away from the time at which we know the number of reconstructed lineages. Comparing these plots allows us to see that as a increases the curvature of the expectation and any given quantile become more positive with respect to time, as originally described by Nee et al. (1994b). In the most extreme case, when $a = 1.3$, a concave or even flat log lineage through time plot is highly unlikely.

We can also look at he effects of varying the parameters using either Assumption 2 or 3. Figure 2.5 shows the plots of the lineage through time null plots for the same parameter values as in Figure 2.4, but assuming that there are forty lineages in the present while the number of lineages at all times in the past are free to vary. Figure 2.6 shows the same plots but assuming both that there are forty lineages in the present and that there are ten reconstructed lineages one time unit before the present. Both of these plots show the same increase in curvature for higher values of a that we saw in the situation in which we assumed ten reconstructed lineages in the past. We also see, not surprisingly that the variance increases as we move away from the times when we know the number of lineages.

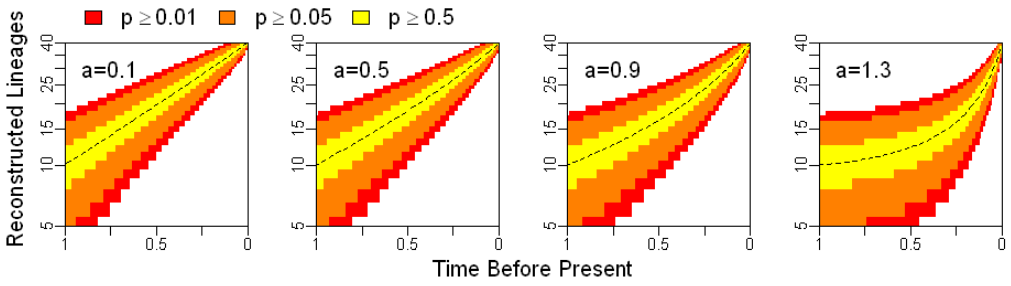

Figure 2.5: Effects of varying the shape parameter, a, on the distribution of reconstructed lineages over time, if we assume that we know the number of lineages in the present. We assume that there are forty lineages in the present. We varied r between plots, so that the expected number of reconstructed lineages one time unit before the present would remain ten as we varied a. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

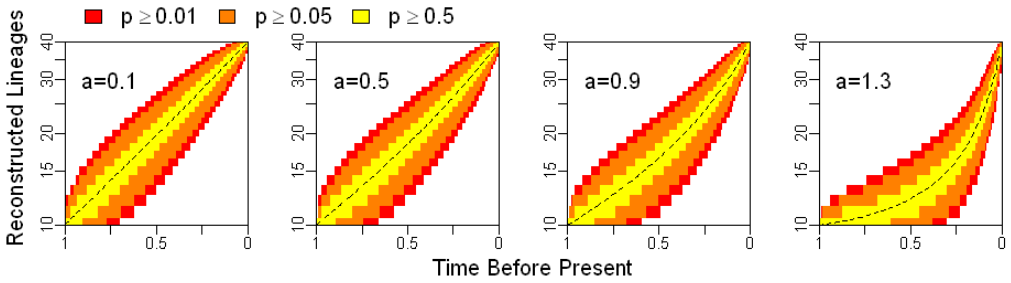

Figure 2.6: Effects of varying the shape parameter, a, on the distribution of reconstructed lineages over time, if we assume that we know the number of reconstructed lineages in the present and the past. We assume that there were ten reconstructed lineages one time unit before the present and forty lineages in the present. We varied r between plots, so that the expected number of lineages in the present would stay forty for a process starting with ten reconstructed lineages as we varied a. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

2.5.2 How Varying Parameters Affects Distribution of Waiting Times

Researchers have used lineage through time plots for the last fifteen years in order to investigate the distribution of branching times. However, lineage through time plots are highly autocorrelated, and this can lead to misleading interpretations of data. For example a single period during which the number of lineages increases rapidly may lead to unexpectedly high numbers of lineages at all subsequent times. Similarly calculating the cumulative probability of a number of lineages, as we did for the null plots in the previous section, will not be appropriate for any random tree at any given time. Trees that start with few lineages early will tend to have few lineages late and those that start with many will tend to be in the upper percentiles later in the process. We do not expect the number of lineages to go careening back and forth between the highest and the lowest percentiles, but rather to generally stay along a given course.

Observing the distribution of waiting times, instead of lineages will alleviate this problem. Waiting times for a given tree will not be distributed independently of each other, but they also will not be autocorrelated. Furthermore, they provide us with much of the information that we expect to extract from a lineage through time plot. Shorter waiting times lead to faster increases in the number of lineages. Waiting times that decrease as we approach the present will create convex lineage through time plots. In essence waiting times are the inverse of the slope of a lineage through time plot. Later I will introduce a graphical tool that allows us to compare waiting times for a specific tree with known branching times to their distribution under the birth death process. Here we will use the equations from subsection 2.4.4 to infer the effects of different parameter values on waiting times without reference to a specific tree.

Figure 2.7 shows the distribution of waiting times as a function of time using the same parameter values as we used in Figure 2.4, and assuming that we start with ten reconstructed lineages. The colors show the two-tailed quantiles, and the solid line shows the present, after which time the process would end and no more new lineages would be observed. The y-axis is cube root transformed, as it will be for all the waiting time plots in this paper, in order to reduce the skew in the distribution. Under all sets of parameter values the branching times decrease as we approach the present. Under the processes with larger as the waiting times decrease more, so that in the process for which a is 1.3 we have the longest waiting times observed in any of the plots at the beginning of the process and the shortest waiting times at the end.

Figure 2.8 shows the distribution of waiting times as a function of time using the same parameter values as we used in Figure 2.7, but assuming that we end with forty reconstructed lineages in the present. Unlike Figure 2.7 there is no line showing the present, because under Assumption 2 we calculate the probabilities of our waiting times going backwards in time from the end of the period, and so those distributions are not constrained by any hard end point as they are under Assumption 1. Other than that these plots are essentially the same

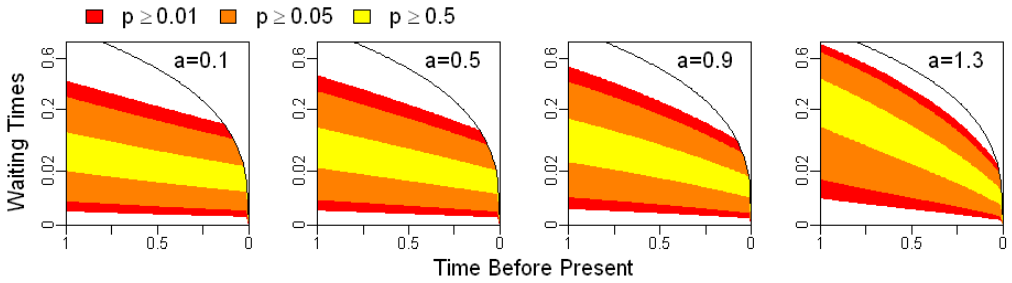

Figure 2.7: Effects of varying the shape parameter, a, on the distribution of waiting times over time, if we assume that we know the number of reconstructed lineages in the past. We assume that there were ten reconstructed lineages one time unit before the present. We varied r between plots, so that the expected number of lineages in the present would stay 40 as we varied a. The colored areas show two-tailed percentiles of the distribution at each time and the solid line represents the present.

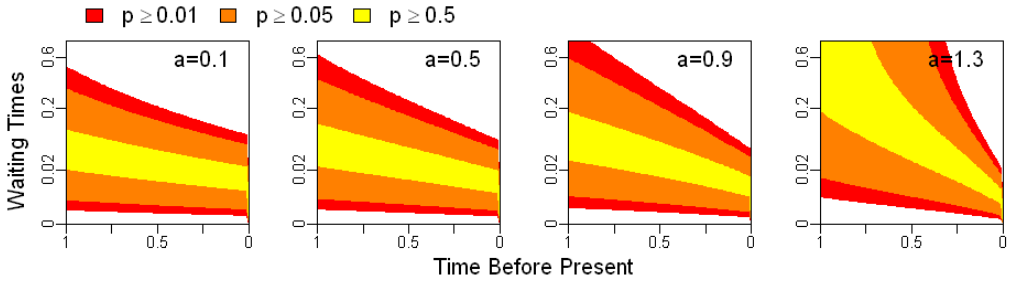

Figure 2.8: Effects of varying the shape parameter, a, on the distribution of waiting times over time, if we assume that we know the number of lineages in the present. We assume that there were forty lineages in the present. We varied r between plots, so that the expected number of lineages in the present would remain forty as we varied a for a process starting with ten reconstructed lineages. The colored areas show two-tailed percentiles of the distribution at each time.

as in Figure 2.7, although the large waiting times at the most ancient times are even more exaggerated when a is 1.3. This is because under Assumption 2 we include the possibility that there is only one lineage or that we may be at a time before the sister clade broke off. There will be no more reconstructed lineage splits preceding the time at which there is one lineage, and so the waiting times will be infinite.

2.5.3 Distributions under the Discrete Time Birth-Death Process

The formulas given in this paper, not only allow us to plot lineage through time null distributions for the CRBD with a variety of assumptions about the number of reconstructed lineages but also for situations in which the birth-death parameters vary as a function of time. Under the DTBD this is a very straight forward calculation given our ability to calculate $B_0(t_j)$ using (2.26), $E_0(t_j)$ using (2.24), and the probability densities and masses for numbers of reconstructed lineages and waiting times demonstrated in section 2.3 and section 2.4.

Figure 2.9a shows the parameter values for a DTBD in which we start with 10 reconstructed lineages one unit of time before the present, and 0.5 units of time after the start of the process r switches from a negative value to a positive value. Figure 2.9c and e show the lineage through time null plot and the waiting time null plot for this same process. Over most of the time that the process proceeds the waiting times decline slowly leading to a slight positive curvature in the lineage through time null plot. However, immediately prior to the change in r, the waiting times increase greatly, leading to a sharp downward turn and an overall concave appearance in the lineage through time plot and a net increase in waiting times over the entire process.

Figure 2.9b also shows the parameter values for a DTBD, only now r switches from negative to positive 0.5 time units before the present, and Figure 2.9d and f show the lineage through time null plot and the waiting time null plot. Not surprisingly there is a sharp decrease in the waiting times immediately before 0.5 time units, which leads to an immediate increase in the slope of the lineage through time null plot.

In both cases the waiting times have large shifts immediately before the shift in r, because the x-axis of these plots is the time at which the waiting period starts. Thus waiting periods closer to the shift in r have less time until r changes and thus a greater chance of ending after r has changed, when a new set of birth death parameters control the waiting times.

2.5.4 Sampling and Mass Extinctions

We can also use the results from subsection 2.2.3 to calculate these distributions when there is taxon sampling or a mass extinction. In order to observe the effects of sampling independently of the large effect it has on the expectation of the number of taxa at the end of the process, we held a and the expected number of sampled lineages at t_0 constant, while we varied the number of lineages sampled, by adjusting r. The lineage through time null plots generated by these processes are shown in Figure 2.10a and the waiting time null plots in Figure 2.10b. As the fraction of extant lineages sampled decreases the slope of the waiting time plots becomes less negative with shorter waiting times in the past and longer waiting times in the present. As a consequence the curvature of the log lineage through time plots decreases, even appearing concave when 90% of lineages are not sampled. A similar phenomenon was originaly described by Slatkin and Hudson (1991) in the context of

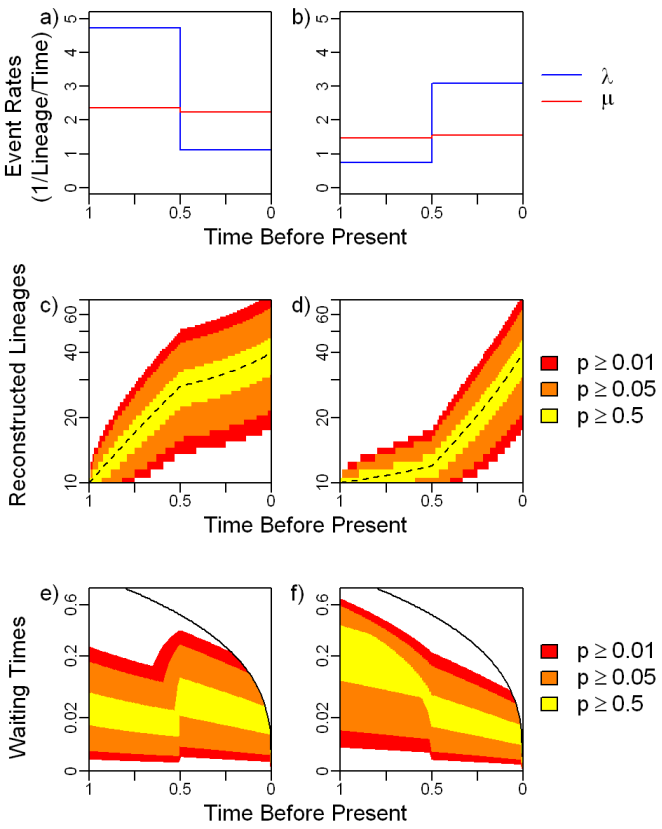

Figure 2.9: Effects of changing r half way through the process on the distribution of reconstructed lineages and waiting times through time. Under Model 1 $a = 0.5$ before 0.5 time units and $a = 2$ after 0.5 time units. Under Model 2 $a = 2$ before 0.5 time units and $a = 0.5$ after 0.5 time units. In both cases values of r are chosen, such that a process that starts with 10 reconstructed lineages one unit of time in the present is expected to have 40 lineages in the present. The values of λ and μ over time for a) Model 1 and b) Model 2. The distribution of reconstructed lineages over time starting with 10 reconstructed lineages one unit before the present for c) Model 1 and d) Model 2. The distribution of waiting times over time starting with 10 reconstructed lineages one unit before the present for e) Model 1 and f) Model 2. The colored areas show two-tailed percentiles of the distribution at each time, the dashed line is the expectation and the solid line represents the present.

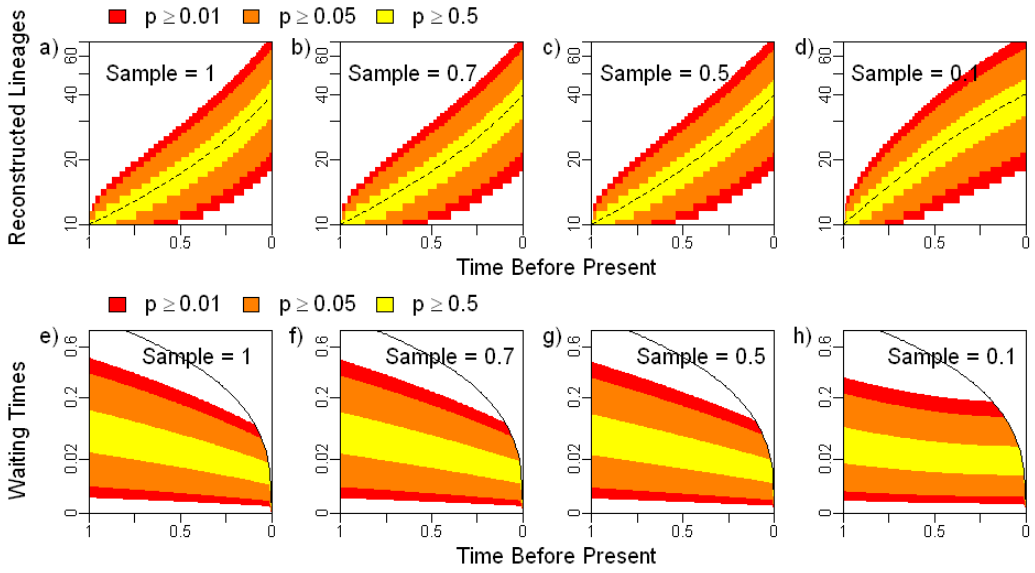

Figure 2.10: Effects of random sampling among extant lineages on the distribution of reconstructed lineages and waiting times over time, when a and the expected number of lineages at t_0 are held constant. Each pair of plots has a different fraction of lineages sampled at the end of the process: (a,e) 100%; (b,f) 70%; (c,g) 50%; (d,h) 10%. For each pair of plots λ and μ were chosen such that $a = 0.8$ and ten reconstructed lineages one time unit before the present are expected to produce forty samlped lineages in the present. (a,b,c,d) Distribution of number of reconstructed lineages through time as the fraction of lineages sampled decreases for a process starting with ten reconstructed lineages one time unit before the present and expected to end with forty sampled lineages in the present. Fewer lineages sampled results in more reconstructed lineages at all times between the start of the process and the present, so that the curvature of the plot decreases. (e,f,g,h) Distribution of waiting times through time as the fraction of lineages sampled decreases for a process starting with ten reconstructed lineages one time unit before the present and expected to end with forty sampled lineages in the present. Fewer lineages sampled results in shorter waiting times early in the process and longer waiting times closer to the present. The colored areas show two-tailed percentiles of the distribution at each time, the dashed line is the expectation and the solid line is the present.

poulation biology.

We can also easily observe the effects of a mass extinction. Figure 2.11 shows null lineage through time and waiting time plots, in which λ and μ are constant for the duration except at 0.5 units of time, when there is a 50% chance of any lineage going extinct. As you can see there is an increase in the slope at the time of the mass extinction, and a rapid decrease in the branching times immediately before the mass extinction. Furthermore, the slope of the waiting times is more negative after the mass extinction than it was before. These plots look a lot like the plots in which the value of r increased at 0.5 time units (Figure 2.9d and f). It is difficult to distinguish the effects of a mass extinction from those of an increasing diversification rate using only data from extant lineages.

2.5.5 Continuously (and Discontinuously) Varying Parameters

Several authors have proposed models for lineage through time plots in which the values of λ and μ do not change in an instant, but instead vary continuously over some period of time. We can use the DTBD to approximate the distributions of number of reconstructed lineages through time and waiting times through time for any TVBD.

First we will examine lineage through time null plots for some common models with continuously varying birth-death parameters. Rabosky and Lovette (2008a) proposed time variable models to describe the concave lineage through time plots seen in adaptive radiations. Under SPVAR the speciation rate decreases exponentially as we approach the present, so that $\lambda(t) = \lambda(0)\exp(kt)$, and under EXVAR the extinction rate starts at zero at time t_k and increases exponentially toward some asymptote, as we approach the present, so that $\mu(t) = \mu^*(1-\exp(k(t-t_k)))$. Here we will investigate two versions of these processes in which $r(t) = r(0)\exp(kt)$, by studying versions of SPVAR in which μ is zero, and versions of EXVAR in which λ is equal to μ^*. I generated parameter values for several processes under both SPVAR and EXVAR that start with ten reconstructed lineages one time unit before the present and are expected to have forty lineages in the present. I chose several values, from 1 to 15, for the ratio between r at the start of this process and r at the end. Maximum likelihood values for $\lambda(0)$ and μ^* were derived for each case using the ml.bd function in the R telos package.

Figure 2.12 shows several lineage through time null plots and waiting time null plots generated under the SPVAR processes described above. We see that for processes with large increases in r, the waiting times increases steadily as we approach the present (Figure 2.12c); we have not seen this under any other model. Furthermore the waiting time plots take on a distinctly convex shape with large ratios. As a consequence, the curvature of the lineage through time plots is smaller for processes with larger changes in r (Figure 2.12a). In comparison to SPVAR, changing the ratio between the r at the end and the beginning has little effect on the distribution of lineages and waiting times under the EXVAR model (Figure 2.13). There is a slight increase in curvature of the lineage through time plot and a slight decrease in the slope of the waiting times plot as the ratio increases, but this

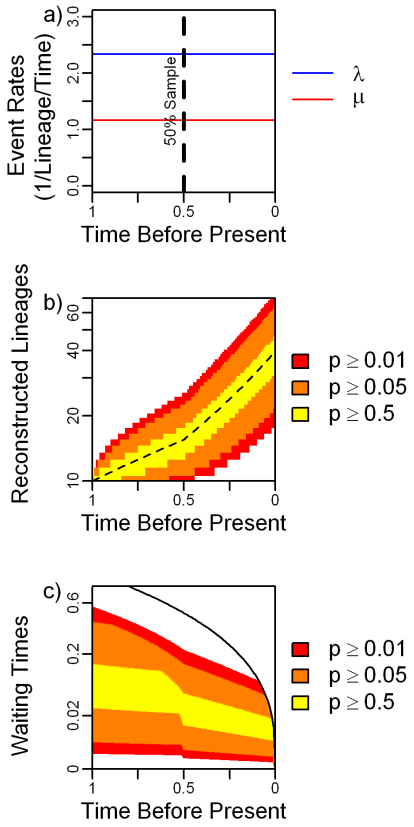

Figure 2.11: Effects of a mass extinction on the distribution of reconstructed lineages and waiting times over time. Distributions were generated with a constant λ and μ and 50% of lineages were removed 0.5 time units before the present. a was held at 0.5 and r was chosen, such that the expected number of lineages in the present is forty for a process that starts with ten reconstructed lineages one time unit before the present. a) The diversification rates used in this process and the timing and magnitude of the sampling event. b) The distribution of reconstructed lineages for a process starting with 10 reconstructed lineages one time unit before the present. c) The distribution of waiting times for a process starting with 10 reconstructed lineages one time unit before the present. The colored areas show two-tailed percentiles of the distribution at each time, the dashed line is the expectation and the solid line is the present.

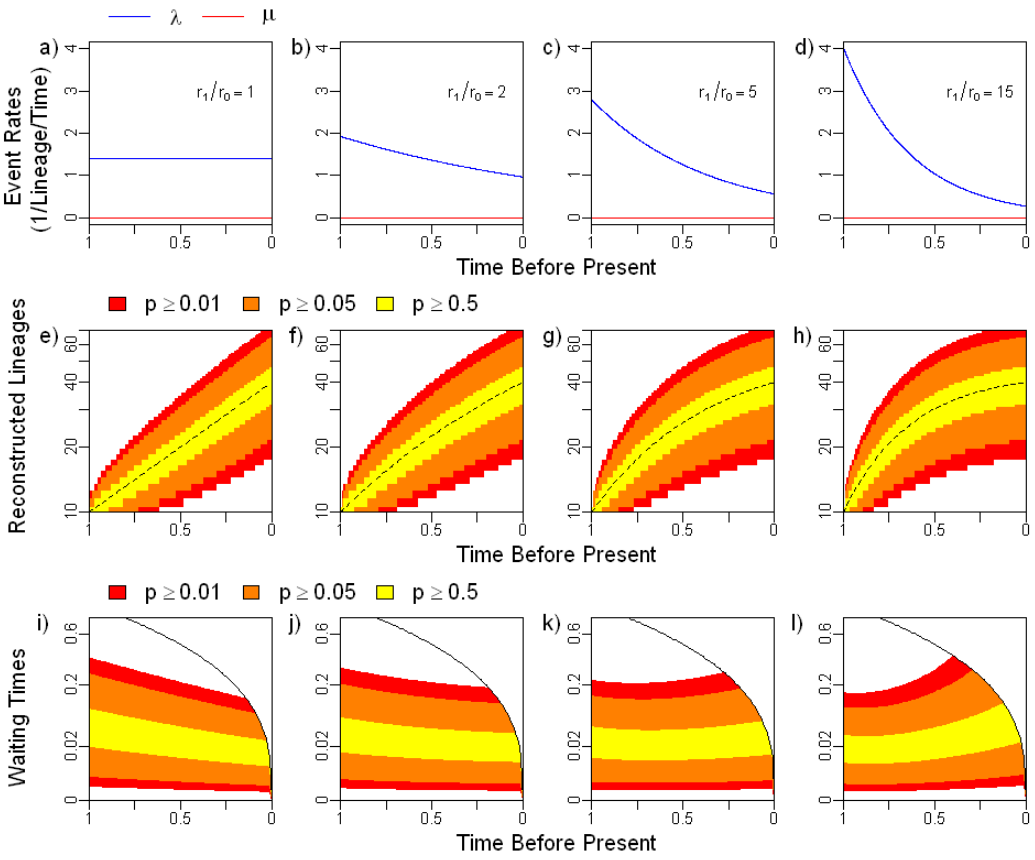

Figure 2.12: Effects of SPVAR birth-death parameters on the distribution of reconstructed lineages and waiting times over time. These figures show the distributions under four different birth-death models in which $\mu = 0$ and $\lambda = \lambda(0)\exp(kt)$. The ratio between $r(1)$ and $r(0)$ is $\exp(k)$ and was fixed at a different number for each model: (a,e,i) 1; (b,f,j) 2; (c,g,k) 5; (d,h,l) 15. λ was chosen such that a process starting with ten reconstructed lineages one time unit before the present is expected to have forty lineages in the present. (a,b,c,d) The values of λ and μ over time for each model. (e,f,g,h) The distribution of reconstructed lineages over time for each model starting with 10 reconstructed lineages one unit before the present. The dashed line is the expectation. (i,j,k,l) The distribution of waiting times over time for each model starting with 10 reconstructed lineages one unit before the present. The colored areas show two-tailed percentiles of the distribution at each time.

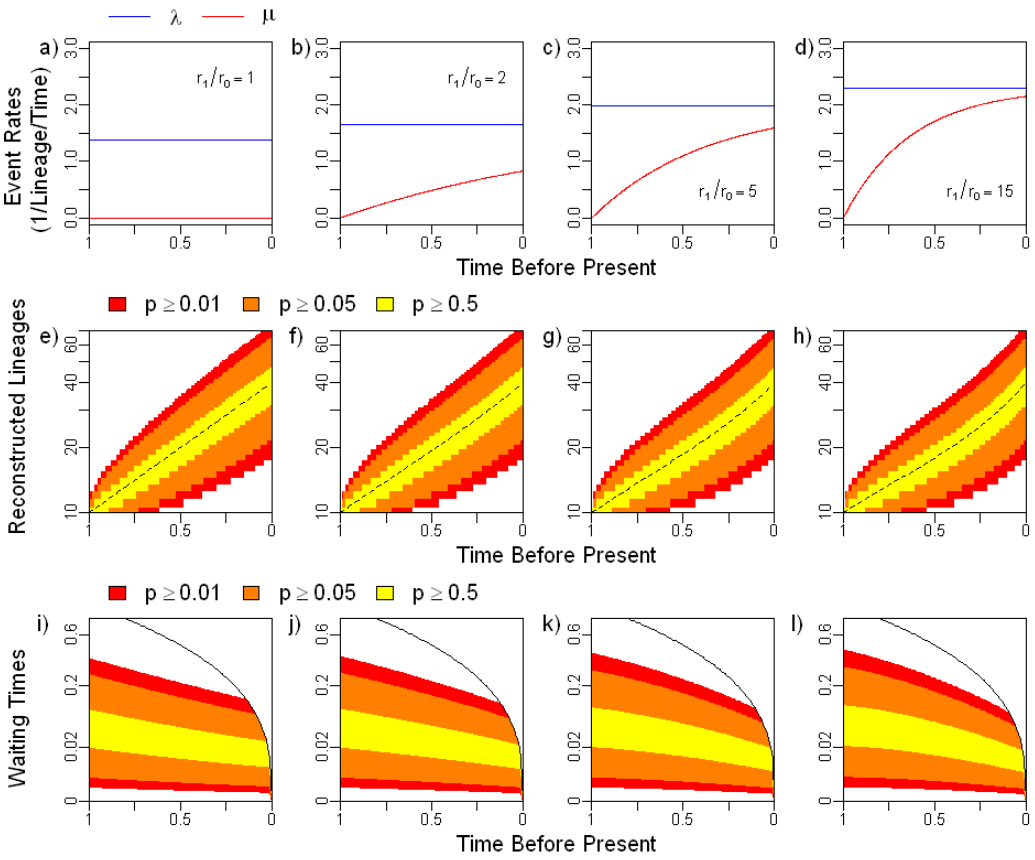

Figure 2.13: Effects of EXVAR birth-death parameters on the distribution of reconstructed lineages and waiting times over time. These figures show the distributions under four different birth-death models in which λ is constant and $\mu = \lambda(1 - \exp(k(t-1)))$. The ratio between $r(1)$ and $r(0)$ is $\exp(k)$ and was fixed at a different number for each model: (a,e,i) 1; (b,f,j) 2; (c,g,k) 5; (d,h,l) 15. λ was chosen such that a process starting with ten reconstructed lineages one time unit before the present is expected to have forty lineages in the present. (a,b,c,d) The values of λ and μ over time for each model. (e,f,g,h) The distribution of reconstructed lineages over time for each model starting with 10 reconstructed lineages one unit before the present. The dashed line is the expectation. (i,j,k,l) The distribution of waiting times over time for each model starting with 10 reconstructed lineages one unit before the present. The colored areas show two-tailed percentiles of the distribution at each time.

effect looks essentially no different from increasing the value of a. These results imply that concave lineage through time plots may be caused by decreasing speciation rates, but not by increasing rates of extinction as originally suggested by Rabosky and Lovette (2008a).

This method also makes it easy to combine continuously varying parameters for one period of time with sampling and constant parameters at other times. Figure 2.14 demonstrates what we might expect to happen, if the survivors of a mass extinction diversified in order to fill the niches left by those lineages who were not so lucky. In these plots there is no expected diversification and a turn over rate of 1.3 lineages per time unit for the first 1/3 of a time unit, then there is a mass extinction in which only one in 15 lineages survives. After the extinction the speciation rate rises instantly and then declines according to the SPVAR model until after another third of a time unit, after that point the parameters remain unchanged until the present. The parameters of SPVAR were chosen using the ml.bd function, such that λ and the expected diversity have returned to their pre-mass extinction levels at the time that the diversification rate returns to zero. This leaves a distinct pattern in which very few reconstructed lineages survive before the mass extinction, after which the plot becomes concave until settling down into a more constant slope (Figure 2.14b). We also see that the waiting times start out very large and then decline rapidly as we approach the mass extinction, they then increase steadily as λ declines and finally decrease over the last third of a time unit as λ returns to its original value (Figure 2.14c).

2.6 Testing the Fit of a Real Tree to a Birth-Death Model

So far I have developed several methods that allow us to explore the distribution of lineages and branching times under any TVBD. Ultimately the purpose of any analytical tool in science is the exploration of real data. Here we will investigate two real phylogenies, that have previously been used to analyze branching times.

The first tree is an ultrametric tree of 69 Australian agamid lizards from Harmon et al. (2003). The phylogeny was reconstructed by maximum likelihood using about 1800 base pairs of mitochondrial DNA and the branch lengths were made ultrametric using non-parametric rate smoothing (Sanderson 2003). This phylogeny was also analyzed in Rabosky (2006a) for temporal variation in birth death parameters. Following Rabosky (2006a) I scaled all the branch lengths so that the basal lineage split occurred 30 million years ago (MYA) (Hugall and Lee 2004).

The second tree is of 26 Plethodon salamander species from Highton and Larson (1979). This tree was constructed using UPGMA on a matrix of electrophoretic genetic distances and immunological distances between proteins. I scaled the branch lengths so that the divergence time of the two basal lineages was 42 MYA (Highton and Larson 1979). The authors made some inferences about the rate of speciation based on the distribution of branching times.

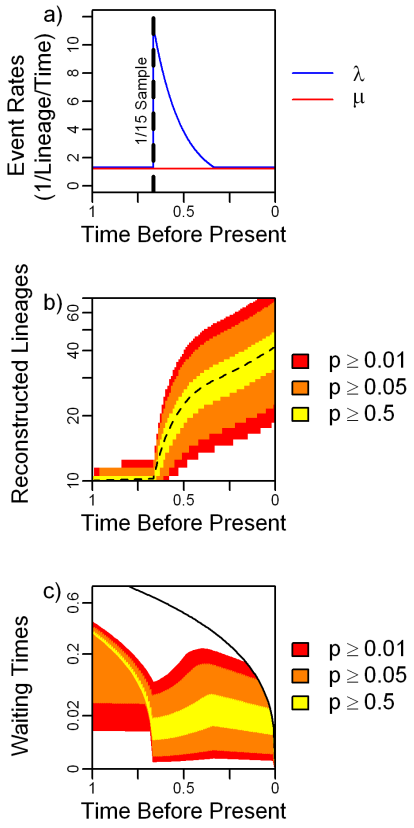

Figure 2.14: Effects of a complex scenario including a mass extinction and the recovery from it on the distribution of reconstructed lineages over time. These figures show the distributions under a complex time variable birth-death model in which μ is unchanged for the entire process and λ is equal to μ for the entire process except from 2/3 to 1/3 of a time unit before the present, so that during those times the number of lineages is not expected to change. 2/3 of a time unit before the present 93.3% of lineages die from a mass extinction. Immediately after the mass extinction λ jumps and then declines exponentially for 1/3 of a time unit until it returns to μ, during that time $\lambda = \mu \exp(k * (t - 1/3))$. k was chosen such that a process starting with ten reconstructed lineages one time unit before the present is expected to have forty lineages in the present. a) The values of λ and μ over time. b) The distribution of reconstructed lineages over time starting with 10 reconstructed lineages one unit before the present. The dashed line is the expectation. c) The distribution of waiting times over time starting with 10 reconstructed lineages one unit before the present. The colored areas show two-tailed percentiles of the distribution at each time.

This same tree was also analyzed in Nee et al. (1994a) and Nee et al. (1995) using the reconstructed birth-death process.

2.6.1 Quantitative Statistics

We can test the fit of a birth-death model to a set of branching times by calculating a single statistic, and then comparing that statistic to its distribution under the model. I will start by comparing these two data sets to a CRBD, as a null model of no variation in the diversification rates. I will do so for all three strict Assumptions about the number of lineages for sake of comparison. I used the maximum likelihood values for λ and μ under each Assumption, which I estimated those values using the equations from subsection 2.4.3 implemented in the ml.bd function from the telos R package. The maximum likelihood parameter values will probably generate better fit statistics than other parameter values and thus are less likely to reject the null hypothethis and represent a conservative test for rejection of the null hypothethis. Furthermore if this data really was generated by a CRBD then the maximum likelihood values are probably close to the actual values that the data was generated under.

One possible statistic is the maximum likelihood value itself. In order to derive an approximation of the distribution of this statistic under the model, one could simulate branching times under the CRBD (subsection 2.4.5) and compare the likelihoods of those simulations to the maximum likelihood for the actual branching times. For each real phylogeny I simulated 1000 sets of branching times with the same depth and the same number of terminal lineages as the tree using the parameters derived under the appropriate assumption. This actually represents the distribution under Assumption 3 although conditioned on the maximum likelihood parameter values from each respective assumption. One could use either of the other Assumptions to generate a distribution, but the depth of the tree and the number of terminal lineages probably have the largest effect on the likelihood and thus may drown out the signal of the distribution of lineages during the process. I then calculated the likelihood for the simulated branching times using the parameters under which they were simulated, and compared them to the maximum likelihood of the actual branching times. I rejected the birth-death process if more than 95% of the simulations had a likelihood larger than the actual data. This test is further biased towards accepting the model as the actual data was given the benefit of using its ML parameters, while the simulations are stuck with the parameter values they were simulated with.

A second possible statistic stems from the fact that $B_0(v)$ for all the branching times will be uniformly distributed under Assumptions 3. This will allow us to use a Kolmogorov-Smirnov test to determine if the data fits the model, by comparing all the $B_0(v)$ as calculated using our maximum likelihood parameter values to the appropriate uniform distribution. The distribution of the D statistic, derived from this test, has already been calculated, so it is unnecessary to simulate trees. I calculated this statistic and its two tailed p-value for each tree using the maximum likelihood parameter values from each assumption. This method

Table 2.1: Birth-death maximum likelihoods parameter values, maximum likelihoods and statistical results for the Agamid phylogeny. The maximum likelihood birth rate (λ) and death rate (μ) are both in units of events/lineage/million years. p(simulation) is the fraction of 1000 simulations using the ML values for λ and μ that had likelihoods less than the actual data. D is the D statistic from the Komolgorov-Smirnov test and p (KS) is the two tailed p-value for this statistic.

	Assumption 1	Assumption 2	Assumption 3
λ	0.07349	0.07458	1.529×10^{-7}
μ	0	0	0.03094
Log Likelihood	-19.956	-18.325	-7.789
p (simulation)	0.004	<0.001	0.485
D	0.22919	0.23221	0.11520
p (KS)	0.002	0.001	0.336

could also be applied to Assumption 2, if the time at which the clade in question diverged from its extant sister clade was included in the data set.

For the Agamid tree maximum likelihood analyses under Assumption 1 and Assumption 2 generated similar parameter values (Table 2.1). Under both assumptions the estimate for μ was zero, and the estimates for λ differed by less than 1.5%. On the other hand under Assumption 3 the estimate of μ is more than 270 thousand times as great as λ. I consider the parameter estimates under Assumption 3 to be unrealistic, as the probability of one lineage at the base of this tree surviving to the present is less than 5×10^{-7}, and the probability of two reconstructed lineages at the base of this tree giving rise to more than 2 lineages in the present is less than 10^{-6}. Thus, unless we accept the highly unlikely proposal that this is one lineage of millions that produced 69 extant lineages, when the vast majority died, we must seek some other explanation. In fact the poor match between the parameter estimates under Assumption 3 and reality is indicative of these branching times not fitting the birth-death process. Indeed the simulated branching times had higher likelihoods than the actual branching times in 996 out of a thousand times for Assumption 1 and a thousand out of a thousand times for Assumption 2, and the Komolgorov-Smirnof test strongly rejected the CRBD under Assumptions 2 (p=0.002) and 3 (p=0.001) parameters . I could not reject the birth-death process under Assumption 3 by comparing the maximum likelihood to simulations or using the Komolgorov-Smirnov test, but given the extremely unrealistic parameters used for comparison these tests can not really be considered valid.

Estimates of λ for the Plethodon tree differed by less than 8% between all three assumptions (Table 2.2). Estimates of μ were more variable. I estimated μ as 72% of λ under Assumption 2, greater than 80% of λ under Assumption 1, and equal to λ under Assuption 3. Nee et al. (1995) also found that the maximum likelihood estimate of a was close to one for this phylogeny. We were unable to reject the CRBD for the Plethodon tree under any set of assumptions by comparing the maximum likelihood to the likelihood of

Table 2.2: Birth-death maximum likelihoods parameter values, maximum likelihoods and statistical results for the Plethodon phylogeny under all three strict assumptions. The maximum likelihood birth rate (λ) and death rate (μ) are both in units of events/lineage/million years. p(simulation) is the fraction of 1000 simulations using the ML values for λ and μ that had likelihoods less than the actual data. D is the D statistic from the Komolgorov-Smirnov test and p (KS) is the two tailed p-value for this statistic.

	Assumption 1	Assumption 2	Assumption 3
λ	0.15108	0.14099	0.14039
μ	0.12599	0.10087	0.14039
Log Likelihood	-25.442	-25.384	-21.569
p (simulation)	0.423	0.465	0.461
D	0.14750	0.13918	0.14193
p (KS)	0.621	0.690	0.667

simulations or using the Komolgorov-Smirnov test.

2.6.2 Visual Evaluations

Visual inspection is the primary method of analyzing lineage through time plots. The analyses described in the previous section give a good description of the fit of an entire data set to a model, but there is much information to be gained from visual inspection, which is lost in a single statistic. Here I will describe some tools to aid in the visual interpretation of lineage through time plots that are based on the distributions described in section 2.5. The first tool is the most obvious one, a comparison between the actual number of lineages at any given time and the distribution of those lineages under the birth-death process.

Figure 2.15 a, b and c show the distributions of the number of lineages over time under all three assumptions conditioned on the maximum likelihood parameters for the Agamid tree under each assumption. The lineage through time plot for the Agamid tree has been added to each plot for comparison. Under Assumption 1, the number of lineages exceed the 99th percentile of the distribution for most of the time that the data is analyzed. Under Assumptions 2, the number of lineages falls within the 95th percentile for most of the duration of the tree. However, the number of lineages does exceed the 99th percentile near the present. Under Assumption 3 the number of lineages never falls outside the 95th percentile. However, I have already stated that the maximum likelihood parameters under Assumption 3 are unrealistic, so I also compared the actual lineage through time plot to the distribution of the number of species under Assumption 3 using the maximum likelihood parameters that I derived under Assumption 1 (Figure 2.15d). In this case the number of lineages stayed within the 95th percentile until near the present when it exceeded the 99th, as under Assumption 2.

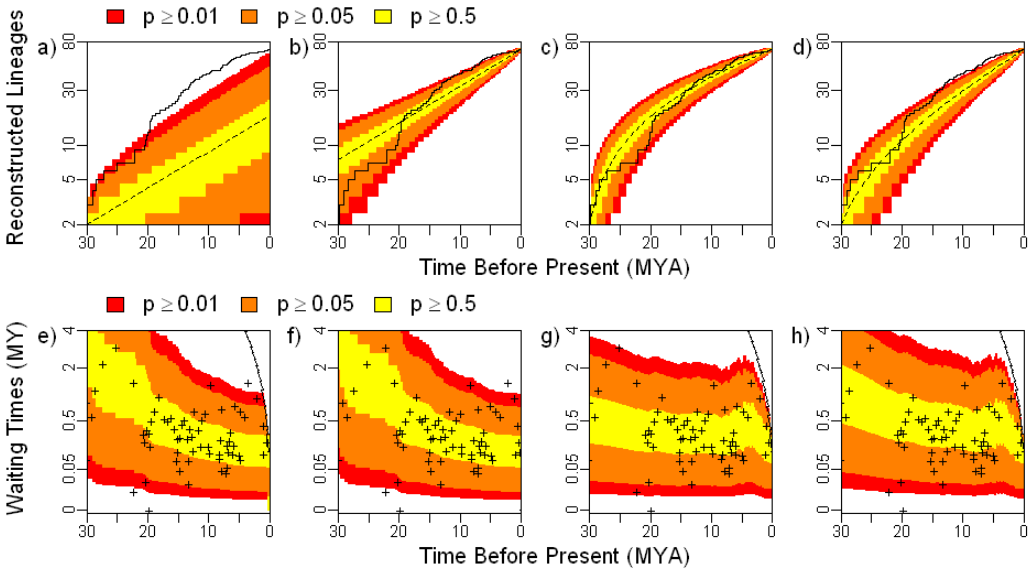

Figure 2.15: Comparison of the number of reconstructed lineages through time and waiting times between reconstructed lineage splits for a phylogeny of Australian agamids to their distribution under the rate constant birth-death process. (a through d) The lineage through time plot (solid line) for the Agamid phylogeny plotted against the distribution of number of lineages through time. (e through h) The waiting times (cross hairs) for the Agamid phylogeny plotted against the distribution of waiting times through time. Each plot uses a different assumption about the number of reconstructed lineages and the maximum likelihoods derived under that same assumption for (a and e) strict Assumption 1, (b and f) strict Assumption 2, and (c and g) strict Assumption 3; with the exception of plots (d and h) that used the assumptions of strict Assumption 3 and the maximum likelihood parameters derived under strict Assumption 1. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

These plots do help us to visualize the relationship between the actual lineage through time plot and the birth-death process, but they seem to be missing something fundamental to the relationship. For example it is not clear from Figure 2.15a why these are the maximum likelihood values, when a different set of parameter values would certainly increase the two-tailed p-values for the number of Agamid lineages under Assumption 1. Furthermore, under Assumption 2 the number of lineages moves from the bottom 5th percentile to the top 5th. Although being at either of these percentiles is acceptable at any given time, going from one to the other seems unlikely. It is important to remember that our maximum likelihood parameter values are based on the waiting times not the number of lineages. Waiting times are better expressed as the inverse of the slope of the lineage through time plots and over most of their durations the slope of our actual lineage through time plots and the expectation

under the birth death process do match up.

In order to explore this relationship we should compare our actual waiting times to the distribution of waiting times under the birth-death process. We could compare waiting times from a real phylogeny to the plots of waiting times as a function of time, which I introduced in subsection 2.4.4. However, when comparing the distribution of waiting times to an actual phylogeny we know not only the timing of the lineage split, but also the number of reconstructed lineages at that time. Therefore, we can calculate the distribution of each waiting time under the birth-death process using (2.43), (2.44) or (2.45) depending on what assumptions we make about the number of lineages. We can then plot certain quantiles of this distribution and compare those to our actual waiting times. I have implemented this plot in plot.wt.null and the calculation of the quantiles in q.wt.null for the telos R package.

Figure 2.15 e to h show these plots of waiting times juxtaposed against their distribution under the birth death process for the Agamid tree. These plots are similar to those introduced in subsection 2.4.4, except here the distribution is based on the number of reconstructed lineages at that time. Figure 2.15 e, f and g show the distribution under Assumptions 1, 2 and 3 respectively using the maximum likelihood parameters under the appropriate assumption, while Figure 2.15h shows the distribution under Assumption 3 using the maximum likelihood parameters from Assumption 1. The crosses show the actual waiting times from our tree. These plots clearly demonstrate the relationship between the actual data and the maximum likelihood parameters better than the distribution of the number of taxa at any time. Approximately half of the points fall inside the 50th percentile and the vast majority are inside the 95th and the 99th percentiles. Thus we can see how the maximum likelihood parameter values are a good fit.

However, what really makes these plots informative is that we can see where the actual data fails to fit the model distribution. Under Assumptions 1 and 2 the waiting times are expected to decrease as we approach the present, but the actual branching times do not (Figure 2.15 e and f). Thus more unexpectedly small waiting times are found early and unexpectedly large waiting times are found late. Under Assumption 3 the waiting times are not expected to decrease very much at all and thus excessively high and low waiting times are found throughout (Figure 2.15 g and h). Furthermore, although slightly more than half fall within the 50th percentile under all assumptions, more branching times fall outside the 95th and 99th percentile than we would expect.

There are four waiting times between 22.2 and 19.7 million years ago (MYA) that are excessively short. These include two waiting times of length zero at 19.7 MYA that appear as only a single cross. These are a consequence of a polytomy involving four lineages at the base of a clade. The authors do not state the origin of this polytomy (Harmon et al. 2003). We will assume for the purpose of this analysis that this is a hard polytomy and that the waiting times are actually zero, although it may in fact be a soft polytomy which would affect our analysis. The period between 22.2 and 19.7 MYA is associated with the large jump in reconstructed lineages that we observed in the lineage through time plots. Assuming that these waiting times are in fact accurate this probably represents a period of excessively high

speciation. Therefor an appropriate model should incorporate a high speciation rate during this period.

Under Assumptions 1 and 2 we see excessively long branching times near to the present (Figure 2.15 e and f). This may imply that a model with decreasing speciation rate, such as SPVAR may be more appropriate. On the other hand under Assumption 3 the branching times appear excessively long in the distant past (Figure 2.15 g and h). This is in keeping with the the overall pattern we see of decreasing expected branching times under Assumptions 1 and 2, and more constant branching times under Assumption 3. In fact we see many of the largest waiting times before 20 MYA as we would expect under Assumption 1 or 2, thus the best model may be one that starts with a normal set of rates then has a dramatic increase in the rate of speciation which declines as time goes on. Another possibility is that, as the longest and shortest waiting times are found early, the accuracy of the reconstructed times decreases the further back in time we go leading to increased variance in the past.

In contrast to the Agamids, the Plethodon phylogeny matches up well with the distributions of reconstructed lineages and waiting times that we would expect under the CRBD (Figure 2.16). This conforms well with the quantitative statistics that we evaluated in the previous section. The one exception is the wait between approximately 5.5 and 13.5 million years ago, when there are 11 reconstructed lineages. This wait is excessively long (one tailed $0.008 < p < 0.02$ for all assumptions) and leads to a subsequent drop in the expected waiting times. This observation is in contrast to the original authors who thought that the fast increase in the number of lineages near the present was remarkable (Highton and Larson 1979); we found this increase was well within the boundaries of our model. Nee et al. (1994a) did recognize that this fast increase was appropriate under the birth-death model, but also failed to recognize that the previous waiting time was excessively long. Overall this excessively long waiting time is probably not a violation that we need to be concerned with.

2.6.3 Comparing Models

One of the advantages of the methods described in this paper is that we are capable of calculating likelihoods and distributions for a variety of TVBD parameters. This allows us to compare how well different models fit the data. All the methods discussed in the previous two sections can be brought to bear on this question. We can compare the maximum likelihoods of the different models using the Akaike Information Criterion (AIC). The AIC is twice the negative log of the maximum likelihood plus twice the number of free parameters, and models with lower AICs are considered better fits for the data (Akaike 1974). We can also calculate the Komolgorov-Smirnov D statistic for the fit of $B_0(v)$ to a uniform distribution. The p-value for this test can not be trusted, as the distribution was chosen to fit the data, but it can be used to compare one distribution to another. We could also use the visual evaluations described above and compare the number of lineages and the waiting times between lineage splits to the expected distributions under different sets of parameter values. It is important to note that one can not make comparisons between different sets of

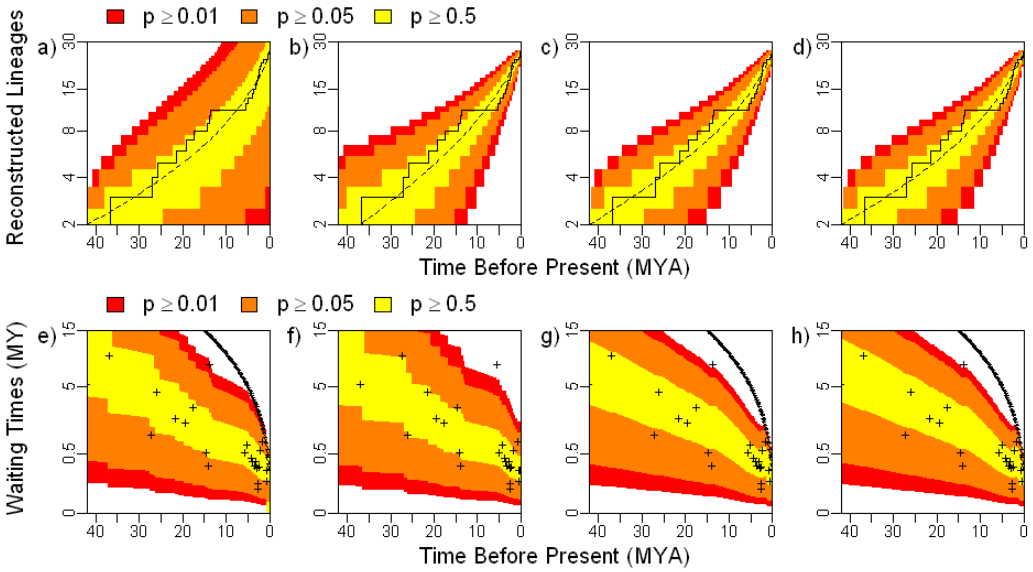

Figure 2.16: Comparison of the number of reconstructed lineages through time and waiting times between reconstructed lineage splits for a phylogeny of Plethodons to the distribution of those values under the rate constant birth-death process. a through d) The lineage through time plot (solid line) for the Plethodon phylogeny plotted against the distribution of number of lineages through time. e through h) The waiting times (cross hairs) for the Plethodon phylogeny plotted against the distribution of waiting times through time. Each plot uses a different assumption about the number of reconstructed lineages and the maximum likelihoods derived under that same assumption for a and e) strict Assumption 1, b and f) strict Assumption 2, and c and g) strict Assumption 3; with the exception of plots d and h that used the assumptions of strict Assumption 3 and the maximum likelihood parameters derived under strict Assumption 1. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

assumptions about the number of lineages, instead a given assumption is a precondition of any analysis. Here I will use strict Assumption 2 for all the models.

In order to investigate the effectiveness of this type of analysis, I have tested several different models to see how well they fit the branching times for the Agamid tree using Assumption 2 (Table 2.3). No one should do this type of mass testing of models in order to identify the processes that have generated a set of branching times. The purpose of this exercise is not to identify by what process the Agamids diversified, instead I did these analyses in order to show that a variety of TVBDs can be fit to a data set and to investigate how each models fits that data. In the future other researchers could use these same methods to investigate a hypothethis that they have a strong a priori reason to believe.

I investigated several models that would classify as DTBDs. The first model, BD_1, is the

Table 2.3: Time Variable Birth-Death Models Used for Comparison

Model	Parameters	Time Period	Splitting Rate	Loss Rate
BD_1	λ, μ	∞ to 0	λ	μ
BD_{21}	λ_2, μ_2, λ_1, μ_1, t_1	∞ to t_1	λ_2	μ_2
		t_1 to 0	λ_1	μ_1
BD_{321}	λ_3, μ_3, λ_2, μ_2,	∞ to t_2	λ_3	μ_3
	λ_1, μ_1, t_2, t_1	t_2 to t_1	λ_2	μ_2
		t_1 to 0	λ_1	μ_1
BD_{121}	λ_2, μ_2, λ_1, μ_1,	∞ to t_2	λ_1	μ_1
	t_2, t_1	t_2 to t_1	λ_2	μ_2
		t_1 to 0	λ_1	μ_1
Y_1	λ	∞ to 0	λ	0
Y_{21}	λ_2, λ_1, t_1	∞ to t_1	λ_2	0
		t_1 to 0	λ_1	0
Y_{321}	λ_3, λ_2, λ_1, t_2, t_1	∞ to t_2	λ_3	0
		t_2 to t_1	λ_2	0
		t_1 to 0	λ_1	0
Y_{121}	λ_2, λ_1,t_2, t_1	∞ to t_2	λ_1	0
		t_2 to t_1	λ_2	0
		t_1 to 0	λ_1	0
BD_{S0}	λ, μ, p	∞ to 0	λ	μ
		0	0	$1-p$
BD_{S1}	λ, μ, p, t_1	∞ to t_1	λ	μ
		t_1	0	$1-p$
		t_1 to 0	λ	μ
$SPVAR_1$	λ_0, μ, k	∞ to 0	$\lambda_0 e^{kt}$	μ
$SPVAR_2$	λ_2, μ_2, λ_0, μ_1,	∞ to t_1	λ_2	μ_2
	k, t_1	t_1 to 0	$\lambda_0 e^{kt}$	μ_1

CRBD, which we already compared to the Agamid tree. I also evaluated a DTBD, BD_{21}, with two time periods each with parameters independent of the other, and a third DTBD, BD_{321}, with three time periods with independent parameters. A fourth DTBD, BD_{121}, has three time periods but λ and μ from the first time period are equal to λ and μ from the third, so that only the parameter values from the middle period are independent. I also included four other DTBDs, Y_1, Y_{21}, Y_{321} and Y_{121}, which are the same as the respective models above, except that μ is constrained to be zero.

In order to demonstrate how these methods can be used to investigate a larger variety of models, I also included four models with either sampling or continuously varying parameter values. BD_{S0} is a CRBD, however there is a free parameter, p, that determines the fraction of lineages randomly sampled for our study. BD_{S1} is the same as BD_{S0}, except the time at which the sampling occurred is also allowed to vary, so that it represents a mass extinction. $SPVAR_1$ is the same as the SPVAR model described before, in which λ decreases exponentially as we approach the present (Rabosky and Lovette 2008a). Under $SPVAR_2$ we divide the process into two time periods. During the first time period the parameters are constant, while during the second λ changes according to SPVAR.

Parameter values for all models were fit by maximum likelihood using the ml.bd function from the R telos package. The maximum likelihood value was used to calculate the AIC and the maximum likelihood parameter values were used to calculate the D statistic and make the null lineage through time plot and the null waiting times plot. I should reiterate here that this is an exercise to investigate how models fit the data not to discover which model fits the data best. The haphazard choice of models used here is not appropriate for a hypothethis test.

Table 2.4 shows the maximum likelihood, AIC and D statistic for each model. There are two patterns that immediately emerge from these numbers. The first is that the CRBD and Yule processes are poor fits for the data as judged by their low maximum likelihoods and high AICs and Ds. The second observation is that just about any model with time variable parameters leads to a large improvement in the fit of the model to the data. This suggests that users of these methods should be cautious in choosing their models, as any randomly chosen model may be a good fit. Model choice should be hypothethis driven. I will now use plotting tools to show how each of these models fits the data.

BD_1 is the CRBD and it produces reasonable parameter values with a μ equal to zero (Figure 2.17a). The other three DTBDs predict unbelievably high values for both λ and μ, especially BD_{321} under which λ and μ are both over 12 events/lineage/million years (ELMY) for most of the duration of the process (Figure 2.17b,c,d). These parameter values lead to impressive contortions of the lineage through time null plots (Figure 2.17 e through h), that allow the model to be closely fit to the data (Figure 2.17 i through l). The unrealistic parameter values in combination with the over tuned null plots make me highly skeptical about the use of these models. These models produce the lowest AICs (Table 2.4), but I would recommend against using them to analyze any data set unless they are supported by an explicit a priori hypothethis and even then one may want to constrain the parameter

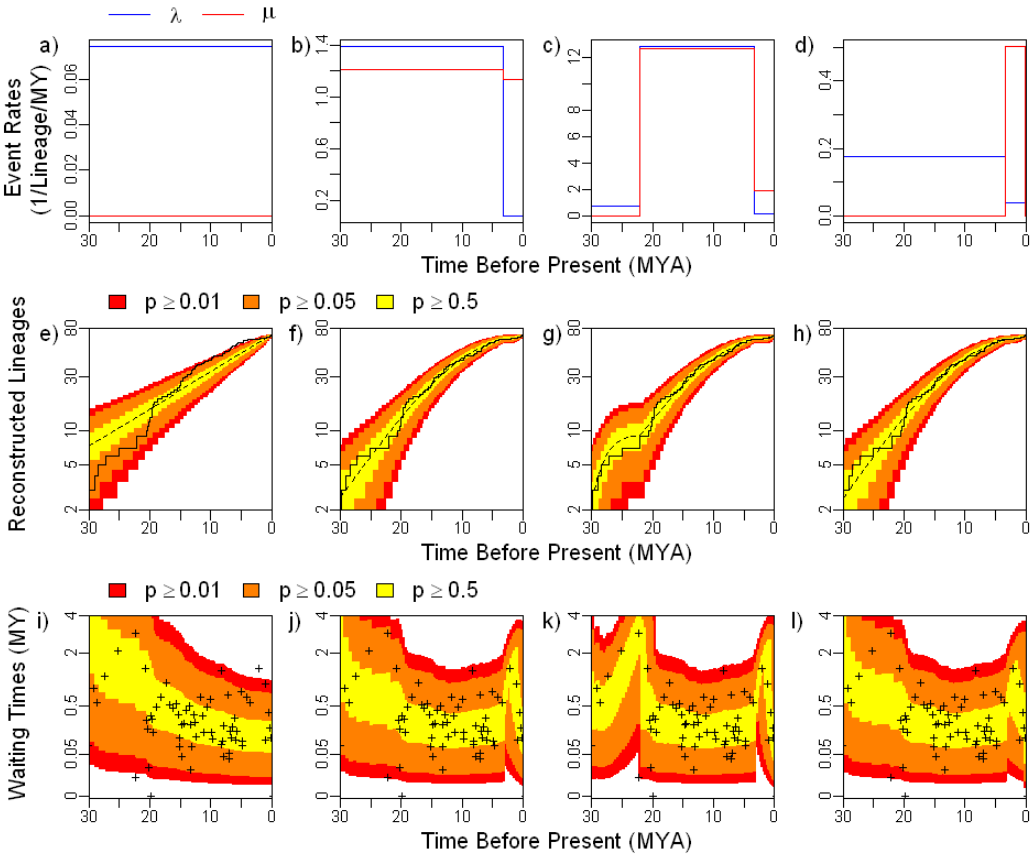

Figure 2.17: Maximum likelihood parameters and lineage through time and waiting time null plots for each of the four DTBD models for the Agamid phylogeny. These plots use the strict Assumption 2 for (a, e and i) BD_1, (b, f and j) BD_{21}, (c, g and k) BD_{321}, or (d, h and l) BD_{121} (see Table 2.3). (a through d) The maximum likelihood values of λ and μ over time for each DTBD. (e through h) The lineage through time plot for the Agamid tree (solid line) compared to the distribution of reconstructed lineages over time for each model. (i through l) The waiting times (cross hairs) of the Agamid phylogeny compared to their distribution over time for each model. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

Table 2.4: Maximum likelihood, Akaike information criterion and Komolgorov-Smirnov D for the Agamid phylogeny when compared to multiple different models

Model	Log Likelihood	AIC	KS D
BD_1	-18.33	40.65	0.19583
BD_{21}	-4.33	18.65	0.06528
BD_{321}	-0.39	16.77	0.0631
BD_{121}	-4.68	21.36	0.06203
Y_1	-18.33	38.65	0.19583
Y_{21}	-7.75	21.50	0.09256
Y_{321}	-3.93	17.85	0.06881
Y_{121}	-7.25	22.49	0.12919
B_{S0}	-7.18	20.36	0.06181
B_{S1}	-6.56	21.13	0.05487
$SPVAR_1$	-7.10	20.20	0.06332
$SPVAR_2$	-4.26	20.53	0.0647

values further.

The fit of the four Yule models, Y_1, Y_{21}, Y_{321} and Y_{121} is more believable than for those models in which μ is also allowed to vary (Figure 2.18). The parameter values are more reasonable (Figure 2.18 a through d), and we can see that in general these models support a λ that decreases in time leading to an overall concave shape for the log lineage through time plots (Figure 2.18 e through h). Furthermore the contortions of the null waiting time plots are not too extreme (Figure 2.18 i through l). Yet these models are able to produce AICs almost as low as the AICs under the models in which μ also varies (Table 2.4). I would still caution against using these models without an explicit hypothethis. As we have seen any model with time variable parameters can lead to a huge improvement in fit, so the success of these models does not necessarily indicate that they are appropriate. Finally it does not seem realistic to constrain μ to be zero, it is highly unlikely that no lineages died during this process. A model in which μ is free to vary but remains constant with time may be more realistic.

The maximum likelihood parameter values for BD_{S0} are not substantially different from those for BD_{S1} (Figure 2.19 a and b). The lineage sampling event under BD_{S1} occurred just 137 thousand years ago, at the time of the last lineage split. Slightly more lineages were sampled under BD_{S1}, and the values of both λ and μ were lower, although r, the difference between them, was essentially unchanged. Both models fit the data well in comparison to the other TVBDs (Table 2.4 and Figure 2.19 c through f). The waiting time null plots are concave and the log lineage through time null plots are convex, matching the patterns seen in the actual phylogeny. It seems more reasonable to include a sampling factor in an analysis than the time variable parameters discussed before. All real phylogenies represent only a

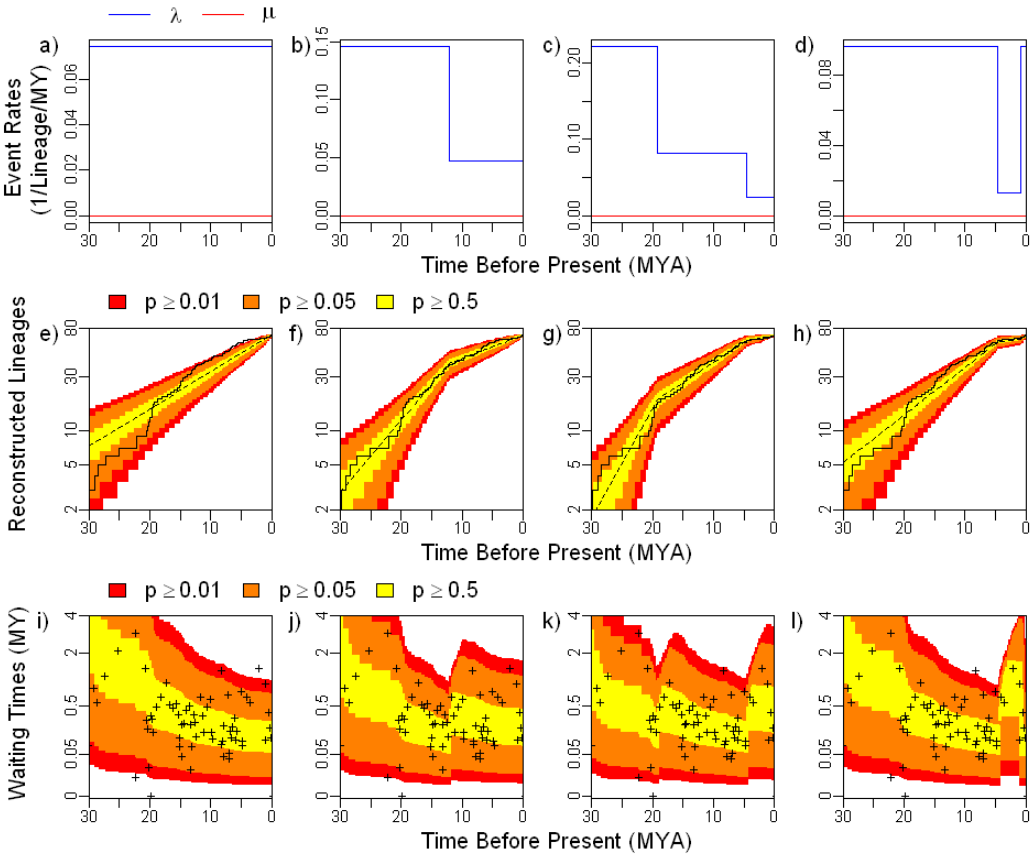

Figure 2.18: Maximum likelihood parameters and lineage through time and waiting time null plots for each of four discrete time Yule models. These plots use the strict Assumption 2 for (a, e and i) Y_1, (b, f and j) Y_{21}, (c, g and k) Y_{321}, or (d, h and l) Y_{121} (see Table 2.3). (a through d) The maximum likelihood values of λ and μ over time for each DTBD. (e through h) The lineage through time plot for the Agamid tree (solid line) compared to the distribution of reconstructed lineages over time for each model. (i through l) The waiting times (cross hairs) of the Agamid phylogeny compared to their distribution over time for each model. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

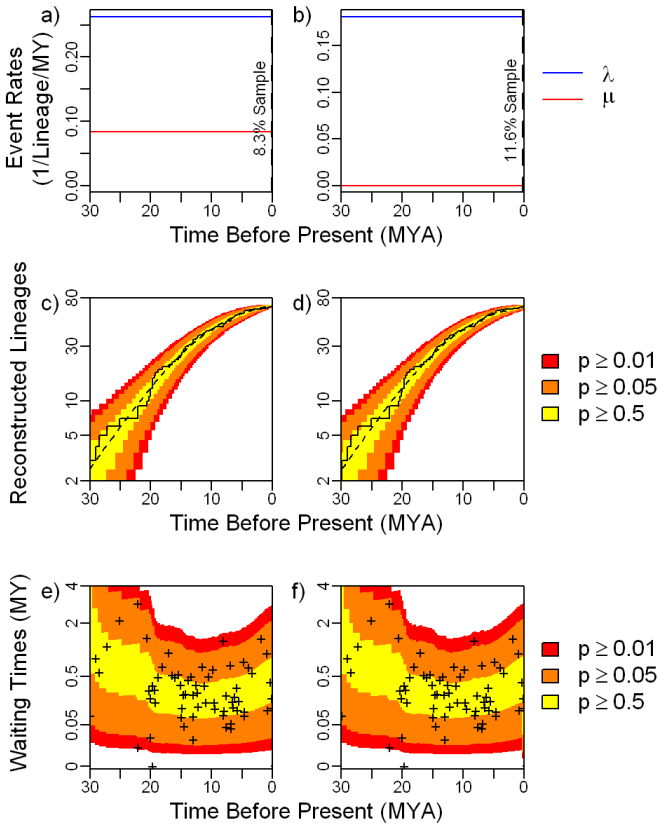

Figure 2.19: Maximum likelihood parameters and lineage through time and waiting time null plots for each of two CRBD models with lineage sampling. These plots use the strict Assumption 2 for (a, c and e) BD_{S0} or (b, d and f) BD_{S1} (see Table 2.3). (a through d) The maximum likelihood values of λ and μ over time and the timing and magnitude of the sampling event for each TVBD. (e through h) The lineage through time plot for the Agamid tree (solid line) compared to the distribution of reconstructed lineages over time for each model. (i through l) The waiting times (cross hairs) of the Agamid phylogeny compared to their distribution over time for each model. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

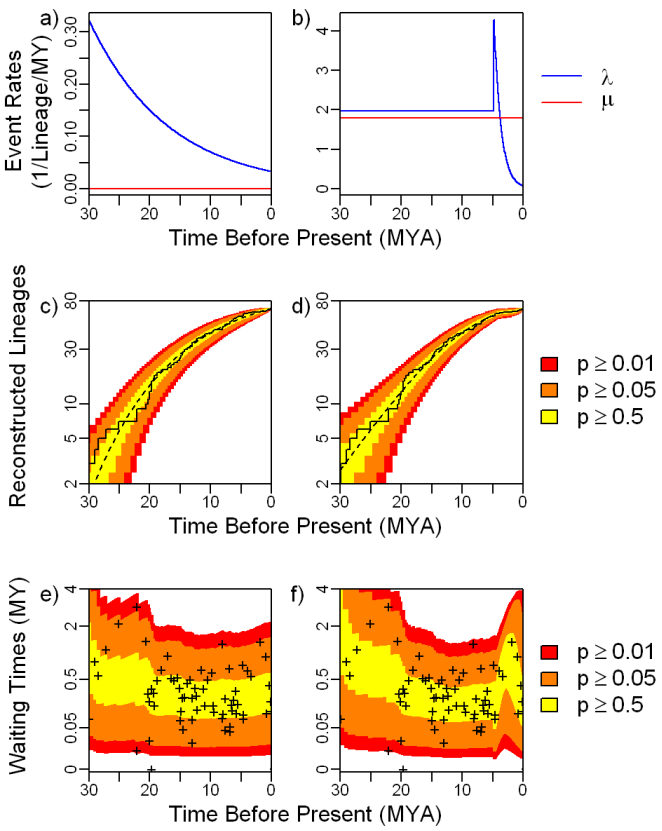

Figure 2.20: Maximum likelihood parameters and lineage through time and waiting time null plots for each of two SPVAR models. These plots use the strict Assumption 2 for (a, c and e) SPVAR_1 or (b, d and f) SPVAR_2 (see Table 2.3). (a through d) The maximum likelihood values of λ and μ over time for each TVBD. (e through h) The lineage through time plot for the Agamid tree (solid line) compared to the distribution of reconstructed lineages over time for each model. (i through l) The waiting times (cross hairs) of the Agamid phylogeny compared to their distribution over time for each model. The colored areas show two-tailed percentiles of the distribution at each time and the dashed line is the expectation.

sampling of the extant taxa and the effects of that sampling should be considered. It is possible that these trees represent only 10% of the extant diversity in this clade; however, the authors stated that this phylogeny represented 87% ofthe extant diversity in the clade (Harmon et al. 2003), so sampling alone is unlikely to explain the pattern we see here. Another possibility is that this sampling could actually represent a recent mass extinction in this clade, although without further evidence such a conclusion is tenuous at best.

The maximum likelihood parameter values for SPVAR_1 projected a μ of zero and a λ

that started at about 0.32 ELMY and declined exponentially to about 0.03 ELMY in the present (Figure 2.20a). $SPVAR_2$ started with λ at almost 2 ELMY and then stayed at that level until about 5 MYA when it shot up to about 4 ELMY and then rapidly declined to about 0.03 ELMY in the present (Figure 2.20b). Both models are good fits for the Agamid phylogeny (Table 2.4 and Figure 2.20 c to f). The null lineage through time plots are concave. The null waiting times plot for $SPVAR_1$ is mostly flat but has a slight negative slope and a slight positive curvature, matching the data reasonably well. The initial constant λ value under $SPVAR_2$ leads to a steady decrease in the waiting times; after the λ increases there is a bump in the null waiting times plot that clings tightly to the data points. There is no solid a priori reason to go around fitting SPVAR models to data sets. It produces concave log lineage through time plots, but this alone is not sufficient to justify its use, as other models can produce concave plots as well. However, if one has good reasons to suspect an adaptive radiation, then it is a reasonable model to consider.

2.7 Discussion

I have shown how to calculate the density, cumulative distribution and quantiles of numbers of reconstructed lineages and waiting times between reconstructed branching events for any time variable birth-death process. I also showed how to generate a random set of branching times under this process. Furthermore I showed how to calculate all these values so long as we assume that we know the number of reconstructed lineages at any time, no matter how many reconstructed lineages there are at that time. I have defined the discrete time birth-death process and shown how it can be used as a simple numerical solution to any birth-death process in which the birth-death parameters vary with time. Finally I developed a number of tools to explore the effects of different parameters on a time variable birth-death process, and to compare actual data to such a process.

The discrete time birth-death process is one in which the parameters of the model are constant over periods of time and then change instantaneously at certain times. Kendall (1948) provided a set of equations that describe any time variable birth death process. He based those equations around the variables $B_j(0)$ and $E_j(0)$, which he called η_{t_j} and ξ_{t_j} respectively, and provided general formulas to solve for those variables. However, he did not discus the discrete time process or attempt to solve $B_j(0)$ or $E_j(0)$ for that process. Rabosky (2006a) did attempt to derive the probability of a set of branching times under the discrete time birth-death process. However, he failed to account for the way in which subsequent changes in λ and μ would affect the probability of a lineage being lost and in turn affect the waiting time. Thus his equations are only appropriate under the Yule model when the probability of a lineage being lost is always zero.

One of the greatest advantages of the discrete time birth-death process over previous implementations of the TVBD is its great flexibility. It can work as a simple and accurate numerical solution to any time variable birth-death process. A numerical solution to these

equations is not always necessary as for some time variable processes the equations given by Kendall (1948) for $B_j(0)$ and $E_j(0)$ can be solved analytically. Even if the equations can not be solved analytically, we could generate a numerical solution to Kendall's equations. However, the discrete time birth-death process has two distinct advantages. It fits into a frame work that I have already established and thus requires no additional calculations or programming and it is trivial to solve for the inverse. Solving for the inverse of these functions is critical for simulating trees and calculating quantiles, and would be difficult for numerical solutions to Kendall's equations.

The discrete time birth-death process can also be used to incorporate random taxon sampling or mass extinctions into an analysis. Stadler (2008) gave some formulas for random sampling when we know the exact number of lineages sampled. Yang and Rannala (1997) and Stadler (2010) gave formulas for random sampling under the CRBD, when starting with one lineage. My results are appropriate for any TVBD and any assumption about the number of lineages at any time. In reality most sampling is non-random and tends to be biased towards incorporating at least one member of all the extant subclades. This leads to a higher retention rate for deeper nodes, which may decrease the curvature of lineage through time plots more than random sampling (Cusimano and Renner 2010). As far as I am aware this is the first attempt to analytically incorporate mass extinctions into the birth-death process. Previous attempts have done so only through simulation (e.g. Crisp and Cook 2009).

I have generalized all the calculations for the reconstructed birth-death process, so that they can be made for any time variable process with any set of assumptions about the number of reconstructed lineages at any time. Nee et al. (1994b) provided versions of several of these equations that work for any time variable birth-death process, but only when we assume that there were one or two lineages at some time in the past. Stadler (2008) calculated the distribution of the number of lineages when we know the number of lineages at some time both before and after the time in question, but only for those cases in which λ and μ are constant and the process starts with one or two lineages and ends in the present. Aldous and Popovic (2005) and Stadler (2008) both provided equations for the density of the time of origin for our process, equivalent to my (2.39). However, both limited their results to the rate constant birth-death process when we know the number of lineages in the present. It is interesting that I got the same results as Aldous and Popovic (2005) and Stadler (2008) for this density, as we made different assumptions about the nature of the origin. These authors assumed that there was one lineage evenly distributed between the present and the infinite past, and that the process started at this point. I assumed that the divergence of the analyzed clade from its extant sister clade was distributed according to the probability of a lineage arising at any time from a single lineage and surviving to the present, and that the reconstructed process would begin after this point. The fact that such different assumptions lead to the same conclusions serves as further justification of that conclusion. My interpretation of the origin corresponds to an actual event, the divergence of the studied clade from its extant sister clade, therefore researchers could gather these data and include

them in their analyses. Stadler (2008) also attempted to calculate the distribution of the number of reconstructed lineages at some time when we know the number of lineages in the present, and she provided the integral found in (2.42), but she was unable to solve it.

Until now I have left open ended the question of which assumption about the number of lineages is appropriate for an analysis. The simple answer is that it depends on the question a researcher is asking. If you want to know about the distribution of lineages or branching times for taxa before the Mesozoic, then you should use Assumption 2 with the number of lineages fixed at the end of the Paleozoic. On the other hand if you want to know about the distribution of clade sizes for clades that started in the Eocene, then you should use strict Assumption 1 starting in the Eocene. So what about a researcher who has a phylogeny and just wants to analyze the distribution of lineages through time for that phylogeny. In some ways it seems that Assumption 3 best captures the intent of the researcher in this situation, as it asks about the distribution of reconstructed lineages without being concerned with the size of the clade or its time of origin. However, as we saw in subsection 2.6.1, this lack of concern can lead to extremely unrealistic parameter estimates. So we should probably limit ourselves to Assumption 1 or 2, at least for parameter estimation, while Assumption 3 can still be used for model comparison. In the past most researchers used Assumption 1, because it was the only option available to them, but I believe that most researchers are studying clades of a given size more so than clades with a most recent common ancestor at a given time. Therefore, in most cases Assumption 2 better represents the state of the researchers knowledge at the beginning of an analysis and thus is the more appropriate assumption.

I have used the discrete time birth-death process to show how to simulate trees under any time variable birth-death process and any set of assumptions about the number of reconstructed lineages at different times. It is trivial to simulate trees of a given depth with an unknown number of species for any process by simply running the process forward until the tree is of the desired length. We can accomplish this same result using Assumption 1, but this is not a great improvement over previously existing methods. Felsenstein (2004) showed how to simulate trees of a given length and a given number of final taxa for a constant rate birth-death model based on the work of Rannala (1997). Here I essentially expanded that result to incorporate a time variable process using Assumption 3 and the discrete time process. Several authors have described a method of simulating trees with a known number of terminal lineages and an unknown length by running the process forward until they achieved the appropriate number of lineages (see Rambaut 2002; Harmon et al. 2008). However Hartmann et al. (2010) showed that this method produces a biased set of branch lengths if μ is greater than zero. They also provided a set of highly inefficient methods for simulating trees of a given size that will work for any birth-death process, and an efficient method for simulating such trees under the constant rate birth-death process. I have expanded this efficient method to include the time variable birth-death process by using Assumption 2 and the discrete time process. However, it would still be necessary to use their inefficient method for any situation in which the rate of splitting or loss varied between lineages. It should be noted that under the time variable process there is a difference

between simulating a tree that starts at a certain time and a tree that ends at a certain time, as the relative timing of the birth-death parameters will differ. The method described by me is appropriate for the case in which the process ends at a certain time; Hartmann et al.'s (2010) inefficient method should be used for the time variable process that starts at a certain time. However, it seems like processes that end at a certain time are more relevant, as all real phylogenies end in the present.

I developed a method of comparing different time variable birth-death models to an actual phylogeny using a likelihood score based on the work of Rabosky (2006a,b) and Rabosky and Lovette (2008b). However, I expanded that work to include any time variable birth-death model, sampling and mass extinctions using the discrete time process. I also showed how to make those calculations for any assumption about the number of reconstructed lineages. Furthermore, I corrected the calculation of the likelihood under the discrete time process (Rabosky 2006a).

I innovated a pair of complementary visual tools, the lineage through time null plot and the waiting times null plot, for investigating the effects of different time variable birth-death models, and for comparing them to real phylogenies. These tools have a distinct advantage over the likelihood methods in that the likelihood methods provide only a single statistic for the comparison of real data to a model, while these two null plots allow an investigator to see where and how their data departs from the predictions of a model. The lineage through time null plot allows a researcher to see not just if the branching times of their phylogeny has diverged from the expectation of a birth-death model but how significant that divergence is. In the last few years several authors have simulated a number of phylogenies under a specific time variable model and overlaid the lineage through time plots for these simulations in order to get a rough idea of what lineage through time plots one should expect from such a model (Rabosky and Lovette 2008a; Crisp and Cook 2009). With the lineage through time null plot it is unnecessary to generate these simulations and it is easier to visualize the distribution. As far as I am aware the waiting times null plot is the first attempt to visualize waiting times. It is much easier to identify when exactly violations of the model occur using waiting time plots than lineage through time plots, as the waiting times are not autocorrelated.

It is easy to infer the effects of different TVBD models using the lineage through time null plots. I quickly confirmed the results of many previous researchers about the effects of different TVBD models on the curvature of lineage through time plots. My lineage through time null plots reinforced the observations of Nee et al. (1994a) that lineage through time plots have more positive curvature when the lineage loss rate is higher (Figures 2.4, 2.5 and 2.6), and more negative curvature when the sampling rate is lower (Figure 2.10). I also demonstrated that lineage through time plots have more negative curvature when the lineage gain rate decreases with time (Figures 2.9 and 2.12), but more positive curvature when the lineage loss rate increases with time (Figure 2.13), as originally suggested by Rabosky (2006a). Furthermore, I made the additional, although probably obvious, observation that lineage through time plots have more positive curvature when the lineage gain rate increases with time (Figure 2.9). On the other hand, I found that mass extinctions alone can not

explain an anti-sigmoidal distribution of lineages through time (Figure 2.11), in contrast to the conclusions of Crisp and Cook (2009). Such a complex curve would probably require a more complex model. It should be noted that although any of these phenomena can affect the curvatures of lineage through time plots, there is a wide range of fairly reasonable curvatures under any TVBD model; thus the importance of comparing real data to a distribution and not just the expectation.

It is unlikely that the CRBD is a completely accurate representation of organismal diversification. However, it is a good null model for investigating changes in the rate of lineage loss or gain. As the CRBD assumes that there are no changes in the rate of diversification, we can interpret clear violations of that model as evidence that rates have in fact changed. We can also use a time variable model to see what effects we would expect certain biological phenomena to have on the shape of phylogenetic trees.

Rates of diversification can vary either temporally or phylogenetically. Phylogenetic changes are usually a consequence of changes in an organisms biology and they lead to clades with rates of diversification that differ from those of their relatives. Researchers must use both topological and branch length data to investigate such processes, and several methods have been developed recently to identify phylogenies with phylogenetically varying rates of diversification (see Agapow and Purvis 2002; McConway and Sims 2004; Moore and Donoghue 2007) and to detect correlations between diversification rates and biological characters (see Maddison et al. 2007; Paradis 2005). On the other hand, temporally varying diversification rates tend to be a consequence of changes in the environment that lead to shifts in the diversification process in a number of lineages simultaneously or of macroevolutionary interactions between lineages such as competition. The study of temporal shifts in diversification rates only requires timing data as the topology will be unaffected (Thompson 1975; Sanderson and Bharathan 1993); however, it would be difficult to distinguish phylogenetic shifts from purely temporal shifts using only temporal data, as phylogenetic shifts will lead to temporal shifts as well. Furthermore, shifts that affect every lineage in a clade simultaneously may also have a phylogenetic pattern, such as when increased speciation in one lineage leads to increased extinction in its close relatives or when a mass extinction affects organisms differently depending on their heritable biological characters. The methods described in this paper assume that their is no phylogenetic effect on diversification rates; this is probably not the case (Savolainen et al. 2002; Blum and Francois 2006). However, even for changes in the pattern of diversification that have a phylogenetic component attempting to identify the temporal signal of this change can be informative.

Estimates of μ for phylogenies under the birth-death process have a large variance (Nee et al. 1994a,b). Rabosky (2010) suggested that estimates of μ may not just be poor but positively biased by phylogenetic variation in diversification rates. We saw in subsection 2.6.3 that allowing μ to vary between different periods can lead to extremely unrealistic maximum likelihood estimates. One solution to the problem of μ estimation is to use the Yule model and assume that your estimate of λ, the lineage splitting rate, is in fact r, the diversification

rate (Paradis 1997). However, this can lead to poor estimates of the diversification rate (Nee 2001). We have seen from our own results that high μ values alone can lead to concave lineage through time plots without requiring any temporal shifts in parameters, as originally pointed out in Nee et al. (1994b). A more practical approach is to attempt to fit a birth-death model using a number of different reasonable values for μ, rather than relying on a single estimate.

One potential risk of the tools described in this paper and available through the R package telos, as well as those made available in LASER (Rabosky 2006b) is that researchers can easily test the fit of many models to their data. Without some sort of a priori hypothethis about what models one might expect, it is not appropriate to go around testing multiple hypotheses and simply picking the best fit (see Freckleton 2009). Some authors have already done this using LASER (e.g. Valente et al. 2010). On the other hand others have provided good examples of how to use these models to test a clearly stated hypothethis about diversification (e.g. Steeman et al. 2009; Egan and Crandall 2008). Furthermore, some authors have used many models to show that in general diversification is slowing down or speeding up without committing to any one model (e.g. Burbrink and Pyron 2010; Dolman and Hugall 2008). In such a case Bayesian model averaging (see Link and Barker 2006) would probably be more appropriate than AIC. Of course an increase or decrease in the rate of diversification could also be identified through simple visual inspection. It may be reasonable to compare multiple models in order to see weather the data fits the CRBD. However, there are many simpler tests that can determine the same thing, such as the gamma statistic (Pybus and Harvey 2000) and the tests described in subsection 2.6.1 and subsection 2.6.2. Rabosky (2006a) showed that his test for changes in the diversification rate was more powerful than the gamma statistic. A comparison between type I and II error for all these tests would be helpful in determining the best way to test for violations of the constant rate birth-death process.

It goes without saying that the interpretations of branching time data using these methods depend heavily on the accuracy of the data. In any phylogeny both the topology and the branch lengths may be improperly estimated. The amount of difference between genetic sequences found in the terminals of phylogenetic trees provides some information about the relative length of branches (Hillis et al. 1996); but rates of sequence evolution vary between lineages and so, much of that information is lost (Gillespie 1994). How exactly to recover that information is a matter of great debate. One option available to researchers is to use an estimate of the branching times that incorporates some measure of error (see Drummond et al. 2002). It would be easy to incorporate error into the visual methods described in this paper. Furthermore, it is possible to use these methods without scaling the branches in absolute time, for many purposes a relative scaling will suffice.

The equations and methods found in this paper serve as a broad generalization of previous work done on the description of the time variable and reconstructed birth-death processes. Furthermore the application of the discrete time birth-death process as a numerical solution to any time variable process makes it relatively straight forward and easy to implement

these methods. Future research can use these tools both to explore data and to deduce how different time variable processes affect the distribution of branching times on a phylogeny. With the implementation of these methods in the telos R package other researchers should be able to benefit from the incites they provide and the flexibility that they allow.

Chapter 3

Comparative Analysis of Chromosome Counts Infers Three Paleopolyploidies in the Mollusca

3.1 Introduction

Polyploidy has long been recognized as an important mechanism of genetic evolution (Taylor and Raes 2004), and over the last decade, as full genome sequences have become available, a great deal of research has been invested into the analysis of whole genome duplications (e.g. Semon and Wolfe 2007; Byrne and Blanc 2006). Polyploid species and individuals are common in plants (Wendel 2000); evidence is accumulating that whole genome duplications also occur in the opisthokonts and have played an important role in the evolution of their genomes (Vandepoele et al. 2004; McLysaght et al. 2002; Wolfe 2004). In contrast to other modes of genome evolution, polyploidy events affect the entire genome at once. By duplicating every gene in the genome a large amount of redundant genetic information is created, which can be used as raw material for evolutionary innovations (Ohno 1967; Haldane 1932). It has been suggested that in several cases the modification of this raw material has been important in the evolution of key innovations, such as glucose fermentation in yeast (Piskur 2001) and the immune system of vertebrates (Kasahara 2007).

Despite the large effort put into the analysis of genome duplications, the identification and confirmation of such duplications, especially ancient ones, has proved problematic. To conclusively demonstrate a paleopolyploidy event, several complete genomes must be sequenced both from taxa that are directly descended from the original polyploid individual and from their relatives that diverged shortly before the whole genome duplication (Wong et al. 2002; Woods et al. 2005). Even to identify likely cases of paleopolyploidy large numbers of genes must be sequenced in several related taxa (Blanc and Wolfe 2004; Spring 1997). Although getting cheaper by the day, such sequencing is still costly in terms of lab time and

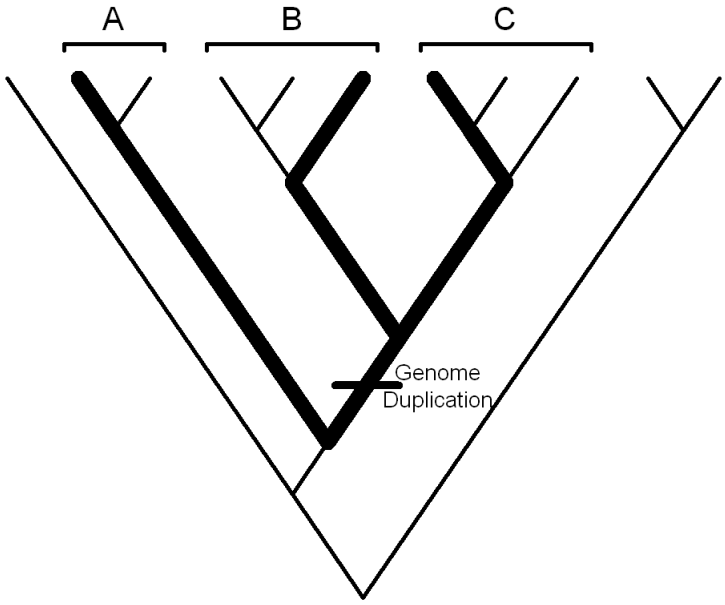

Figure 3.1: An ideal sampling of three taxa for investigating the genome duplication indicated by the dash. By selecting one taxon each from clades A, B and C the investigator would minimize the amount of shared history either before or after the event and thus have the best chance of accurately reconstructing the effects of the duplication on the genome.

materials (Ansorge 2009).

In order to study any evolutionary transition it is best to identify lineages that diverged as shortly before and after the evolutionary event as possible (Figure 3.1). Information regarding the exact state of the organism before and after the transition is lost as time passes. By choosing taxa that diverged shortly after such an event the amount of shared information loss is minimized, and researchers are better able to reconstruct the genome of the organism immediately after duplication. Similarly, choosing taxa that diverged as shortly before an event as possible allows us to more accurately reconstruct the state of the organism before undergoing the transition.

It would be beneficial to identify the exact phylogenetic position of a whole genome duplication, before embarking on a course of research into the effects of paleopolyploidy. In the past these events have initially been identified haphazardly as an unintended consequence of a more general investigation into the genomes of the organisms in question (Wolfe and Shields 1997; Lundin 1993). Furthermore the identification of the branch on which the duplication has actually occurred can only be achieved through the unguided sequencing of the relatives of these taxa (e.g. Irvine et al. 2002). This approach is highly inefficient for

the study of individual genome duplications, and untenable for a program of research into genome duplications in general.

Here we demonstrate a method for identifying the phylogenetic position of genome duplications using only chromosome counts and a well resolved phylogeny. A genome duplication will double the number of chromosomes in a genome. If the rate of aneuploidy fixation is sufficiently low relative to the amount of time since the duplication occurred, then this doubling should leave a signal that will be detectable in the extant descendants of the original polyploid organism. Indeed Ohno (1970) originally proposed that there was a whole genome duplication in a vertebrate ancestor based on placental mammals having genomes two to tree times as large as *Ciona intestinalis*. By using karyotypic data we are able to increase taxon sampling in order to accurately identify the branches on which the duplication occurred and thus guide the selection of taxa for genome sequencing.

We used the birth death process as a null model for the distribution of chromosome numbers among taxa. The birth death process is a stochastic model often used to compare the numbers of genes in gene families between organisms (Novozhilov et al. 2006; Lynch and Conery 2003; Hahn et al. 2005). Under this process the rate of chromosome duplication or loss is proportional to the number of chromosomes in the genome. We developed a second model in which the number of chromosomes could double in a single stochastic event in addition to evolving by the birth death process. The maximum likelihoods were calculated under each model and used to compare the fit of the data to the models. Furthermore the likelihoods calculated under the duplication model were used to calculate the posterior probability that a duplication occurred on each branch of the tree. By using a likelihood model we were able to compare the background rate of change in chromosome number by the birth-death process to doublings caused by whole genome duplications on certain branches. Mayrose et al. (2010) recently published a similar likelihood method that assumed that genomes added or lost chromosomes at a constant rate or at a rate linearly related to the number of chromosomes in the genome.

We demonstrate the utility of our model on a phylogeny of Molluscan taxa. Mollusca is a large and disparate clade with members that play an important role in marine and terrestrial ecosystems. Furthermore they are one of the most diverse groups within the Lophotrochozoa, the least studied of the three major clades of Bilaterian animals. Currently genomics studies in Mollusca are still in their infancy; only one mollusc genome has so far been sequenced (Chapman et al. 2007). Natural polyploid species of molluscs have been recognized for decades (e.g. Patterson 1969; Goldman et al. 1983), but as of yet no hypotheses of paleopolyploidy have been proposed. By determining which molluscan clades are polyploid we can help to guide future research into mollusc genomics. Here we identify several potential candidates.

3.2 Methods

3.2.1 Phylogeny

No single phylogenetic study included all the taxa in our analysis, so we used phylogenies from several different papers to piece together a tree for the Mollusca, and pruned taxa for which we did not have chromosome data. For the relationship among molluscan classes we used Haszprunar (2000). The internal structure of the Polyplacophora (Okusu et al. 2003), the Bivalvia (Giribet and Wheeler 2002), and the Cephalopoda (Lindgren et al. 2004) were each derived from separate papers. We used the phylogeny found in Ponder and Lindberg (1997) for the relationships among the major clades of gastropods. Barker (2001) provided the back bone for the phylogeny within the Heterobranchia, while Wade et al. (2006) filled in the more detailed relationships among the pulmonate families. We took Wägele et al. (2008) to be the basis for our branching patterns within the Opisthobranchia, and used Colgan et al. (2007) to add families that were not found in Wägele et al. (2008) to our tree.

For any likelihood model used to analyze phylogenetic data the probability of a transition along a branch depends on the length of that branch, and any rates used in such models are proportional to the units of branch length. We assumed that under our null model rates of chromosome duplication and loss are constant with respect to time and therefore the branch lengths for our tree should be in years between speciation events. We deduced the timing of each node in our tree by identifying the first appearance for each of the terminal taxa in our study as well as for larger clades containing one or more of these taxa. For most taxa we used the earliest fossil found in the Paleobiology Database, however we used dates found in Nishiguchi and Mapes (2008) for the cephalopods, and in Solem and Yochelson (1979) and Zilch (1959-1960) for the pulmonates. Every node was fixed at the oldest date for the first appearance of any of its descendants.

This method resulted in several branches of length zero. We could not assign a length to a terminal branch if it lead to one of two sister taxa with no fossil record, nor to an internal branch if one of the two clades immediately descended from it had a longer fossil history than the internal branch's sister clade. For example imagine a tree with two sister taxa and their closest relative. If one of the sister taxa has an older fossil record than either of the other two taxa, then both nodes will be forced down to the same time, at the first appearance of the oldest taxon. This is likely to be a consequence of the incompleteness of the fossil record, and not to represent the actual time between lineage splitting events. Therefore we assumed each node had to precede both its daughter nodes by at least some minimum number of years. Three trees were constructed with zero length branches expanded to 10^4, 10^5 and 10^6 years; we will refer to these as Tree-10^4, Tree-10^5 and Tree-10^6 respectively.

3.2.2 Chromosome Number

Chromosome counts were derived from the literature. The majority of data for Bivalves and Gastropods were found in Thiriot-Quievreux (2002, 2003). Additional data was found by searching Biosis for keywords "Karyo* or Chromosome* or Cytolog*" and taxonomic data"Mollusc". When we had chromosome number data for multiple species within a terminal taxon we used the mode for that taxon (White 1973). If there were multiple modes, we used the mode closest to the median, and if there were two modes equal distance from the median, we chose one at random.

3.2.3 Phylogenetic Signal

One should only evaluate data phylogenetically, when that data has a strong phylogenetic signal; that is to say closely related taxa are more similar to each other than would be expected at random. In order to test for signal we randomized the data among the tips 9999 times and calculated the number of steps for the randomized data sets using ordered parsimony and a parsimony model for which the cost for going from m chromosomes to n chromosomes along a branch is the absolute value of $\ln(n) - \ln(m)$. Under this second parsimony model the cost of adding or loosing chromosomes is proportional to the number of chromosomes in the genome (as it is under the birth-death process) but it still retains the quality of an ordered parsimony model that the cost of going from state x to state z equals the cost of going from state x to state y plus the cost of going from state y to state z, when y is intermediate in value between x and z. We compared the number of steps for our randomized data sets to the number of steps for our actual data. One should expect fewer steps under either parsimony model for a data set with high phylogenetic signal than for a random data set, as there should be fewer steps between closely related individuals.

3.2.4 Likelihood Model

For our null hypothesis we assumed that each chromosome has an equal probability of duplicating or splitting and an equal probability of being lost at any time. Thus there is a constant duplication rate, λ, and loss rate, μ, for each chromosome in the genome. On any branch of the taxon tree the time constant birth-death process operates and we can use (2.37) and (2.20) to calculate the probability that $\mathring{n}_i$ chromosomes at the beginning of branch i of the taxon tree will leave $\overset{\star}{N}_i$ chromosomes at the end of that branch, assuming that all $\mathring{n}_i$ chromosomes survive to the end of that branch.

$$P(N_0|\mathring{n}_i, \lambda, \mu) = \binom{\overset{\star}{N}_i - 1}{\mathring{n}_i - 1}(1-u(t_i))^{\mathring{n}_i}(u(t_i))^{\overset{\star}{N}_i - \mathring{n}_i} \tag{3.1}$$

Where t_i is the length of branch i and we can use (2.7) or (2.8) to solve for $u(t_i)$.

Using this result we can now calculate the probability that $\mathring{N}_i$ chromosomes at the beginning of the branch leave $\dot{N}_i$ chromosomes at the end of branch i, allowing for the possibility that any of the initial chromosomes could be lost. In this case $\mathring{n}_i$ is the number of chromosomes in the initial group of $\mathring{N}_i$ chromosomes that give rise to the $\dot{N}_i$ chromosomes at the end of the process. The other $\mathring{N}_i - \mathring{n}_i$ chromosomes are lost. There are $\binom{\mathring{N}_i}{\mathring{n}_i}$ different ways to arrange $\mathring{N}_i - \mathring{n}_i$ lost chromosomes among $\mathring{N}_i$ initial chromosomes. Therefore, we can calculate this probability by summing over all the possible numbers of surviving chromosomes.

$$\begin{aligned} P(\dot{N}_i|\mathring{N}_i,\lambda,\mu) &= \sum_{\mathring{n}_i=1}^{\min(\mathring{N}_i,\dot{N}_i)} \binom{\mathring{N}_i}{\mathring{n}_i} (E_0(t_i))^{\mathring{N}_i-\mathring{n}_i} P(\dot{N}_i|\mathring{n}_i,\lambda,\mu) \\ &= \sum_{\mathring{n}_i=1}^{\min(\mathring{N}_i,\dot{N}_i)} \binom{\mathring{N}_i}{\mathring{n}_i}\binom{\dot{N}_i-1}{\mathring{n}_i-1} [(1-au(t_i))(1-u(t_i))]^{\mathring{n}_i} (au(t_i))^{\mathring{N}_i-\mathring{n}_i} u(t_i)^{\dot{N}_i-\mathring{n}_i} \end{aligned} \tag{3.2}$$

This equation is equivalent to the one given by Bailey (1964, pg. 94) and Foote et al. (1999), although it takes a different form and here we provided an alternative derivation. We can assume that all the chromosomes were not lost on any branch of the tree, so that the transition probability between any two states $\mathring{N}_i$ and $\dot{N}_i$ is the probability of going from $\mathring{N}_i$ to $\dot{N}_i$ chromosomes along branch i of length t_i, given that $\dot{N}_i$ is greater than one.

$$P(\dot{N}_i|\mathring{N}_i,\dot{N}_i>0,\lambda,\mu) = \frac{P(\dot{N}_i|\mathring{N}_i,\lambda,\mu)}{1-(E(t_i))^{\mathring{N}_i}} \tag{3.3}$$

We used this equation to calculate the probability of a set of chromosome counts at the tips of a phylogeny conditioned on a particular value for λ, μ and the number of chromosomes at the root. We calculated the likelihood for the entire tree by proceeding down from the tips of the tree to the root and marginalizing over all the possible states at the internal nodes (Felsenstein 1973, 1981), such that we calculated the probability of the data above branch i given $\mathring{N}_i$ by summing over all the possible values of $\dot{N}_i$.

$$P(C_i|\mathring{N}_i) = \sum_{\dot{N}_i=1}^{\infty} p_i(\dot{N}_i|\mathring{N}_i) P(C_i^+|\dot{N}_i) P(C_i^-|\dot{N}_i) \tag{3.4}$$

where $p_i(\dot{N}_i|\mathring{N}_i) = P(\dot{N}_i|\mathring{N}_i,\dot{N}_i>0,\lambda,\mu)$, C_i are the chromosome counts on the tips above branch i and C_i^+ and C_i^- are the chromosome counts on the tips above each of the branches descended from branch i. Each of those descendant branches will obviously start with $\dot{N}_i$ chromosomes and so the probability of the data above them can in turn be calculated with (3.4). As there are an infinite number of possible states at each internal node we

excluded from our calculation those internal states that were far outside the range of observed chromosome counts.

We added whole genome duplications to this model by assuming that they occur at some constant rate, δ. We will call the number of genome duplications that happened on branch i, D_i. The transition probability of going from $\mathring{N}_i$ to $\dot{N}_i$ chromosomes along a branch with no full genome duplications is simply the transition probability from the birth-death process times the probability that no genome duplication occurred.

$$P(\dot{N}_i, D_i = 0 | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu, \delta) = \exp(-\delta t_i) P(\dot{N}_i | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu) \tag{3.5}$$

In order to calculate the transition probability of going from $\mathring{N}_i$ to $\dot{N}_i$ chromosomes along a branch with one full genome duplication, we divide the branch into two discrete time periods at the time of the duplication, t_δ. In that case the number of chromosomes before a duplication is N_δ and the time between the start of the branch and the duplication is $t_i - t_\delta$; after the duplication, there will be $2N_\delta$ chromosomes. We can calculate the transition probability by summing over all the possible values of N_δ and integrating over t_δ from zero to t_i.

$$\begin{aligned} &P(\dot{N}_i, D = 1 | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu, \delta) \\ &\quad = \delta \exp(-\delta t_i) \int_0^{t_i} \sum_{N_\delta = 1}^{\infty} P(N_\delta | \mathring{N}_i, N_\delta > 0, \lambda, \mu) P(\dot{N}_i | 2N_\delta, \dot{N}_i > 0, \lambda, \mu) dt_\delta \end{aligned} \tag{3.6}$$

We calculated this integral using numerical integration, breaking each branch into sections and assuming that the duplication happened in the middle of that section. We used three sections, as we found that three sections yielded the same results as ten sections. We calculated the sum by treating the duplication event as an internal node and limiting the number of possible states as above.

Thus we can calculate the likelihood of specific values of λ, μ and δ conditioned on the numbers of chromosomes at the root, except we now calculate the transition probability as $p_i(\dot{N}_i | \mathring{N}_i) = P(\dot{N}_i | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu, \gamma)$, where:

$$P(\dot{N}_i | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu, \delta) \approx P(\dot{N}_i, D_i = 0 | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu, \delta) + P(\dot{N}_i, D_i = 1 | \mathring{N}_i, \dot{N}_i > 0, \lambda, \mu, \delta) \tag{3.7}$$

This excludes the possibility of multiple duplications on any branch, which is in general very small. These calculations were performed using the program GDCN 1.0.

3.2.5 Model Comparison

We compared three different models by fitting the parameters of those models to our data set by maximum likelihood. The "equilibrium" model is the simplest model, it assumes that $\lambda = \mu$ and that $\delta = 0$. In order to show that our data is better explained by including

whole genome duplications we compared the equilibrium model to a "duplication" model in which δ is allowed to take values greater than 0. It is possible that the superior fit of the duplication model is a consequence of a strong tendency for the number of chromosomes to increase rather than to paleopolyploidy events. To account for this possibility we also compared these two models to the "trend" model which allows λ and μ to take different values, but assumes that $\delta = 0$. We examined all three models on all three trees in order to assure that no set of branch lengths was unduly affecting our conclusions.

Our calculation of the likelihood is conditioned on the number of chromosomes at the root. We do not know the actual number of chromosomes in the common ancestor of all molluscs, but there are several ways to estimate our likelihood without knowing this value. One way is to assume that the prior probability of any chromosome count is the same, and to calculate the likelihood as the sum of the likelihoods for every possible chromosome count at the root (Pagel 1999). This method would place undo weight on chromosome counts that are in fact highly unlikely to be the true chromosome count, and would favor hypotheses that are robust to assumptions about the ancestral chromosome count, even though there is no a priori reason to do so. For other types of discrete characters the equilibrium frequencies of the character states under the model are commonly used as priors for the root of the tree (Maddison and Maddison 2007), but the birth death process has an infinite number of possible states and thus the equilibrium frequencies for each state are either 0 or 1 depending on the values of λ and μ and the starting state. We could also assign a complex prior to the states at the root, but we have no basis for doing so without a phylogenetic analysis of chromosome number in non-molluscan Lophotrochozoa. A fourth possibility is to treat the ancestral chromosome count as another parameter to be set by our maximum likelihood search. For most of our analyses we settled on this approach. In order to justify it, we initially calculated maximum likelihoods conditioned on a range of ancestral chromosome values from 1 to 80. By comparing a large range of ancestral mollusc chromosome counts we could see how strongly our model supported the maximum likelihood reconstruction of the root, and be certain that our model choice was not overly biased by our reconstruction of the ancestral chromosome count.

The maximum likelihoods for all models under all sets of root values were used to calculate the Akaike Information Criterion (AIC) for those models. The AIC is defined as twice the number of parameters in the model minus twice the maximum likelihood. The AICs of the different models were compared in order to select the best fit model for the data. Models that fit the data better should have lower AICs (Akaike 1974).

3.2.6 Identifying Branches with Duplications

The ultimate purpose of this exercise was to identify branches on which it is likely that a full genome duplication occurred. In order to accomplish this we calculated the posterior probability that there was a duplication on each branch of the tree. For each branch, i, we used the maximum likelihood values of λ, μ, δ and the ancestral chromosome count, to

calculated two likelihoods: L_i^δ, in which $p_i(\overset{\star}{N}_i|\overset{\circ}{N}_i) = P(\overset{\star}{N}_i, D_i = 1|\overset{\circ}{N}_i, \overset{\star}{N}_i > 0, \lambda, \mu, \delta)$ on that branch; and $p_j(\overset{\star}{N}_j|\overset{\circ}{N}_j) = P(\overset{\star}{N}_j|\overset{\circ}{N}_j, \overset{\star}{N}_j > 0, \lambda, \mu, \delta)$ on every other branch, j; and L_i^0, in which $p_i(\overset{\star}{N}_i|\overset{\circ}{N}_i) = P(\overset{\star}{N}_i, D_i = 0|\overset{\circ}{N}_i, \overset{\star}{N}_i > 0, \lambda, \mu, \delta)$ on that branch; and $p_j(\overset{\star}{N}_j|\overset{\circ}{N}_j) = P(\overset{\star}{N}_j|\overset{\circ}{N}_j, \overset{\star}{N}_j > 0, \lambda, \mu, \delta)$ on every other branch, j. In that case the maximum likelihood should equal $L_i^\delta + L_i^0$ and the posterior probability of a duplication on that branch will be $P(D_i = 1|C, \lambda, \mu, \gamma) = L_i^\delta/(L_i^\delta + L_i^0)$. These values were calculated for all three trees, using the maximum likelihood values for λ, μ, δ and the root.

3.2.7 Simulations and Model Fit

It is important to show not just that one of these models fit the data better than another, but that these models provide a reasonable description of the data themselves. In order to do make this comparison, maximum likelihood values of λ, μ and δ for our data set were derived for the duplication and the equilibrium models on Tree-10^6 conditioned on the maximum likelihood count of ancestral chromosomes. Approximately 1000 data sets were then simulated using each set of maximum likelihood parameters and starting with the maximum likelihood number of chromosomes at the root. Maximum likelihoods were then calculated for each of these data sets under the models used to generate the data. These maximum likelihoods were used as statistics to reject the hypothesis that the actual data was generated by this model. Failure to reject this hypothesis was taken to indicate that the model was a reasonable approximation of the actual process by which the real data were generated.

The duplication model has one more free parameter than the equilibrium model, as the equilibrium model can be considered a special case of the duplication model in which $\delta = 0$. Thus twice the difference between the logs of the maximum likelihoods for these models should correspond to a chi-squared distribution with one degree of freedom. We tested the fit of this statistic to the chi-squared distribution by calculating maximum likelihoods using both the equilibrium model and the duplication model for each of the data sets simulated under the equilibrium model. Twice the difference between the logs of these likelihoods was calculated and compared to a one tailed chi-squared distribution with one degree of freedom by a quantile-quantile plot.

We especially wanted to show that the birth-death process was an appropriate model for the background rate of chromosome number evolution independently of any whole genome duplications. First we simulated 1000 data sets using maximum likelihood values for the root count, λ, μ and δ but we also placed one duplication at random on a minimum length branch. Then we simulated another 1000 data sets on Tree-10^6 starting with 15 chromosomes and using the maximum likelihood values of λ and μ calculated under the duplication model. However, under these simulations, whole genome duplications did not occur at random, but instead always occurred on the branches for which our actual data showed a high posterior probability of a whole genome duplication. We calculated maximum likelihoods using the

duplication model and compared those likelihoods to the one calculated for our actual data in order to reject the use of the birth-death process for our background rate of chromosome evolution.

3.3 Results

3.3.1 Phylogeny, Chromosome Counts and Signal

The topology for the phylogeny we used is shown in Figure 3.2a. Most traditional Molluscan clades are monophyletic in this tree, but both Caenogastropoda and Sigmurethra are paraphyletic. Figure 3.2b shows the same tree with the branch lengths scaled to match the branch lengths we derived from paleontological data. 69 of 123 internal branches had length zero, as a consequence of basal branching taxa in a clade having a more recent fossil record than a taxon nested well within that clade, and 6 of 125 terminal branches had length zero, because two sister taxa had no fossil record. Zero length branches are shown as polytomies.

We obtained chromosome counts for 997 species of Mollusca from 125 terminal taxa including members of all five major extant classes (Polyplacophora, Bivalvia, Cephalopoda, Scaphopoda and Gastropoda). We used the mode of the chromosome number for each terminal clade in our phylogenetic analysis (Table A.1). In most cases all the chromosome counts in each terminal clustered around the modes, but several clades have members with highly divergent counts. Within the Anomoidea, Loliginidae, Viviparoidea, Thiaridae, Anclidae, and Planorbidae there are species with chromosome counts two times the mode for the entire clade implying recent polyploidies. In the Planorbidae genus *Bulinus* there are species with three and even four times the mode of the clade. Furthermore within the Unionoida, Cardioidea, Turbinoidea, Cerithiidae, Pleuroceridae, Littorinidae, Muricidae, Conoidea, and Succineidae there are species with chromosome counts approximately half of the mode. Most of these clades are nested in larger clades with modes similar to their own.

Chromosome number has very high phylogenetic signal among the molluscan taxa studied, as demonstrated by two different parsimony statistics we derived from our data set and compared to the same statistics calculated for data sets generated by randomizing our data over the tips of the trees. Our data set had 250 steps for ordered parsimony and 13.36148 steps for our weighted step matrix. Both these values were highly significant ($p<<0.0001$), as they were less than the equivalent value for any of our 9999 randomized data sets. Furthermore the parsimony scores for our randomized data sets fell between 558 and 749 steps for the ordered parsimony and between 27.47879 and 37.12357 steps for the weighted parsimony, implying that the actual p-value is much lower. Thus there is a very clear phylogenetic pattern to the distribution of our data on the tree.

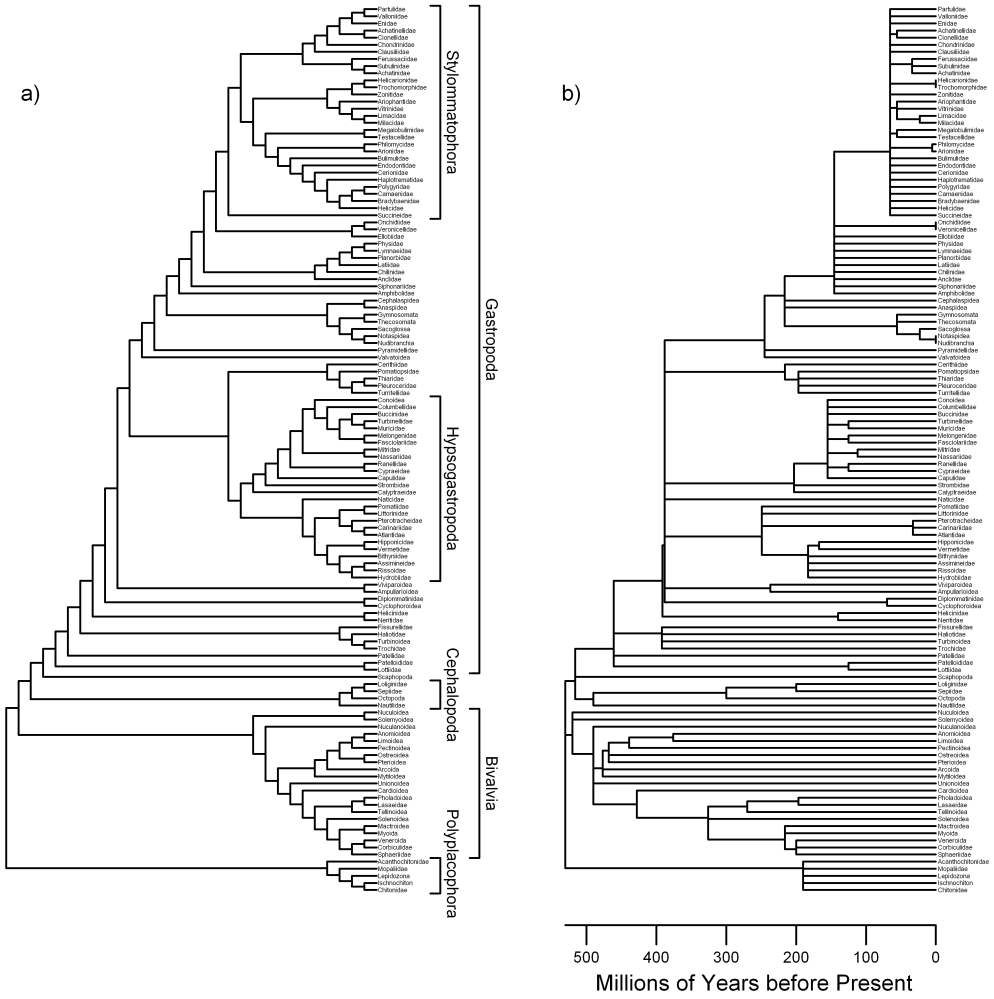

Figure 3.2: a) The cladogram of the Mollusca used in this study. Several clades important to this study are identified with brackets. b) The phylogeny of Molluscan taxa used in this study. The topology of this tree is the same as the tree in Figure a), but here branch lengths are shown in millions of years as derived from the fossil record using the Paleobiology Database. Zero length branches are shown as polytomies.

Table 3.1: Maximum likelihood parameter values for all three models and sets of branch lengths. All rates are in units of Events/Chromosome/Billion Years.

Branch Lengths	Model	AIC	Root Count	$\lambda+\mu$	$\lambda-\mu$	δ
Tree-10^6	Equilibrium	364.540	17	5.0359	0*	0*
	Trend	366.504	16	4.9772	0.06851	0*
	Duplication	357.508	15	3.0555	0*	0.12304
Tree-10^5	Equilibrium	366.792	17	5.2310	0*	0*
	Trend	368.719	20	5.4529	-0.34818	0*
	Duplication	361.744	15	3.4038	0*	0.11661
Tree-10^4	Equilibrium	369.134	17	5.2577	0*	0*
	Trend	371.043	20	5.4787	-0.34931	0*
	Duplication	364.529	15	3.6157	0*	0.10860

* Rate is assumption of model, not set by Maximum Likelihood.

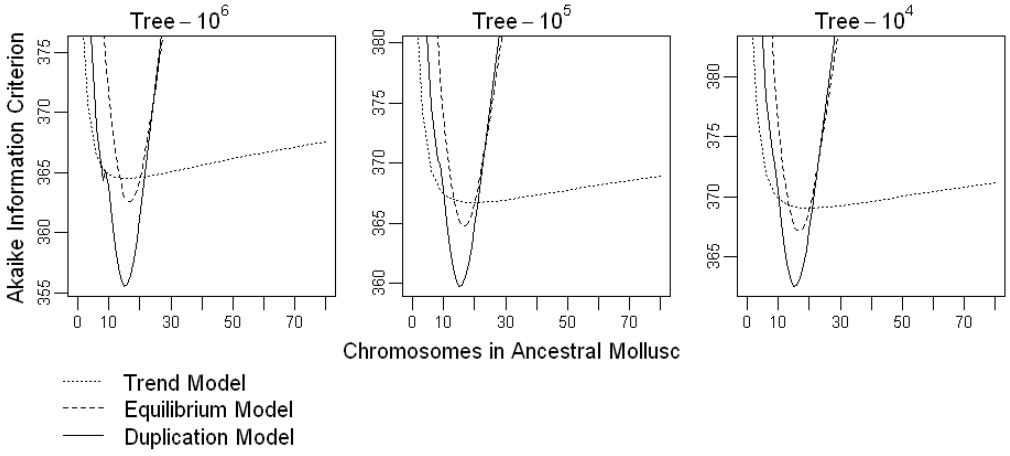

Figure 3.3: The Akaike Information Criterion (AIC) for each of three models of chromosome number evolution to chromosome counts found in extant Mollusc families. Each figure shows the AICs for all three models conditioned on the number of chromosomes found in the last common ancestor of all Mollusca for a given set of branch lengths.

3.3.2 Model Choice

The duplication model achieved its maximum likelihood at 15 ancestral chromosomes for all three trees while the equilibrium model achieved its maximum likelihood at 17 (Table 3.1). For both these models likelihoods fell off precipitously for all trees when conditioned on ancestral chromosome counts that differed from their maximum likelihood chromosome count by more than 2 (Figure 3.3). The trend model reached its maximum likelihood

at chromosome counts between 16 and 20 for the three different trees, and while the likelihoods decreased precipitously for lower ancestral chromosome counts, they decreased more gradually for higher counts.

The AIC was much lower for the duplication model than it was for the equilibrium model when conditioned on ancestral chromosome counts near the maximum for all four trees (Figure 3.3). Although the equilibrium model did have a lower AIC than the duplication model when conditioned on ancestral chromosome counts that are far from their maxima, these ancestral chromosome counts are themselves poorly supported by either model. When the ancestral chromosome count was treated as an additional maximum likelihood parameter the AIC was lower for the duplication model than it was for the equilibrium model for all sets of branch lengths by at least 4.6 and as much as 7.0 (Table 3.1). It is reasonable to conclude that the duplication model is a much better fit for the data than the equilibrium model.

The duplication model had much lower AICs than the trend model when the ancestral chromosome count was fit by maximum likelihood for all sets of branch lengths (Table 3.1); the difference between AICs ranged from 6.51 for Tree-10^4 to 9.00 for Tree-10^6. The equilibrium model also had lower AICs than the trend model, when the ancestral chromosome count was fit by maximum likelihood, although the differences were all less than 2 (Table 3.1). However, the trend model achieved relatively high maximum likelihoods over a much larger range of ancestral chromosome counts, and when conditioned on more extreme chromosome counts it had lower AICs than either of the other two models (Figure 3.3). Nevertheless, the duplication model is clearly a better fit for the data as its AICs were so much lower for all four trees when the ancestral chromosome counts were fit by maximum likelihood.

There was little variation in the maximum likelihood estimates of parameter values between the three sets of branch lengths (Table 3.1). Maximum likelihood estimates for the total rate of change for the birth-death process were much smaller under the duplication model than under the other two models, because under the duplication model a great deal of the change in chromosome numbers can be accounted for by whole genome duplications. The estimate of net change in chromosome number under the trend model was slightly positive for tree-10^6; the estimate of net change was negative for the other two trees, but still only represented 6.4% of the total rate. Our estimates of the whole genome duplication rate under the duplication model ranged from 0.109 duplications/billion years for Tree-10^4 to 0.123 duplications/billion years for Tree-10^6. It should be noted that this is the whole genome duplication rate we would expect per lineage, not over the whole tree, thus we observed several whole genome duplications in a clade with a 530 million year history.

3.3.3 Branches with Duplications

Three branches had posterior probabilities of whole genome duplications greater than 0.67 for at least one set of branch lengths when conditioned on their maximum likelihood parameter values. These included the branch at the base of the Coleoidea, a branch within the

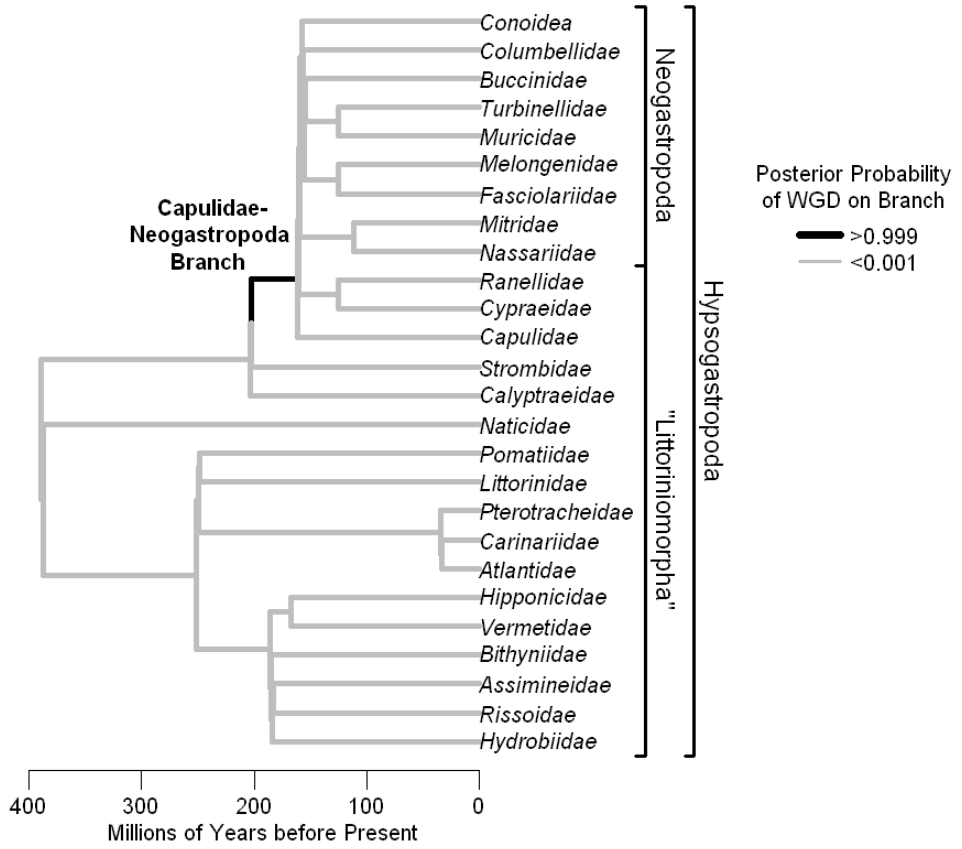

Figure 3.4: The phylogeny of all the terminal taxa in the Hypsogastropoda using the branch lengths from Tree-10^6 showing the posterior probability of a whole Genome Duplication (WGD) on each branch. The Capulidae-Neogastropoda branch is marked, as it has a very high posterior probability of a WGD, while all the other branches have essentially none. Support for a WGD on the Capulidae-Neogastropoda branch is greater than 0.999 for all sets of branch lengths.

Stylommatophora at the base of a clade containing the Sigmurethra and the Orthurethra, and a branch within the Hypsogastropoda at the base of an unnamed clade. Several branches phylogenetically close to these well supported branches also showed some support for a whole genome duplication (Figures 3.4, 3.5 and 3.6). No other branch had a posterior probability greater than 0.03 for any of the trees.

The branch within the Hypsogastropoda at the base of a clade sister to the Strombidae and containing the Neogastropoda and several families of Littorinimorpha - hereafter to be referred to as the Capulidae-Neogastropoda branch - had posterior probabilities of a whole genome duplication greater than 0.999 for all three trees (Figure 3.4). This is extremely strong support for an evolutionary scenario in which the number of chromosomes doubled

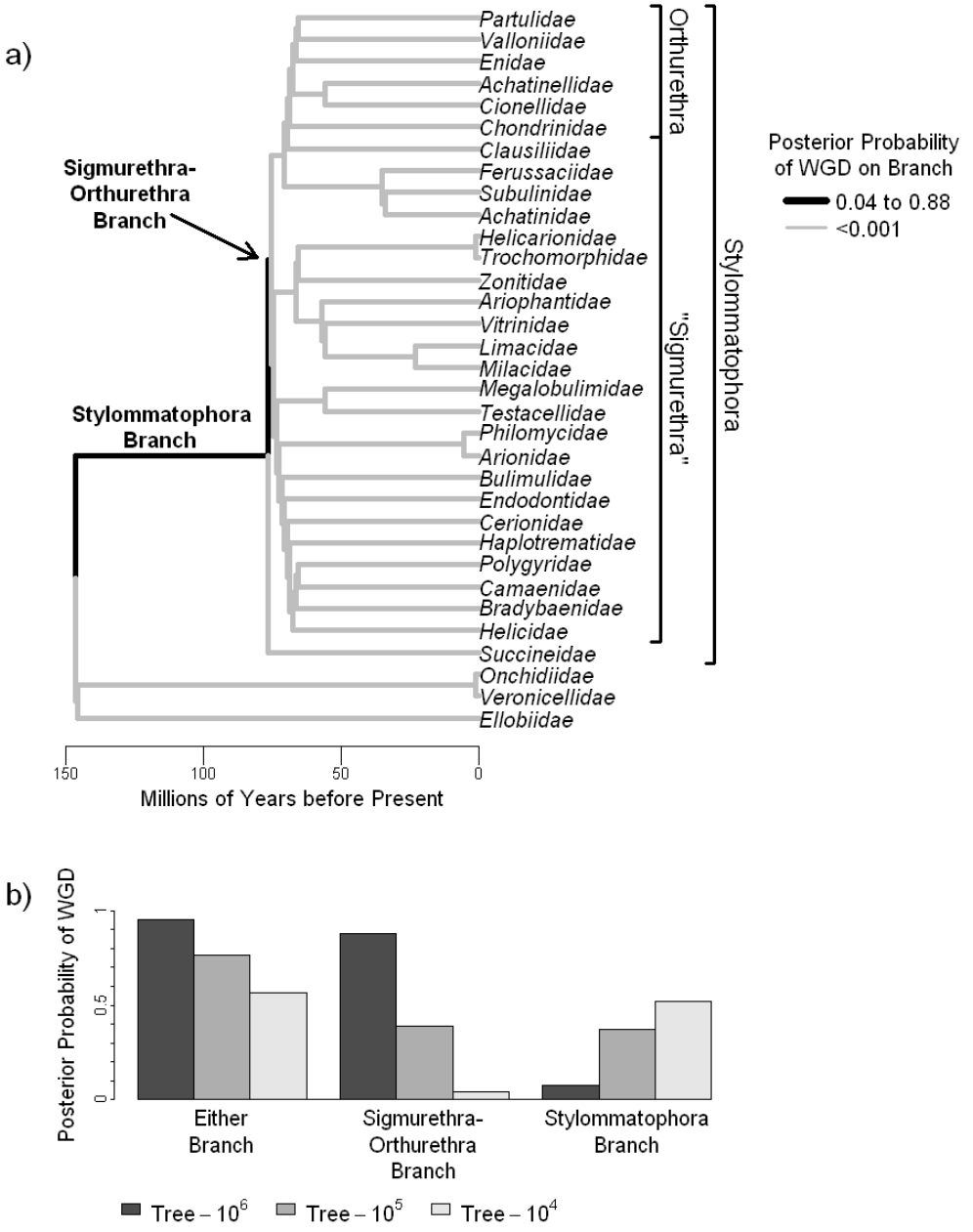

Figure 3.5: Posterior probabilities for a whole genome duplication on two branches near the base of the Stylommatophora. a) The phylogeny of all the terminal taxa in a clade containing the Stylommatophora, the Systellommatophora and the Ellobiidae using the branch lengths from Tree-10^6 showing the posterior probability of a whole Genome Duplication (WGD) on each branch. The Stylommatophora branch

and the Sigmurethra-Orthurethra branch are marked, as they have a relatively high posterior probability of a WGD, while all the other branches have essentially none. b) A bar plot showing the posterior probability of a whole genome duplication on either of these branches as well as on each of these branches individually under each of the three different sets of branch lengths. Support for a WGD on the Sigmurethra-Orthurethra branch as well as on either of the two branches decreases as the minimum branch length decreases, while support for a WGD on the Stylommatophora branch increases.

on this branch.

We calculated the posterior probability for a whole genome duplication on a branch within the Stylommatophora at the base of a clade containing the Sigmurethra and the Orthurethra of 0.881 for Tree-10^6 (Figure 3.5). On the other hand Tree-10^5 had much less support for a duplication on this branch (posterior probability=0.389), and Tree-10^4 had even less (posterior probability=0.043). Tree-10^4 and Tree-10^5 alternatively showed some support (posterior probabilities 0.520 and 0.374 respectively) for a whole genome duplication on the branch at the base of the Stylommatophora, while Tree-10^6 supported a whole genome duplication on this branch more weekly (posterior probability=0.074). There was no support for whole genome duplications occurring on both these branches; the posterior probability was less than $1x10^{-9}$ for all sets of branch lengths. The Sigmurethra-Orthurethra branch has a minimum branch length; as a consequence the likelihood of a duplication on this branch was less on trees with smaller minimum branch lengths. This greatly decreased the support for a doubling on this branch and instead compensated in part by increasing the support for a doubling on the ancestral stylommatophoran branch, which is 80 million years long and thus much more likely to have a whole genome duplication on it. This appears to be an effect of our branch lengths, thus the analysis on Tree-10^6, which is less effected by branch lengths, is the most reasonable one and it is likely that a paleopolyploidy event occurred on the Sigmurethra-Orthurethra branch.

The data clearly supported at least one paleopolyploidy event in the Cephalopods, but it was difficult to distinguish whether there were one or two whole genome duplications and the exact branch on which they occurred (Figure 3.6a). All three sets of branch lengths had a posterior probability of a whole genome duplication on the Coleoidea branch between 0.675 and 0.687 and on the Decapodiformes branch between 0.288 and 0.303 (Figure 3.6b), with weak support for duplications occurring on both of the branches (posterior probability between 0.030 and 0.046), meaning that the posterior probability that a whole genome duplication occurred on one of these branches was greater than 0.93 for all three sets of branch lengths. These trees also had posterior probabilities less than 0.09 that a whole genome duplication occurred either at the base of the Cephalopods or on the Nautilidae branch with an overall posterior probability between 0.078 and 0.091 that a whole genome duplication occurred on more than one branch in this clade.

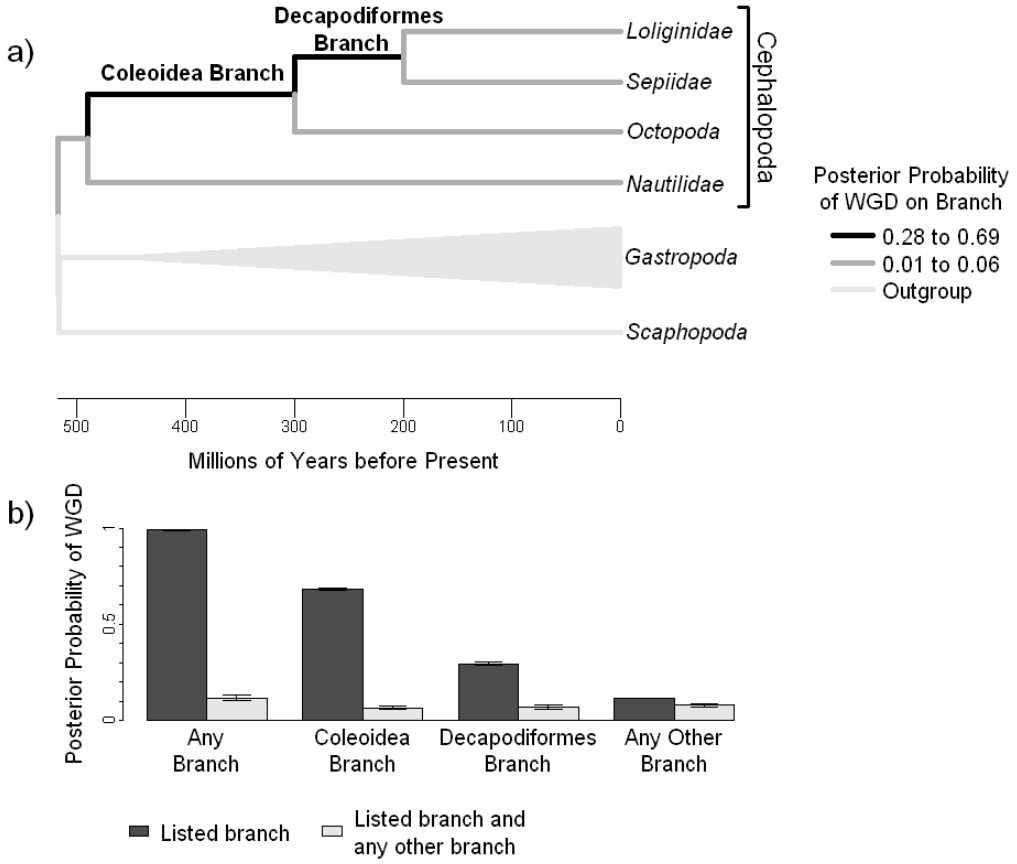

Figure 3.6: Posterior probabilities for a whole genome duplication on branches within the Cephalopoda. a) The phylogeny of all the terminal taxa in the Cephalopoda with Gastropoda and Scaphopoda included as an outgroup using the branch lengths from Tree-10^6 showing the posterior probability of a whole Genome Duplication (WGD) on each branch. Two branches with relatively high posterior probability of a WGD are labeled above the branch. b) A bar plot showing the posterior probability of a whole genome duplication on various branches in the phylogeny. The dark bars show the posterior probability of a WGD on the specified branch or branches. The light bars show the posterior probability that there were WGDs on any pair of branches in the Cephalopoda including the specified branch or branches. The posterior probability is shown by the large bar for Tree-10^5, by the lower error bar for Tree-10^4, and by the upper error bar for Tree-10^6. Support for a WGD somewhere in this clade is strong, but the exact location is not clear. There is little variation between the different sets of branch lengths.

3.3.4 Simulations to Evaluate Model Fit

We ran 1000 simulations on Tree-10^6 under the equilibrium model starting with 17 chromosomes using the maximum likelihood value of λ calculated for our actual data set. When evaluated under the equilibrium model, 160 of these simulations had maximum likelihoods lower than our data set, implying that we can not reject this model as an explanation for our data (p=0.160). We also calculated the maximum likelihood for all simulated data sets under the duplication model, and used that value to calculate a log likelihood ratio. The largest log likelihood ratio for our simulated data sets was 5.506 and the likelihood ratio for over 96% of our simulations was less than 0.001, while the likelihood ratio for our actual data was 18.064. Thus we can strongly reject our null hypothesis in favor of our alternative hypothesis of genome duplications ($p<<0.001$). Comparing the log likelihood ratios of our simulated data to a one tailed chi-squared distribution with one degree of freedom indicates that the chi-squared distribution is an extremely conservative basis upon which to estimate the level of significance for rejecting our null hypothesis.

We also ran 1000 simulations on Tree-10^6 under the duplication model using the maximum likelihood parameter values for a process conditioned on 15 ancestral mollusc chromosomes. When evaluated under the duplication model 40 of these simulations had maximum likelihoods less than our actual data, indicating that we can also tentatively reject our alternative hypothesis as a fit for our model (p=0.040).

However, the major reason that the simulations had higher maximum likelihoods than the actual data is that one of the reconstructed genome duplications from our actual data appears to have occurred on a minimum length branched. Given the estimated value for δ of 0.123 whole genome duplications/ billion years, the probability of a whole genome duplication on so short a branch is approximately 1.23 x 10^{-4}. Thus any data set with a duplication on a minimum length branch will obviously have a lower maximum likelihood than one without. Given this estimated value of gamma we would expect whole genome duplications to occur on a minimum length branch in 0.943% of our simulations, and indeed they actually did occur in only 9 of our simulations. Since the vast majority of our simulations did not have a duplication on a minimum length branch we would expect them to have higher likelihoods than our actual data. To confirm this we did 1000 simulations in which at least one duplication occurred on a minimum length branch using the maximum likelihood parameters from the duplication model, 166 had a maximum likelihood lower than our data set, indicating that we can not reject the birth-death process as a model for the background rate of chromosome change.

In order to get a better idea of the fit of our data to the birth death process as the background rate of chromosome change, we ran 1000 simulations starting with 15 chromosomes on Tree-10^6 with maximum likelihood values for λ, as found under the duplication model, in which duplications always occurred on the Capulidae-Neogastropoda branch, the Coleoidea branch, and the Sigmurethra-Orthurethra branch, and nowhere else. When these simulations were evaluated under the duplication model, 25.1% of the data sets

simulated with the deduced set of whole genome duplications had maximum likelihoods less than our actual data set. Thus we could not reject the birth death process as a model for our background rate of chromosome evolution.

3.4 Discussion

We developed a likelihood model based on the birth-death process that allows us to predict the phylogenetic position of paleopolyploidy events through comparative analysis of chromosome counts in extant species. We applied this model to a data set of chromosome counts in molluscan taxa and found that a model in which the total number of chromosomes occasionally doubled explained our data much better than a model in which chromosome counts only evolved via the birth-death process (Figure 3.3). We identified three potential instances of paleopolyploidy within the Mollusca (Figures 3.4, 3.5 and 3.6). In one case we could clearly identify the branch on which the whole genome duplication occurred; in the other two cases we could narrow down the position of the whole genome duplication to one of two branches. Based on the assumptions inherent to our model, support for whole genome duplications within the Mollusca in general and in these three clades in particular is very strong.

Comparative analysis of chromosome counts suggests that a whole genome duplication occurred in the common ancestor of a clade containing the Capulidae, the Ranellidae, the Cypraeidae and the Neogastropoda after their divergence from the Strombidae and the other hypsogastropod families included in our analysis (Figure 3.4). This paleopolyploidy event was strongly supported by all three sets of branch lengths (posterior probability>0.999) and there was no support for a whole genome duplication on any other branch in the Hypsogastropoda. Our interpretation of the fossil record indicates that this whole genome duplication occurred at some point between the beginning of the Jurassic (203 MYA) when the first Strombidae fossils appear and the lower Cretaceous (155 MYA) when the Neogastropoda initially radiated.

We identified another likely whole genome duplication early in the history of the Stylommatophora: either at the beginning of the Cenozoic (65 MYA) in the common ancestor of the Sigmurethra and the Orthurethra after they diverged from the Succineidae; or in the common ancestor of all the Stylommatophora after they diverged from the other Pulmonates in the lower Cretaceous (138 MYA) and before they radiated at the beginning of the Cenozoic (65 MYA) (Figure 3.5a). We could not establish a length for the Sigmurethra-Orthurethra branch or many of the branches immediately descended from it, as several extant Stylommatophoran families appear at the beginning of the Cenozoic and the phylogenetic position of Stylommatophora fossils appearing before then is uncertain (Solem and Yochelson 1979; Zilch 1959-1960) (Figure 3.2); we treated each of these branches as one branch of minimum lengths within a relatively fast radiation. As a consequence, support for a paleopolyploidy event on the Sigmurethra-Orthurethra branch decreased for

shorter minimum branch lengths, as the posterior probability of a whole genome duplication is proportional to the length of the branch. There was a concomitant increase in the posterior probability on the Stylommatophora branch, but it was not sufficient to compensate for the decrease on the Sigmurethra-Orthurethra branch, thus total support for a whole genome duplication on either of these branches decreased with shorter minimum branch lengths (Figure 3.5b). As the length of the Sigmurethra-Orthurethra branch affects the posterior probability of both branches and the maximum support on the Sigmurethra-Orthurethra branch is higher than the maximum support for the Stylommatophora branch, we conclude that the lack of support for whole genome duplications on the Sigmurethra-Orthurethra branch is in fact an artifact of our method of assigning branch lengths, and that it is more reasonable to conclude that a whole genome duplication occurred there than on the Stylommatophora branch. It should be noted that the number of chromosomes was highly variable among the Succineidae species that we have data for (Table A.1), thus it is possible that selecting a different chromosome number to represent the Succineidae would lead to different conclusions about the location of the paleopolyploidy event. A well resolved phylogeny of the Succineidae should be used in order to better reconstruct the number of chromosomes in their last common ancestor, but such a phylogeny is not currently available. Vinogradov (2000) detected a general tendency to large genome sizes in terrestrial pulmonates. However as his analysis was not phylogenetic, the signal may be a consequence of large chromosome counts in the Stylommatophora.

Our data suggested that a third whole genome duplication occurred within the Cephalopoda, although the exact location of this paleopolyploidy event is not clearly reconstructed (Figure 3.6a). All sets of branch lengths support a duplication in the common ancestor of the Coleoidea after they diverged from the Nautilidae in the lower Ordovician (490 MYA) but before the Decapodiformes split from the Octopoda in the Carboniferous (300 MYA). On the other hand they also show some support for a duplication occurring in the common ancestor of the Decapodiformes after they split from the Octopoda, but before the Sepiidae and the Loliginidae split in the lower Jurassic (200 MYA). There is also some support for multiple whole genome duplications within the Cephalopoda either on both these branches or on some combination of these branches and other branches within the Cephalopoda (Figure 3.6b). Inspection of chromosome counts within the Cephalopoda implies another scenario. The Octopoda have 30 chromosomes in their haploid genome on the other hand the Sepiidae and the Loliginidae both have 46. Given that we reconstructed 15 chromosomes in the ancestral Mollusc, this implies that the Octopoda are tetraploid, while the Decapodiformes are hexaploid. Our model does not account for changes in ploidy other than doublings and thus would be limited to identifying tetraploids and octoploids. The hexaploid Decapodiformes likely caused the uncertainty in reconstructing the whole genome duplication within the Cephalopoda. Although variation in chromosome counts within both the Sepiidae and the Loliginidae is overall very low, we have data for one species in each family with many more chromosomes than the mode (Table A.1), thus it is possible that the ancestral Decapodiformes had many more chromosomes than are shown by our

data, which would increase support for a whole genome duplication on the Decapodiformes branch. Increasing the taxon sampling on each of our terminal branches would allow us to better reconstruct the ancestral chromosome counts at the tips and reduce the length of the terminal branches. However, there are no extant taxa that branch off of any of the internal branches within our cephalopod phylogeny, and so we can not break up any of the long internal branches.

It has been suggested that whole genome duplications can lead to large morphological and physiological innovations, as redundancy eases the constraints on genes throughout the genome (Ohno 1967; Haldane 1932). Indeed the Stylommatophora, the Coleoidea and the Decapodiformes are all characterized by a number of important morphological synapomorphies. The Stylommatophora are the dominant group of land snails and slugs and are characterized by a long pedal gland placed beneath a membrane and retractile tentacles (Mordan and Wade 2008). The Coleoidea in particular represent a large jump in morphological complexity, as they are characterized by an internalized shell, a muscular mantle for locomotion, chromatophores, ink sacs, an eye lens and complex morphologically distinct arms (Nishiguchi and Mapes 2008). The Decapodiformes are a large group with a diversity of body forms and ecologies and are characterized by having two of there arms modified as tentacles and highly complex suckers (Nishiguchi and Mapes 2008). A paleopolyploidy event may have contributed to any or all of these innovations. On the other hand, the Capulidae-Neogastropoda clade is unnamed and has not been recognized for any major innovations. The Neogastropoda are a large and diverse clade found within this group, and according to our estimates the whole genome duplication may have preceded the radiation of the neogastropods by only a short time, and so may have played a major role in their evolution.

Although this is the first suggestion of paleopolyploidy in the Mollusca that we are aware of, it is not a surprising finding, as more recent polyploidies have long been recognized in many species of Molluscs. Several polyploid species have been identified in the wild (e.g. Patterson and Burch 1978; Barsiene et al. 1996; Park 2008) and in at least one case the origin of a polyploidy has been analyzed karyologically (Goldman et al. 1983). Polyploidy has also been artificially induced in several species of commercial molluscs (e.g. Beaumont and Fairbrother 1991; Le Pennec et al. 2007; Yang and Guo 2006). Furthermore, polyploidy is common in the somatic tissue of many Molluscs (Tokmakova et al. 2006; Tabakova et al. 2005; Anisimov et al. 1995). Within several of the terminal taxa in our analysis, we recognized members with approximately twice the number of chromosomes as the mode for that family; these are probably also a consequence of more recent whole genome duplications. We could expand our analysis to include these events by using a more refined phylogeny with genera or species at the tips, instead of the terminal taxa we used.

Overall we observed a strong phylogenetic signal in mollusc chromosome number with no tendency for the number of chromosomes to increase or decrease. Some researchers had previously noticed that taxonomic groups within the Mollusca tended to have similar numbers of chromosomes (Patterson and Burch 1978; Nakamura 1985), as our analysis

confirmed ($p<<0.001$). Many researches have observed a tendency for chromosome numbers to either increase (e.g. Patterson and Burch 1978) or decrease (e.g. Butot and Kiauta 1969; AHMED 1976). Maximum likelihood analysis of our data set suggested that the rate of chromosome addition was less than the rate of chromosome loss for two sets of branch lengths and greater for the third set. However, in no case was the trend greater than 6.4% of the total rate. Furthermore a likelihood model in which the chromosome numbers were expected to stay stable was a better fit for our data than one in which the chromosome loss and addition rates were allowed to vary independently and thus create a trend, although the whole genome duplications could be construed as representing a positive trend. We also observed that several families contained species with approximately one half the number of chromosomes as the mode. In contrast to doubling, there is no known biological mechanism to halve the number of chromosomes. It is possible that shortly after a whole genome duplication some lineages shed the redundant half of their genome, or that the half chromosome counts actually represent the primitive state for these families and the larger chromosome counts are actually derived as a consequence of a whole genome duplication. However, both these scenarios are contradicted by the fact that these terminals are nested in larger clades in which the other terminals have modes similar to their own. A more intensive phylogenetic analysis of *Nucella* also determined that the half chromosome count in *N. lapillus* is the derived state (Collins et al. 1996). This phenomenon bears further investigation.

We used the birth-death process as a null-model and as a model for the background rate of change in chromosome numbers. This seems an appropriate model as aneuploidy is usually a consequence of nondisjunction; thus both gains and losses are likely to occur at approximately equal rates for any random chromosome. Furthermore, as the birth-death process assumes equal rates of change, it is an appropriate null model for detecting variation in such rates. From a practical stand point the birth-death model allows us to calculate transition probabilities for meristic data using relatively few parameters. However, there are several ways in which the evolution of chromosomes does not reflect the birth-death process. Selective and mechanical constraints will place both upper and lower bounds on the process of chromosome number evolution (Roth et al. 1994), while the birth-death process will allow the chromosome numbers to vary from zero to infinity. Also it is highly unlikely that rates of chromosome duplication and loss would remain constant through 500 million years of evolution; especially after whole genome duplications, when duplicate chromosomes would likely be lost, as they would be redundant and thus not protected by natural selection (Lynch and Conery 2000, 2003). It would be nice to use a more complex model that accounts for some of these possibilities, but transition probabilities are very difficult to calculate under such models. Nevertheless, we were unable to reject the birth-death process as a model for the back-ground rates of chromosome evolution in our data set by comparing its maximum likelihood to the maximum likelihoods of 1000 simulations with whole genome duplications on the same branches on which we deduced duplications occurred from our actual data and 1000 simulations with random duplications, in which at least one occurred on a minimum length branch. Thus, the birth death process is a sufficient approximation of the background

rate of chromosome number evolution for our data set.

Mayrose et al. (2010) described a similar method to deduce the phylogenetic location of paleopolyploidies, and used a maximum likelihood implementation of it to analyze chromosome counts from several plant taxa. Rather than analytically calculating the probability of a transition under a birth-death process, they established a rate matrix and approximated the exponent of that rate matrix to calculate the transition probabilities. Thus they had to limit the number of possible states, but as a consequence were able to easily use a number of models. They used models in which the rate of increase and decrease were independent of the number of chromosomes and models in which those rates were linearly related to the number of chromosomes; the birth-death process is a specific case of the latter model. They found that the models with constant rate were a slightly better fit for the data. They were also able to include rates for genome doubling, genome halving and genome increasing by 50%. As they used a rate matrix, they were able to integrate over al possible timings of a genome duplication and include the possibility of multiple duplications on a single branch, while we used numerical integration to make our calculations. Finally they were able to exclude the possibility of reaching zero chromosomes from their models, while we merely assumed that the process did not reach zero chromosomes on any branch in the phylogeny. They compared models using the AIC and calculated the posterior probability of a duplication on any branch of the tree in the same manor that we did.

We used a mix of maximum likelihood and Bayesian methods in our analysis. Parameter values were fit by maximum likelihood, and the maximum likelihood values were used to compare models. However, we calculated the posterior probability of a duplication on each branch, as in a Bayesian analysis, but when conditioned on the maximum likelihood parameter values. This is similar to the method used by Pagel (1999) to reconstruct the state of discrete characters at the nodes of a phylogeny and by Yang (1994) and Nielsen and Yang (1998) to determine in which evolutionary rate category different nucleotides belonged. In this analysis the posterior probability for a whole genome duplication is a measure of our confidence that the number of chromosomes doubled at that location in the phylogeny within the context of our model given our observed chromosome counts (see Ellison 2004).

Branch lengths are critical to any phylogenetic method that relies on probabilistic models of character change, as changes on short branches will be considered less probable than changes on long branches (e.g. Diaz-Uriarte and Garland 1998; Mayrose et al. 2004). In this study we used a set of branch lengths derived exclusively from the fossil record. This provided us with a set of branch lengths based entirely on time. We were lucky to be studying a group with a relatively strong fossil history. Nevertheless, the fossil record is incomplete and can only supply minimum ages for nodes; it may be more accurate to include data from molecular branch lengths as well (see Rambaut and Bromham 1998; Drummond et al. 2006; Sanderson 2003), but this data was not available to us. Although these branch lengths are imperfect they provide a strong general outline for the major features of the tree, such as the relatively long terminal branches. We ended up with several zero length branches as a consequence of the incomplete fossil record and created three different sets of branch lengths

with those zero length branches expanded by different amounts. All three sets of branch lengths had essentially the same results, only disagreeing on the exact location of the whole genome duplication in the Stylommatophora.

It is possible that the events we identified within our phylogeny are not in fact whole genome duplications. They may represent large increases in chromosome number that were not actual doublings. However, if this were the case, then the posterior probabilities would likely be diffusely spread over several branches, whereas in our analysis the posterior probabilities are quite strong on individual branches. In fact the one situation in which the distribution of the posterior probabilities is more diffuse, in the Cephalopods, is likely a consequence of a different process operating on the genome. Even if these events do actually involve an exact doubling in the number of chromosomes, they are not necessarily whole genome duplications. To confirm that these instances are in fact paleopolyploidy events will require intensive analysis of whole genome sequences (e.g. Wong et al. 2002; Woods et al. 2005).

This method could also be applied to infer the phylogenetic position of paleopolyploidies in other taxa. As long as the background rate of chromosome number evolution is low enough relative to the time spans involved, it should be possible to detect a whole genome duplication from comparative analysis of chromosome counts. However, background rates of change in chromosome number may be particularly low in gastropods (Chambers 1987); in other taxa the background rate may drown out the signal of whole genome duplications. It would also be possible to apply this method to counts of gene family members or large syntenic regions, which we would also expect to double after a whole genome duplication. Much support for paleopolyploidy has come from the identification of repeated syntenic regions (e.g. Holland et al. 1994; Abi-Rached et al. 2002; Wong et al. 2002). To better detect the signal of whole genome duplication, one should simultaneously analyze counts for some combination of multiple regions, multiple gene families and chromosomes.

This study only represents the first step in the study of molluscan paleopolyploidies. This analysis can help to guide future researchers in the selection of molluscan taxa for whole genome sequencing (Figure 3.1). The study of whole genome duplications requires multiple genome sequences (Wong et al. 2002; Woods et al. 2005). Currently *Lottia gigantea* is the only mollusc for which a whole genome sequence is available (Chapman et al. 2007). In the future researchers interested in paleopolyploidy in the Mollusca should select taxa that diverged shortly after and shortly before the genome duplications we have identified (Figure 3.1). Researchers interested in paleopolyploidy in non-molluscan taxa could use the methods that we have developed in this paper to guide them in their selection of taxa for whole genome sequencing by identifying branches on which the number of chromosomes or members of gene families have doubled.

Chapter 4

Using the Gene Phylogeny to Detect Changes in Gene Duplication and Loss Rates on a Taxon Phylogeny

4.1 Introduction

Gene families are clades of related genes. Some gene families are found in only a single taxon while some are found in every taxon in the tree of life. In any given genome a gene family may consist of only a single gene or as many as hundreds. Gene family diversity is created by two basic processes, gene duplication and taxon speciation. Genes can be duplicated within a genome either through tandem duplication, duplications of large pieces of chromosomes or even whole chromosomes and through the integration of reverse transcribed mDNA sequences; all the descendants of one copy of a gene created by a duplication are paralogous to all the descendants of the other copy (Fitch 1970). When two taxon lineages diverge each gene in the genome of their last common ancestor also diverges into two lineages, one in each descendant taxon lineage; two genes with a single common ancestor in the most recent common ancestor of the genomes in which they are found are called orthologous (Fitch 1970). Orthologs are by definition always found in different genomes. On the other hand all members of a gene family in a single genome are by definition paralogs, but not all paralogs are found in the same genome. Paralogs may occur in different genomes if the most recent common ancestor of those genomes had a pair of paralogous genes, A and B. After undergoing a speciation each of those paralogs would be passed on to both descendant lineages, the descendants of A in one taxon lineage would be orthologous to the descendants of A in the other taxon lineage and paralogous to all the descendants of B in either taxon lineage.

Ohno (1967, 1970) proposed that changes in the size of gene families are extremely important for the evolution of organisms throughout the tree of life, but only recently, as a

number of fully sequenced genomes have become available, has the evidence accumulated to support that view (Lynch and Conery 2003; Taylor and Raes 2004). Several gene families have expanded in a single taxon clade in which they have come to fill a new biological role (Nei and Rooney 2005; Taylor and Raes 2004; Lespinet et al. 2002). Furthermore rates of gene loss are better correlated with expression levels and the severity of knock out mutations than are rates of nucleotide substitution (Krylov et al. 2003). Therefore, it is important to develop methods that allow us to test for different rates of gene lineage gain and loss in order to identify gene families that have undergone major changes in evolutionary patterns and to confirm correlations between gene gains and losses and other biological or ecological changes.

Several methods have been developed that use the birth-death process to calculate the likelihood of different rates of gene duplications and losses on branches of a taxon phylogeny given counts of gene family members in different taxa as data. Lynch and Conery (2003) used the birth-death process to estimates rates of gene gain and loss from gene counts and identified significant interspecies differences in those rates. Hahn et al. (2005) showed how to use gene counts to calculate the likelihood of a set of birth death parameters for which λ and μ are equal and infer branches of a taxon tree on which an exceptionally large number of gene duplications occurred in a given gene family. They developed a program to execute this model (De Bie et al. 2006), and using that software have detected a great deal of variation in the rate of gene turn over between taxon lineages and taxon lineage specific differences in the rates of expansion and loss between different gene families (Demuth and Hahn 2009). Iwasaki and Takagi (2007) developed a birth-death model of gene count evolution with rate variation between branches which could calculate the likelihood of ancestral gene content reconstructions at the nodes of a taxon phylogeny; and they showed that birth-death rates differ between lineages. Cohen and Pupko (2010) have developed a likelihood model based on gene counts that includes parameters for birth, death and horizontal transfer between taxon lineages; and can calculate the posterior probability of the number of gains and losses along a branch of the taxon tree.

It would be ideal if we could include not only gene counts but also the gene tree in our analysis of gene gain and loss. The gene tree contains much information that is lost in gene counts especially regarding the number of reconstructed gene lineages present at the internal nodes of the taxon tree. Analyses done with gene counts alone have difficulty distinguishing whether differences in gene content between two sister taxa are a consequence of gene loss in the lineage with fewer genes or gene gains in the lineage with more genes. However, the pattern of orthologous gene relationships should improve our ability to infer ancestral gene content and thus distinguish gene gains from gene losses, as well as more accurately infer birth-death parameters. In fact, when we compare a gene tree and a taxon tree we can often infer several gene lineages that must have been lost. This inference of lost lineages is possible when we analyze a gene phylogeny, but not when we only use gene counts, and should greatly improve our ability to estimate μ.

In order to compare a gene tree to a taxon tree one must first generate a reconciliation,

a hypothethis about where on the taxon tree the internal nodes of the gene tree occurred. For any particular reconciliation, the nodes of the gene tree that separate paralogous genes occurred on branches of the taxon tree, while those nodes that separate orthologous genes occurred at internal nodes of the taxon tree. The term reconciliation was originally defined by Goodman et al. (1979), who provided an algorithm for inferring the maximum parsimony reconciliation, that is to say the reconciliation with the fewest number of gene duplications and losses. They also noted that incongruence between a gene tree and a taxon tree may be from an incorrect gene phylogeny, and suggested that the number of gene duplications and losses should be added to the number of nucleotide changes when inferring a gene tree by parsimony. Page (1994) pointed out that the gene tree-species tree problem is the same as the host-parasite problem and the taxon tree-area cladogram problem in biogeography and described a parsimony algorithms for comparing the fit of trees. Page and Charleston (1997) suggested that the number of gene duplications and losses inferred from a maximum parsimony reconciliation of multiple gene trees be used as an optimality criterion for inferring the best taxon tree. They also distinguish between paralogous splits that we can be certain happened because of multiple paralogs in a single genome, as oppose to those that we can only infer as a consequence of incongruity between the gene tree and the taxon tree, which may actually be a consequence of either of those trees being incorrect. Cotton and Page (2005) made an interesting use of the maximum parsimony reconciliation, in which they constrained timing of potentially orthologous nodes of many gene trees to the appropriate branches of the taxon tree and used those dates for rate smoothing. By comparing the distribution of node times in those gene trees to simulated gene trees they concluded that birth-death rates have not been constant through time.

Most modern methods of phylogenetic statistical analysis rely on likelihood calculations that allow us to incorporate much of the uncertainty in parameter estimation, phylogeny estimation and character reconstruction into an analysis in either a Bayesian or maximum likelihood context. To incorporate a reconciliation into an analysis of the birth-death process we would have to calculate the probability of a reconciliation given a taxon tree and a set of birth death parameters. Arvestad, Lagergren and their collaborators have shown how to calculate the probability of a reconciliation (Arvestad et al. 2003, 2009), calculate the probability of a gene tree by summing over all possible reconciliations, calculate the maximum likelihood reconciliation and sample from the distribution of possible reconciliations (Arvestad et al. 2004, 2009). They have used their models to infer a gene tree based on a gene sequence alignment and a taxon tree either by summing over all possible reconciliations for each gene tree topology (Arvestad et al. 2004) or by considering the reconciliation a nuisance parameter and juxtaposing the distance between gene nodes in the reconciliation and the number of nucleotide changes along that branch with a relaxed clock (Åkerborg et al. 2009). They have also used their model to infer the posterior probability that a given node in a gene tree is in fact orthologous (Arvestad et al. 2003; Sennblad and Lagergren 2009).

In order to infer differences in the birth-death process between branches of a taxon tree I

implemented a reversible-jump Markov Chain Monte Carlo (MCMC) analysis to estimate the posterior probability of a particular assignment of birth-death parameters to the branches of a taxon tree given the taxon tree and a gene tree. An assignment is the particular combination of taxon branches with the same birth-death rates. I essentially used the same model as (Arvestad et al. 2009) to calculate the probability of a gene tree given a taxon tree and a set of birth-death parameters by summing over every possible reconciliation, except that I dealt with the prior distribution of genes at the root differently. I treated the number of reconstructed gene lineages at the root and the values of the birth-death rates as nuisance parameters and summed over them in the MCMC. In section 4.2 I derive the probability of a gene tree given a taxon tree using a different method than (Arvestad et al. 2009), and a different prior for the root. In section 4.3 I describe the reversible-jump MCMC used to estimate the posterior probability of rate assignments to the branches of the taxon tree.

I wanted to demonstrate that a model which relied on a gene tree had more power to detect differences in the birth-death process than did a model which used only gene counts as data. Therefore I simulated a number of gene trees on a very simple taxon tree using a wide range of birth-death parameters and analyzed them for differences in the birth-death rate using both the gene tree model around which this paper is based and a gene count model. The gene count model calculated the probability of a particular set of gene counts in the same way as Hahn et al. (2005). However, I evaluated the likelihoods using a reversible-jump MCMC that closely matched the one used to evaluate the gene tree model, and I included the parameter space in which λ and μ are not equal. In section 4.4 I describe these simulations, their analyses and the results.

Phylogenies of real gene families rarely have a fully resolved topology. Therefore, it is important to consider the uncertainty in the topology reconstruction, when evaluating the distribution of birth-death rates. I accomplished this by including the gene tree topology as a nuisance parameter in my reversible-jump MCMC. Rather than calculate the probability of the gene tree given the birth-death parameters and the taxon tree, I calculated the probability of the gene sequences, such that the probability of the gene sequences given the gene tree was calculated in the usual way and the prior distribution of the gene tree topology was calculated from the taxon tree and the birth-death parameters. Thus, in estimating the posterior distribution of the assignment of birth-death rates to the branches of the taxon tree, I summed over the uncertainty in the gene tree topology. In section 4.5, I describe how this calculation was made as well as analyses I did on two real gene families, a clade of protein tyrosine kinase genes found in nematodes, fruit flies and humans, and the posterior hox genes from nine bilaterian taxa.

4.2 Model

4.2.1 Definitions

Definition of the Trees

The goal here is to calculate the probability that a phylogeny of genes, G_0, evolved on a given phylogeny of taxa, T_0, given a set of birth-death parameters that apply to the evolution of G_0 on each branch of T_0. Each tree has a topology and may or may not have a set of branch lengths; furthermore, every gene that defines the terminals of G_0 has been found in some taxon that defines the terminals of T_0. We assume that all the genes in G_0 are monophyletic with respect to the other genes found in the taxa of T_0. Let T be any subtree of T_0 and G be any subtree of G_0.

Let g be some point anywhere on G_0. γ is any element of G_0, and any G has three types of elements: branches, $\vec{\gamma}$; nodes, $\overset{\circledast}{\gamma}$; and tips, $\overset{\times}{\gamma}$ (Figure 4.1a). A fourth type of element is called a connector $\overset{\star}{\gamma}_i$ and can be either a tip or a node. Each γ is a set of points, g, on the gene tree. For any given branch, $\vec{\gamma}_i$, the last point in the branch is referred to as $\overset{\star}{g}_i$, and the first point is $\overset{\circ}{g}_i$; at the end of that branch is a connector $\overset{\star}{\gamma}_i$. For any internal branch $\vec{\gamma}_i$ there are two descendant branches, $\vec{\gamma}_{i^+}$ and $\vec{\gamma}_{i^-}$, and the connector at the end of the branch is a node, $\overset{\circledast}{\gamma}_i$. Each node is a set of only three points the last point of the branch immediately preceding it and the first point of each of its descendant branches, $\overset{\circledast}{\gamma}_i = \{\overset{\star}{g}_i, \overset{\circ}{g}_{i^+}, \overset{\circ}{g}_{i^-}\}$. Therefore we see that $\vec{\gamma}_i \cap \overset{\circledast}{\gamma}_i = \{\overset{\star}{g}_i\}$, and $\vec{\gamma}_{i+} \cap \overset{\circledast}{\gamma}_i = \{\overset{\circ}{g}_{i+}\}$. Furthermore the connector at the end of any terminal branch, $\vec{\gamma}_i$, is a tip, $\overset{\times}{\gamma}_i$, which has only one element the last point in $\vec{\gamma}_i$, so that $\overset{\times}{\gamma}_i = \{\overset{\star}{g}_i\}$. I will also define $\vec{\gamma}_r$ as the root of the gene tree, so that $\overset{\circledast}{\gamma}_r$ is the basal node of the gene tree. We can see how these elements are sufficient to describe the entire gene tree, so that $G_0 = \{\text{All } \gamma\}$. Furthermore $\vec{G}$ is the set of all the branches in G, $\overset{\circledast}{G}$ is the set of all the nodes in G, $\overset{\times}{G}$ is the set of all the tips in G, and $\overset{\star}{G}$ is the set of all the connectors in G, so that $\overset{\star}{G} = \overset{\circledast}{G} \cup \overset{\times}{G}$ and $G = \vec{G} \cup \overset{\star}{G}$. $G(\gamma)$ is the subtree of G_0 above γ, including γ, so that $G_0 = G(\vec{\gamma}_r)$.

The taxon tree, T_0, has the same descriptors as the gene tree (Figure 4.1b). t is any point on T_0. τ is an element of T_0: $\vec{\tau}$ is any branch in T_0, $\overset{\circledast}{\tau}$ is any node in T_0, $\overset{\times}{\tau}$ is any tip in T_0, and $\overset{\star}{\tau}$ is any connector in T_0. An internal branch $\vec{\tau}_i$ has two descendant branches, $\vec{\tau}_{i^+}$ and $\vec{\tau}_{i^-}$, and a node at its terminal end, $\overset{\circledast}{\tau}_i$. The last point on branch $\vec{\tau}_i$ is $\overset{\star}{t}_i$ and the first point on the branch is $\overset{\circ}{t}_i$, and $\overset{\circledast}{\tau}_i = \{\overset{\star}{t}_i, \overset{\circ}{t}_{i^+}, \overset{\circ}{t}_{i^-}\}$. $\vec{T}$ is the set of all the branches in T, $\overset{\circledast}{T}$ is the set of all the nodes in T, $\overset{\times}{T}$ is the set of all the tips in T, and $\overset{\star}{T} = \overset{\circledast}{T} \cup \overset{\times}{T}$. $\vec{\tau}_r$ is the root of the taxon tree and $\overset{\circledast}{\tau}_r$ is the basal node. $T(\tau)$ is the subtree of T_0 above τ including τ.

Definition of a Reconciliation

Although we know at what point the terminals of G_0 occurred on T_0, we do not know where exactly all the other gs occurred (Figure 4.2). We will call a particular set of assumptions

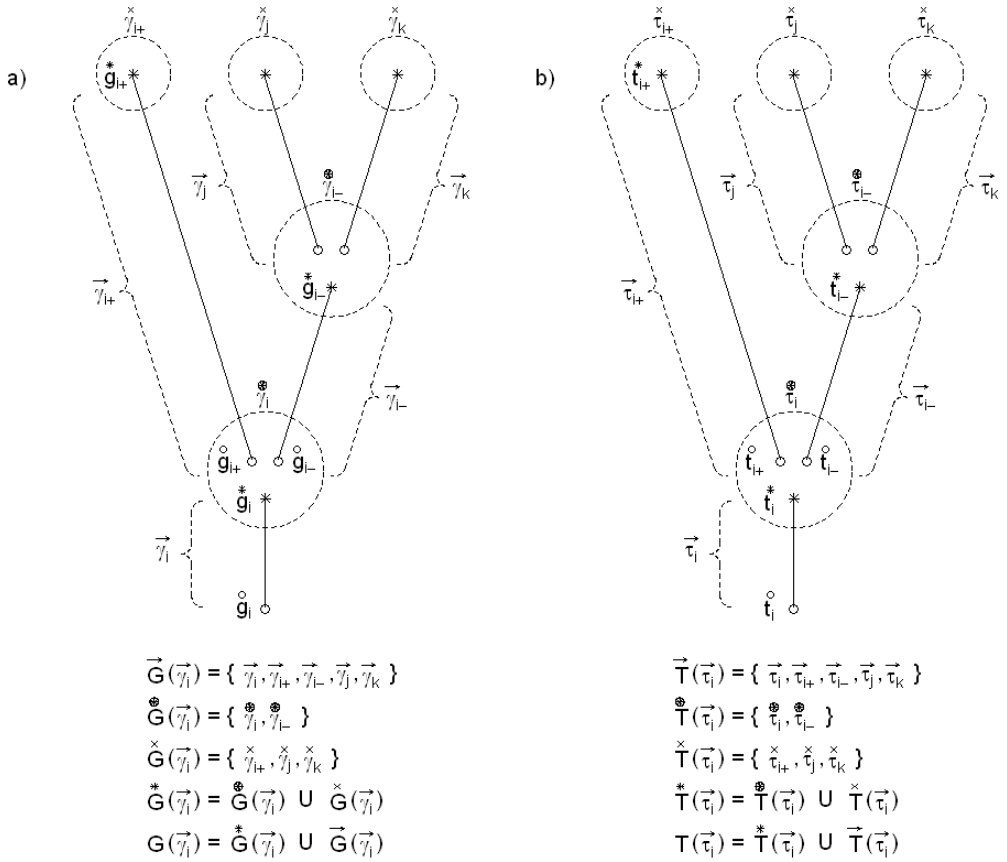

Figure 4.1: The definition of the elements for a) a gene tree, $G(\vec{\gamma}_i)$, and b) a taxon tree tree, $T(\vec{\tau}_i)$. See text for a description of the various elements. The branches $\vec{\gamma}_i$, $\vec{\gamma}_j$, $\vec{\gamma}_k$, $\vec{\tau}_i$, $\vec{\tau}_j$ and $\vec{\tau}_k$ were all named arbitrarily, but the name for every other element of either phylogeny is necessitated by the topology and the names of the other elements.

about where each γ occurred on T_0 a reconciliation, ρ, and we will define $\rho(g)$ as the t where g occurred for reconciliation ρ. $\rho(\gamma)$ is the set of all the elements in T_0 through which γ passes under reconciliation ρ, $\rho(\gamma) = \{\tau : t \in \tau, t = \rho(g), g \in \gamma\}$. Finally $\rho(G)$ is the set of all the $\rho(\gamma)$ for every element in G, $\rho(G) = \{\rho(\gamma) : \gamma \in G\}$, thus $\rho(G_0)$ is a sufficient description of an entire reconciliation. The reconciliations of the tips of the gene tree are taken as data and thus are fixed for all reconciliations, such that $\rho_i(\overset{\times}{\gamma}) = \rho_j(\overset{\times}{\gamma})$ for any particular $\overset{\times}{\gamma}$ if ρ_i and ρ_j are different reconciliations.

We will only concern ourselves with the branch lengths of the taxon tree, and assume that the lengths of the gene tree have no effect on our reconciliation. A more precise way to say it is that no matter where exactly a particular point of the gene tree lies along a single

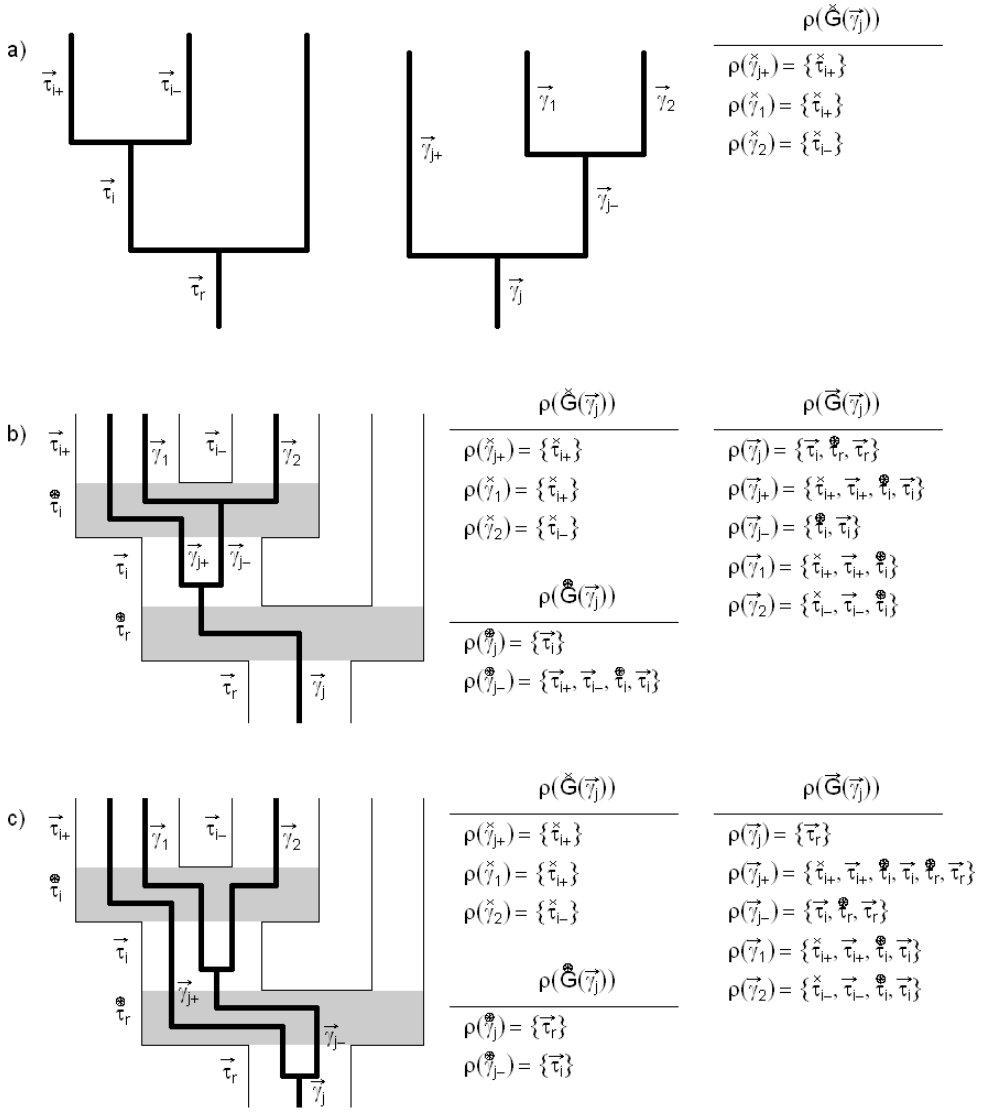

Figure 4.2: Alternative reconciliations of the nodes and branches of a given taxon tree and gene tree and the reconciliations of their tips. a) The taxon tree, T_0, with branch labels to the left of each branch, and a gene tree, $G(\vec{\gamma}_j)$, with branch labels to the right of each branch. The tip reconciliations that define the relationship between the trees are shown in table form. b and c) Two possible reconciliations of these trees based on the reconciliation between the tips of the trees. The gene tree is shown within the taxon tree. Labels for nodes and branches of the taxon tree are shown to the left of each element and labels for the branches of the gene tree are shown to the right of each branch. The tips of the taxon tree are not labeled and are presumed

to occur at the ends of the terminal branches. Nodes of the taxon tree are shaded gray. The reconciliations of all the elements of the gene tree are shown in table form. The reconciliations of the tips of the gene tree do not vary. b) The maximum parsimony reconciliation in which each node of the gene tree occurs in the most recent element of the taxon tree that it could possibly occur. c) An alternative reconciliation in which $\overset{\circledast}{\gamma}_{j}$ occurs on $\vec{\tau}_i$ and $\overset{\circledast}{\gamma}_{j}$ occurs on $\vec{\tau}_r$.

branch of the taxon tree the probability of a reconciliation will be the same; $P(\rho(g)=t_1)=P(\rho(g)=t_2)$, if $\{t_1,t_2\}\subset\vec{\tau}$. It is possible to use a similar method to calculate the probability of the reconciliation in which we do not consider this to be true, but instead calculate the probability of the reconciliation of a node of the gene tree to each point along the node of a taxon tree separately (Arvestad et al. 2003, 2009). Calculating the probability of each of these reconciliations separately would allow us to compare the number of nucleotide changes along a branch of the gene tree to the time at which it occurred on the taxon tree, which is presumably an approximation of the true time (Åkerborg et al. 2009). However, it is computationally much more burdensome, as it requires that the timing of each node of the taxon tree be treated as an independent parameter. Under our method we calculate only the probability that $\{\vec{\tau}\}=\rho(\overset{\circledast}{\gamma})$, and not that $t=\rho(\overset{\star}{g})$, where:

$$P(\{\vec{\tau}\}=\rho(\overset{\circledast}{\gamma}_i))=\int_{t\in\vec{\tau}}P(t=\rho(\overset{\star}{g}_i))\partial t$$

Possible Reconciliations

We will define R as the set of every possible ρ, and $R(\gamma)$ as the set of all elements of the taxon tree in which γ could be found, $R(\gamma)\equiv\bigcup_{\rho\in R}\rho(\gamma)$. Any particular element of the gene tree can not be found on every element in the taxon tree. In particular we know that γ could not be found on any subtree of T_0 where any of the terminals descended from γ are not found in the terminals of that subtree, so that $R(\gamma)\subseteq\{\tau:\bigcup\rho(\overset{\times}{G}(\gamma))\subseteq\overset{\times}{T}(\tau)\}$. In other words any particular element of G_0 must have occurred in a common ancestor of all the taxa in which the descendants of that element were found. There is a hierarchy of dependence for the reconciliations of the different types of elements (Figure 4.2): the reconciliations of the tips of a gene tree, $\rho(\overset{\times}{G})$, are considered part of the data; the reconciliations of the nodes of a gene tree, $\rho(\overset{\circledast}{G})$, are restricted to certain elements of the taxon tree by $\rho(\overset{\times}{G})$; and the reconciliations of the branches of a gene tree, $\rho(\overset{\circledast}{G})$, are completely determined by $\rho(\overset{\circledast}{G})$. Therefore $\rho(\overset{\star}{G})$ is a sufficient description of an entire reconciliation.

Under any particular reconciliation each node in the gene tree, $\overset{\circledast}{\gamma}$, may be either orthologous or paralogous (Figure 4.2). When a gene is duplicated within a genome during the evolution of a lineage, it creates a paralogous node in the gene tree. A $\overset{\circledast}{\gamma}$ is considered paralogous on a $\vec{\tau}$ if $\rho(\overset{\circledast}{\gamma})=\{\vec{\tau}\}$. As with all elements, a $\overset{\circledast}{\gamma}$ can only be paralogous in a common

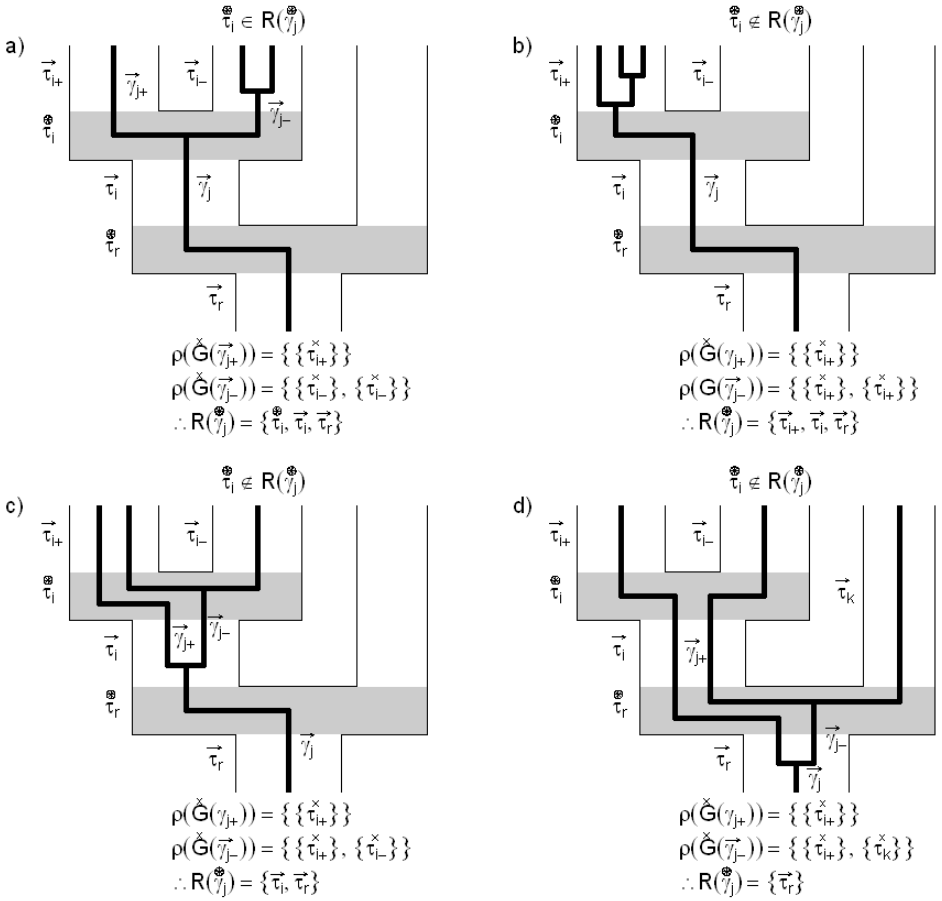

Figure 4.3: Different reconciliations of the tips of a gene tree constrain the possible reconciliations available to the nodes of that tree. Each figure shows a maximum parsimony reconciliation of a three terminal gene tree, $G(\vec{\gamma}_j)$, on a three terminal taxon tree, when we assume different reconciliations for the tips. The reconciliations of the tips of the gene tree and the possible reconciliations of a node of the gene tree, $\overset{\circledast}{\gamma}_j$, are described below each plot. $\overset{\circledast}{\gamma}_j$ can only be orthologous on taxon node $\overset{\circledast}{\tau}_i$ in figure a); in all the other figures $\overset{\circledast}{\gamma}_j$ must be a paralogous divergence, although it could fall on multiple branches. Labels for nodes and branches of the taxon tree are shown to the left of each element and labels for the branches of the gene tree are shown to the right of each branch. The tips of the taxon tree are not labeled and are presumed to occur at the ends of the terminal branches. Nodes of the taxon tree are shaded gray. a) The descendants of each basal branch descended from $\overset{\circledast}{\gamma}_j$ are found in a different descendant clade from $\overset{\circledast}{\tau}_i$, so that $\overset{\circledast}{\gamma}_j$ can be found in that node or in any of the branches

below it. b) All the tips descended from $\overset{\circledast}{\gamma}_j$ are found in $\overset{\times}{\tau}_{i^+}$, so that $\overset{\circledast}{\gamma}_j$ can be found in $\vec{\tau}_{i^+}$ or in any of the branches below it. c) The tips descended from $\overset{\circledast}{\gamma}_j$ are all found in $\overset{\times}{\tau}_{i^+}$ and $\overset{\times}{\tau}_{i^-}$, but the descendants of $\overset{\circledast}{\gamma}_{j^-}$ are also found in both $\overset{\times}{\tau}_{i^+}$ and $\overset{\times}{\tau}_{i^-}$. Therefore $\overset{\circledast}{\gamma}_j$ can be found only in $\vec{\tau}_i$ and the the branch below it. d) The tips descended from $\overset{\circledast}{\gamma}_j$ are found in $\overset{\times}{\tau}_k$ as well as the descendants of $\vec{\tau}_i$. These tips are arranged such that $\overset{\circledast}{\gamma}_j$ can only be found in $\vec{\tau}_r$ and not $\overset{\circledast}{\tau}_r$.

ancestor of all the taxa in which its descendant genes are found. On the other hand, when two lineages diverge in the taxon tree from a speciation, they create an orthologous divergence in every lineage of the gene tree present at that time, such that one of the two new gene lineages is passed on in one of the descendant taxon lineages created by the the speciation and the other new gene lineage is passed on in the other descendant of the speciation. We consider a $\overset{\circledast}{\gamma}$ orthologous in a $\overset{\circledast}{\tau}_i$ for a particular reconciliation if $\overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma})$; in such a case $\rho(\overset{\circledast}{\gamma}) = \{\overset{\circledast}{\tau}_i, \vec{\tau}_i, \vec{\tau}_{i^+}, \vec{\tau}_{i^-}\}$. Any given $\overset{\circledast}{\gamma}$ can only be orthologous in the one $\overset{\circledast}{\tau}$ that is the most recent common ancestor of all the taxa in which its descendant genes are found; and can only be orthologous in even that one node if that $\overset{\circledast}{\gamma}$ creates two clades, one of which has all of its descendants in $T(\vec{\tau}_{i^+})$ and the other of which has all its descendants in $T(\vec{\tau}_{i^-})$ (Figure 4.3).

$$R(\overset{\circledast}{\gamma}_j) \cap \overset{\circledast}{T}_0 = \{\overset{\circledast}{\tau}_i : \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^+})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^+}), \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^-})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^-})\}$$
$$\cup \{\overset{\circledast}{\tau}_i : \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^+})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^-}), \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^-})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^+})\}$$

If a node of the gene tree can be found in any element of the taxon tree, then it can be found in any branch of the taxon tree that is ancestral to that element, but not in any node ancestral to that element.

Every connector in the gene tree, $\overset{\star}{\gamma}$, has a most recent element of the taxon tree on which all of its points can be reconciled (Figures 4.2b and 4.3); we will call that element $\rho_{\text{MRC}}(\overset{\star}{\gamma})$. For every tip of the gene tree $\rho_{\text{MRC}}(\overset{\star}{\gamma})$ is obviously the tip of the taxon tree in which that that gene was found. For a node of the gene tree $\rho_{\text{MRC}}(\overset{\circledast}{\gamma})$ can either be a node of the taxon tree in which it is orthologous or a branch of the taxon tree, if it can not be orthologous.

An individual branch of the gene tree, $\vec{\gamma}$, can be found in any number of elements of the taxon tree. Imagine that a lineage of the gene tree is present during a speciation in the taxon tree, so that it is split into two descendant lineages, one in each of the taxon lineages created by the speciation. If one of these gene lineages is completely lost, then there will be only a single lineage in the reconstructed gene tree spanning the branch of the taxon tree before the speciation, the node of the taxon tree that represents the speciation and the branch of the taxon tree in which the gene lineage survives. In this way, a branch of the gene tree can theoretically pass through many consecutive elements of the taxon tree.

In order for a branch of a gene tree, $\vec{\gamma}_j$, to be present in a given branch of the taxon tree, $\vec{\tau}_i$, it need only be true that all the descendants of $\vec{\gamma}_j$ are found in the descendants of

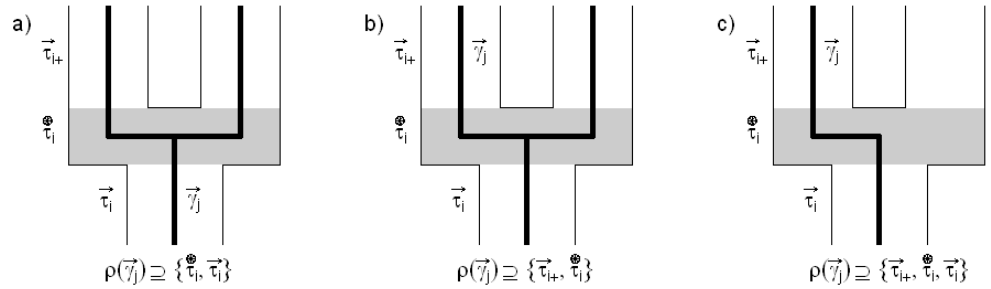

Figure 4.4: Different ways in which a branch of the gene tree can be reconstructed in a node of a taxon tree. Labels for nodes and branches of the taxon tree are shown to the left of each element and labels for the branches of the gene tree are shown to the right of each branch. The tips of the taxon tree are not labeled and are presumed to occur at the ends of the terminal branches. Nodes of the taxon tree are shaded gray. The elements of the taxon tree immediately surrounding $\overset{\circledast}{\tau}_i$ in which $\vec{\gamma}_j$ is reconciled are listed below the figure. a) $\vec{\gamma}_j$ ends in an orthologous gene split in $\overset{\circledast}{\tau}_i$. b) $\vec{\gamma}_j$ starts in an orthologous gene split in $\overset{\circledast}{\tau}_i$. c) $\vec{\gamma}_j$ passes through $\overset{\circledast}{\tau}_i$ as it goes from $\vec{\tau}_i$ to $\vec{\tau}_{i^+}$

$\vec{\tau}_i$. However, in order for $\vec{\gamma}_j$ to be present in a given node of the taxon tree, $\overset{\circledast}{\tau}_i$, one of two things must also be true (Figure 4.4). $\vec{\gamma}_j$ may be found in $\overset{\circledast}{\tau}_i$, if $\overset{\circledast}{\gamma}_j$, the node at the end of $\vec{\gamma}_j$, is orthologous in $\overset{\circledast}{\tau}_i$; we have already discussed what conditions must be met in order for this to be a possibility. $\vec{\gamma}_j$ could also be present in $\overset{\circledast}{\tau}_i$ if it passed from that node into one of its descendant branches, either $\vec{\tau}_{i^+}$ or $\vec{\tau}_{i^-}$; we have also already discussed what must be true in order for $\vec{\gamma}_j$ to be found in a given branch of the taxon tree. These conditions can be summarized as so:

$$R(\vec{\gamma}_j) \cap \overset{\circledast}{T}_0 = \{\overset{\circledast}{\tau}_i : \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^+})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^+})\} \cup \{\overset{\circledast}{\tau}_i : \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^+})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^-})\}$$
$$\cap \{\overset{\circledast}{\tau}_i : \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^-})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^-})\} \cup \{\overset{\circledast}{\tau}_i : \bigcup \rho(\overset{\times}{G}(\vec{\gamma}_{j^-})) \subseteq \overset{\times}{T}(\vec{\tau}_{i^+})\}$$

If a branch of the gene tree can be found in any element of the taxon tree, then it can be found in any other element of the taxon tree that is ancestral to that element.

If for a particular reconstruction $\vec{\gamma}$ starts in $\vec{\tau}_i$ and passes through $\overset{\circledast}{\tau}_i$ into $\vec{\tau}_{i^+}$, then $\{\vec{\tau}_i, \overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma})$ (Figure 4.4c). On the other hand if $\vec{\gamma}$ has its base in a reconstructed orthologous gene split in $\overset{\circledast}{\tau}_i$ and continues on in $\vec{\tau}_{i^+}$, then $\{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma})$, but $\vec{\tau}_i \notin \rho(\vec{\gamma})$ (Figure 4.4b). Furthermore, if $\vec{\gamma}$ is found in $\vec{\tau}_i$ and ends in an orthologous gene node at $\overset{\circledast}{\tau}_i$, then $\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma})$, but $\{\vec{\tau}_{i^+}, \vec{\tau}_{i^-}\} \subseteq \rho^C(\vec{\gamma})$ (Figure 4.4a). Therefore, if $\vec{\gamma}$ is found at the end of $\vec{\tau}_i$, then $\{\vec{\tau}_i, \overset{\circledast}{\tau}_i\} \subseteq \rho(\vec{\gamma})$, whether it continues through into a descendant taxon lineage or ends in an orthologous gene node at $\overset{\circledast}{\tau}_i$ (Figure 4.4a and c); and if $\vec{\gamma}$ is found at the beginning

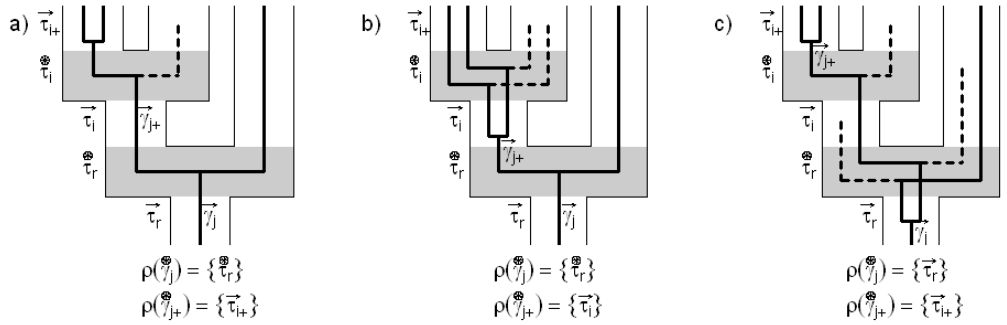

Figure 4.5: Moving nodes of the gene from their maximum parsimony reconciliation position increases the number of events necessary to explain the evolution of the gene tree. All three plots show the same gene tree and taxon tree with the same reconciliation of the gene tree tips, but each plot shows an alternative reconciliation of the gene tree nodes. Labels for nodes and branches of the taxon tree are shown to the left of each element and labels for the branches of the gene tree are shown to the right of each branch. The tips of the taxon tree are not labeled and are presumed to occur at the ends of the terminal branches. Nodes of the taxon tree are shaded gray. Gene lineages which did not survive to the present but we can infer must have been present at some time and then lost are shown as dashed lines. The reconciliations of the nodes are shown in text form below each tree. a) The maximum likelihood reconciliation, we can infer at minimum one gene duplication and one gene loss. b) $\overset{\circledast}{\gamma}_{j^+}$ has been moved down to $\vec{\tau}_i$ from a paralogous position on $\vec{\tau}_{i^+}$. This increases the number of inferred gene losses by one. c) $\overset{\circledast}{\gamma}_j$ has been moved down to $\vec{\tau}_r$ from an orthologous position on $\overset{\circledast}{\tau}_r$. This increases the number of inferred gene losses by two and the number of gene duplications by one.

of $\vec{\tau}_{i^+}$, then $\{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma})$, whether $\vec{\gamma}$ survived through from an ancestral taxon lineage or starts in an orthologous gene node at $\overset{\circledast}{\tau}_i$ (Figure 4.4b and c).

Maximum Parsimony Reconciliation

If we were to compare reconciliations using maximum parsimony as an optimum criterion it is clear which reconciliation would be best (Goodman et al. 1979) (Figure 4.5). Every potentially orthologous gene node would be found as an ortholog in the most recent common ancestor of all the taxa within which the descendants of the gene node were found. Gene nodes which can not be orthologous would be found in the branch immediately ancestral to that most recent common ancestor. Branches would obviously be spread between the nodes that define them. Any other reconciliation would be worse as it would add unnecessary events. If a node is found on a branch that is not the most recent possible, then there will

have to be additional deaths of gene lineages because as a branch of the gene tree passes through an additional node of the taxon tree, a gene lineage must die in the taxon clade in which the reconstructed gene lineage is not found. Thus if ρ is the maximum parsimony reconciliation, then $\rho_{\text{MRC}}(\dot{\gamma}) \in \rho(\dot{\gamma})$ for every $\dot{\gamma}$.

Number of Reconstructed Lineages

A value of particular interest is the number of reconstructed lineages in a gene subtree that are alive at a certain point on the taxon tree under a given reconciliation. I will call the number of lineages in tree $G(\gamma_j)$ alive at the end of taxon branch $\vec{\tau}_i$ that survive to the present, $\dot{n}_i(\gamma_j)$, and the number alive at the start of that branch $\mathring{n}_i(\gamma_j)$. So that $\dot{n}_i(\gamma_j) = \left|\{g : g \in \bigcup G(\gamma_j), \rho(g) = \dot{t}_i\}\right|$, and $\mathring{n}_i(\gamma_j) = \left|\{g : g \in \bigcup G(\gamma_j), \rho(g) = \mathring{t}_i\}\right|$. Furthermore, I will call the number of tips in $G(\gamma_j)$ alive in the present $\check{n}_0(\gamma_j)$, so that $\check{n}_0(\gamma_j) = \left|\check{G}(\gamma_j)\right|$. I will also introduce a term $\dot{N}_i(\gamma)$, the number of gene lineages descended from γ, that are reconstructed or not and that are alive at the end of taxon branch τ_i. This value has no equivalent relationship to elements of the gene tree as the gene tree only consists of reconstructed lineages; however we can be certain that $\dot{N}_i(\gamma) \geq \dot{n}_i(\gamma)$.

It is also important to know the minimum and maximum values that $\dot{n}_i(\gamma_j)$ can take. At any point on the taxon tree in which γ_j could be reconstructed $\dot{n}_i(\gamma_j)$, can be no greater than $\check{n}_0(\gamma_j)$; if it were, then some lineages would have to be lost between that time and the present and that is impossible for reconstructed lineages. I will call the minimum number of reconstructed lineages descended from $\vec{\gamma}_j$ present at $\dot{t}_i$, $\dot{n}_i^{\min}(\gamma_j)$. If $\vec{\gamma}_j$ could be reconstructed at any element descended from $\dot{t}_i$ then the node at the end of $\vec{\gamma}_j$ could also have occurred on that later element, and so there may still be as few as one gene lineage at $\dot{t}_i$. On the other hand, if $\vec{\gamma}_j$ could not be reconstructed at $\circledast\vec{\tau}_i$, then $\circledast\vec{\gamma}_j$ must also occur before $\circledast\vec{\tau}_i$, and so $\vec{\gamma}_j$ must have at least two descendants at the end of $\vec{\tau}_i$. Furthermore, if the descendant branches from $\vec{\gamma}_j$ must also have more than one descendant each at $\dot{t}_i$, then $\vec{\gamma}_j$ can have no fewer descendants, than the sum of the minimum number of descendants of each of its descendant branches.

$$\dot{n}_i^{\min}(\gamma_j) = \begin{cases} 1 & \text{if } \circledast\vec{\tau}_i \in R(\vec{\gamma}_j) \\ \dot{n}_i^{\min}(\gamma_{j^+}) + \dot{n}_i^{\min}(\gamma_{j^-}) & \text{if } \circledast\vec{\tau}_i \notin R(\vec{\gamma}_j) \end{cases}$$

Equivalent Trees

The taxon tree is a labeled phylogeny, meaning that each tip of the tree is different from every other tip, so that two trees with the same topology can be different if the distribution of their tip labels are different and no two different subtrees of the same tree can be equivalent as they must have different labels. In contrast to a labeled phylogeny, in an unlabeled phylogeny, all tips are only distinguished by their topology, so that all trees and subtrees of any one tree

with the same topology are equivalent to each other. The gene tree falls somewhere between a labeled and an unlabeled phylogeny, so that we might call it "semi-labeled". Not every tip of the gene tree is completely different from every other tip, but at the same time not every tip is the same. Tips of the gene tree are only equivalent if they are found in the same tip of the taxon tree. Trees can also be distinguished from each other by their topology. Thus, different subtrees can be equivalent if they have the same topology and their terminals were found in the same terminals of the taxon tree.

We can determine if two subtrees of the gene tree are equivalent using these simple rules. Tips of the gene tree are equivalent if the genes that define those tips were found in the same taxon, $G(\overset{\times}{\gamma}_i) \cong G(\overset{\times}{\gamma}_j) \leftrightarrow \rho(\overset{\times}{\gamma}_i) = \rho(\overset{\times}{\gamma}_j)$. Trees that start in a node are equivalent if each of the clades descendant from one of those nodes are equivalent to one of the clades descended from the other node,

$$G(\overset{\circledast}{\gamma}_i) \cong G(\overset{\circledast}{\gamma}_j) \leftrightarrow G(\vec{\gamma}_{i^+}) \cong G(\vec{\gamma}_{j^+}) \text{ and } G(\vec{\gamma}_{i^-}) \cong G(\vec{\gamma}_{j^-})$$
$$\text{or } G(\vec{\gamma}_{i^+}) \cong G(\vec{\gamma}_{j^-}) \text{ and } G(\vec{\gamma}_{i^-}) \cong G(\vec{\gamma}_{j^+})$$

Of course trees descended from a branch are equivalent if the trees descended from their connectors are equivalent, $G(\vec{\gamma}_i) \cong G(\vec{\gamma}_j) \leftrightarrow G(\overset{\star}{\gamma}_i) \cong G(\overset{\star}{\gamma}_j)$. It goes without saying that a subtree is equivalent to itself.

4.2.2 The Probability of a Gene Being Lost

The first step in calculating the probability of a gene tree evolving on a taxon tree is to calculate the probability that a single gene present at some point on the taxon tree left no descendants in the sampled taxa. We will define $\overset{\star}{E}_i(t_i)$ as the probability that a single gene lineage alive at t_i on $\vec{\tau}_i$ will have no descendants alive at $\overset{\star}{t}_i$. Furthermore we will define $E_0(t_i)$ as the probability that a gene at t_i will leave no descendants in any of the terminals of T_0; $E_0(t_i)$ is the same as the probability that no descendant genes were found in the tips of $T(\vec{\tau}_i)$, because we are assuming that there is no horizontal transfer of genes, so that a gene at t_i could not leave descendants at any point of T_0 that is not descended from t_i. The reasoning behind (2.10) assumes only that a TVBD applies between t_k and t_j and that all lineages alive at time t_j will have the same probability of being lost by time t_i. Therefore:

$$1 - E_0(t_i) = \frac{(1 - E_0(\overset{\star}{t}_i))(1 - \overset{\star}{E}_i(t_i))}{1 - \overset{\star}{B}_i(t_i)E_0(\overset{\star}{t}_i)} \tag{4.1}$$

where $\overset{\star}{B}_i(t_i)$ is the probability of one reconstructed lineage at t_i leaving more than one lineage at $\overset{\star}{t}_i$. If we assume that a CRBD operates on each branch of the taxon tree, then we can use (2.19) and (2.20) to calculate $\overset{\star}{E}_i(t_i)$ and $\overset{\star}{B}_i(t_i)$ respectively. We could of course also use the DTBD to solve for these values under any TVBD.

When $\overset{\star}{\tau}_i \subset \overset{\times}{T}_0$ then $E_0(\overset{\star}{t}_j) = 0$, but when $\overset{\star}{\tau}_i \subset \overset{\circledast}{T}_0$ then $E_0(\overset{\star}{t}_j)$ can not be calculated so easily. The probability that a single gene lineage present at the end of some internal branch on the taxon tree, $\overset{\star}{t}_i$, is not found in any of the terminals of the taxon tree is the probability that the gene was lost in both descendant lineages.

$$E_0(\overset{\star}{t}_i) = E_0(\mathring{t}_{i^+}) \times E_0(\mathring{t}_{i^-}) \tag{4.2}$$

Once we know $E_0(\overset{\star}{t}_i)$, we can use (4.1) to calculate $E_0(t_i)$ for any t_i in $\vec{\tau}_i$. Furthermore, $\mathring{t}_{i^+}$ and $\mathring{t}_{i^-}$ will at the base of $\vec{\tau}_{i^+}$ and $\vec{\tau}_{i^-}$ respectively, and so we can calculate the probability of a lineage being lost from either of those points using (4.1). Therefor, we can calculate $E_0(t)$ for any t by starting at the tips of the taxon tree and proceeding backwards down the tree, using (4.1) to calculate the values of E_0 for the points at the base of the branches and (4.2) to calculate E_0 for the points at the end of each internal branch.

4.2.3 Probability of a Reconciliation

The next step is to show how to calculate the probability of a gene tree and a given reconciliation of that tree on the taxon tree. As I mentioned before, it is possible to calculate this probability assuming that the paralogous nodes of the gene tree occurred at specific times along the branches of the gene tree (Arvestad et al. 2009), which would be useful for comparing the amount of genetic change in the genes to the actual time that has elapsed (Åkerborg et al. 2009). However, here I will focus only on the branch of the taxon tree on which a certain gene node occurred, not on the exact time along that branch. I will show how to calculate certain elements of this probability. These elements could be combined into the total probability of a reconciliation fairly easily, but I will not show how to make that calculation here, as we are uninterested in that probability. These elements will instead be used in the next section to show how to calculate the probability of a gene tree summed over all possible reconstructions.

Probability Mass of Reconstructed Gene Lineages at a Taxon Node

Let us imagine a reconciliation of G_0 on T_0 in which a given $\vec{\gamma}$ is present at the base of some branch on the taxon tree $\vec{\tau}_{i^+}$. We want to know the probability that $\vec{\gamma}$ left $\overset{\star}{n}_{i^+}(\vec{\gamma})$ reconstructed lineages at the end of $\vec{\tau}_{i^+}$. We can see that (2.35), (2.36) and (2.11) will apply along $\vec{\tau}_{i^+}$ so long as a TVBD operates between the times when we count the number of reconstructed lineages, even if the TVBD does not operate between the end of that branch and the time at which we observe the lineages. Therefore, we can use (2.36) to calculate the probability that a single reconstructed gene lineage alive at the base of $\vec{\tau}_{i^+}$ will leave $\overset{\star}{n}_{i^+}$ reconstructed lineages at the end of that branch.

$$P(\overset{\star}{n}_{i^+}(\vec{\gamma}) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma})) = (1 - B_0(\mathring{t}_{i^+}, \overset{\star}{t}_{i^+}))(B_0(\mathring{t}_{i^+}, \overset{\star}{t}_{i^+}))^{\overset{\star}{n}_{i^+}(\vec{\gamma})-1} \tag{4.3}$$

We can then use (2.11) to calculate $B_0(\mathring{t}_{i^+}, \overset{\star}{t}_{i^+})$.

$$B_0(\mathring{t}_i, \overset{\star}{t}_i) = \frac{\overset{\star}{B}_i(\mathring{t}_i)(1 - E_0(\overset{\star}{t}_i))}{1 - \overset{\star}{B}_i(\mathring{t}_i)E_0(\overset{\star}{t}_i)} \tag{4.4}$$

If we assume that a CRBD operates on each branch, then we can calculate $\overset{\star}{B}_i(\mathring{t}_i)$ using (2.20) and we can calculate $E_0(\overset{\star}{t}_i)$ using the method described in subsection 4.2.2.

Probability of a Gene Topology on a Taxon Branch

The probability in section 4.2.3 does not solve the entire probability for the portion of a gene tree reconstructed on a branch of the taxon tree. We not only need to calculate the probability of a single gene lineage leaving a certain number of reconstructed lineages at some later time, but also the probability of the branching pattern for that portion of the tree. The probability that a node with n descendant lineages splits into two clades with m and $n-m$ lineages is the same for any positive value of m less than n (Slowinski and Guyer 1993). Therefore the probability of any particular split between lineages for a given node of size n is $1/(n-1)$, when the two descendant gene lineages are distinguished from each other.

$$P(\overset{\star}{n}_i(\vec{\gamma}_{j^+})|\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j), \rho(\overset{\circledast}{\gamma}_j) = \vec{\tau}_i) = \frac{1}{\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j) - 1} \tag{4.5}$$

We can multiply together this probability for each node and (4.3) in order to calculate the probability of the portion of the gene tree starting with one lineage at $\mathring{t}_i$ and ending at $\overset{\star}{t}_i$.

Probability of a Reconciliation at a Taxon Node

If we have a reconstructed gene lineage at some point t_i in the middle of a branch $\vec{\tau}_i$, then we know that that lineage will still be present at the instant immediately after t_i, as it must survive from then until the present. However, if that same lineage is present at the base in a node $\overset{\star}{t}_i$, then there are two points immediately after that point, $\mathring{t}_{i^+}$ and $\mathring{t}_{i^-}$. It need not survive from the base of both those branches until the present; it only has to survive in one in order to be observed. For a reconciliation in which $\{\vec{\tau}_i, \overset{\circledast}{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)$, if $\overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma}_j)$, then $\overset{\circledast}{\gamma}_j$ is an orthologous divergence and its descendants must survive to the present in the descendants of $\vec{\tau}_{i^+}$ and $\vec{\tau}_{i^-}$. On the other hand if $\vec{\tau}_{i^+} \in \rho(\vec{\gamma}_j)$, then its descendants will survive in the descendants of $\vec{\tau}_{i^+}$, but not the descendants of $\vec{\tau}_{i^-}$; the opposite would obviously be true if $\vec{\tau}_{i^-} \in \rho(\vec{\gamma}_j)$.

For a given $\vec{\gamma}$ that occurs at a $\overset{\circledast}{\tau}_i$ we can calculate the probability of $\overset{\circledast}{\gamma}$ being an orthologous duplication at $\overset{\circledast}{\tau}_i$ as the probability that it survived in both descendant lineages, given that it survived in at least one (Figure 4.4a).

$$P(\overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma})|\{\vec{\tau}_i, \overset{\circledast}{\tau}_i\} \subseteq \rho(\vec{\gamma})) = \frac{(1 - E_0(\mathring{t}_{i^+}))(1 - E_0(\mathring{t}_{i^-}))}{(1 - E_0(\overset{\star}{t}_i))} \tag{4.6}$$

On the other hand, if we have a $\vec{\gamma}$ that occurs at a $\overset{\circledast}{\tau}_i$, we can calculate the probability that it would survive in the descendants of $\vec{\tau}_{i^+}$ but not the descendants of $\vec{\tau}_{i^-}$, and thus not form a node in the reconstructed gene tree (Figure 4.4c), as:

$$P(\vec{\tau}_{i^+} \in \rho(\vec{\gamma}) | \{\vec{\tau}_i, \overset{\circledast}{\tau}_i\} \subseteq \rho(\vec{\gamma})) = \frac{(1 - E_0(\mathring{t}_{i^+}))E_0(\mathring{t}_{i^-})}{(1 - E_0(\overset{\star}{t}_i))} \tag{4.7}$$

It is obvious how we can reverse this equation for the case in which $\vec{\gamma}$ survived in the descendants of $\vec{\tau}_{i^-}$ but not the descendants of $\vec{\tau}_{i^+}$.

4.2.4 Probability of a Gene Tree

We now have enough information to calculate the probability of the entire reconstructed gene tree given the species tree by summing over all the possible reconstructions.

$$P(G_0|T_0) = \sum_{\rho \in R} P(\rho(G_0)|T_0)$$

However, summing over each individual reconstruction would be computationally burdensome. Instead here I will demonstrate a more efficient way of calculating that sum.

Probability of a Gene Tree Starting at the Base of a Branch of the Taxon Tree

We will start with a reconstructed gene lineage that is present at the base of a branch of the taxon tree. In order to calculate the probability of the gene tree descended from this lineage, we can sum the probability of this gene tree evolving with a certain number of reconstructed lineages at the end of this branch over all the possible numbers of lineages at the end of the branch. We already determined the maximum and minimum number of descendant lineages at the end of a branch in section 4.2.1.

$$P(G(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_j)) = \sum_{m=\overset{\star}{n}_{i^+}^{\min}(\vec{\gamma}_j)}^{\overset{\times}{n}_0(\vec{\gamma}_j)} P(G(\vec{\gamma}_j), \overset{\star}{n}_{i^+}(\vec{\gamma}_j) = m | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_j)) \tag{4.8}$$

The probability that a gene subtree, $G(\vec{\gamma}_j)$, which starts with a single lineage at the base of $\vec{\tau}_{i^+}$, will evolve with $\overset{\star}{n}_{i^+}(\vec{\gamma}_j)$ reconstructed lineages at the end of the branch can then be broken down into the probability of the gene tree evolving given that there were $\overset{\star}{n}_{i^+}(\vec{\gamma}_j)$ reconstructed lineages at the end of $\vec{\tau}_{i^+}$, and the probability of their being $\overset{\star}{n}_{i^+}(\vec{\gamma}_j)$ reconstructed lineages at the end of $\vec{\tau}_{i^+}$.

$$\begin{aligned} &P(G(\vec{\gamma}_j), \overset{\star}{n}_{i^+}(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_j)) \\ &\quad = P(G(\vec{\gamma}_j) | \overset{\star}{n}_{i^+}(\vec{\gamma}_j), \vec{\tau}_{i^+} \in \rho(\vec{\gamma}_j)) P(\overset{\star}{n}_{i^+}(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_j)) \end{aligned} \tag{4.9}$$

We can calculate the probability of their being $\overset{\star}{n}_{i^+}(\vec{\gamma}_j)$ reconstructed lineages at the end of $\vec{\tau}_{i^+}$ using (4.3). The probability of the gene tree evolving given that there were $\overset{\star}{n}_{i^+}(\vec{\gamma}_j)$ reconstructed lineages at the end of $\vec{\tau}_{i^+}$ does not depend on where exactly $\vec{\gamma}_j$ started on the branch only that it is on the branch, because we have already determined the number of lineages at the end of $\vec{\tau}_{i^+}$, so all that affects the portion of the probability on that branch is the topology of the gene tree, which will not be affected by the amount of time that has passed.

If we know that a gene lineage $\vec{\gamma}_j$ is present in $\vec{\tau}_i$ and has left a single reconstructed lineage at the end of that branch, then that single lineage is also $\vec{\gamma}_j$ and we can assume that it was present in $\overset{\star}{\tau}_i$, the element at the end of $\vec{\tau}_i$. On the other hand if we know that a gene lineage $\vec{\gamma}_j$ was present in $\vec{\tau}_i$ and has left more than one reconstructed lineage at the end of that branch, then we know that $\overset{\circledast}{\gamma}_j$ must also have occurred in $\vec{\tau}_i$ and that it will have the same number of descendant lineages at $\overset{\circledast}{\tau}_i$, as $\vec{\gamma}_j$ has. Therefore:

$$P(G(\vec{\gamma}_j)|\overset{\star}{n}_i(\vec{\gamma}_j),\vec{\tau}_i \in \rho(\vec{\gamma}_j)) = \begin{cases} P(G(\vec{\gamma}_j)|\{\overset{\star}{\tau}_i,\vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) & \text{if } \overset{\star}{n}_i(\vec{\gamma}_j)=1 \\ P(G(\overset{\circledast}{\gamma}_j)|\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j), \rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\}) & \text{if } \overset{\star}{n}_i(\vec{\gamma}_j)>1 \end{cases} \tag{4.10}$$

I will show how to calculate the probability of a gene tree starting at the end of a taxon tree branch in section 4.2.4.

We can calculate the probability of a gene tree descended from a node $G(\overset{\circledast}{\gamma}_j)$ given that the node is found on $\vec{\tau}_i$ and has $\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j)$ reconstructed lineages descended from it at $\overset{\circledast}{\tau}_i$, by summing over all the possible number of descendant lineages at $\overset{\circledast}{\tau}_i$ left by each of the branches descended from $\overset{\circledast}{\gamma}_j$. The number of lineages descended from $\overset{\circledast}{\gamma}_{j^+}$ and $\overset{\circledast}{\gamma}_{j^-}$ must sum to $\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j)$, furthermore neither branch can have fewer than one descendant or more descendants than it has descendant genes in the present.

$$P(G(\overset{\circledast}{\gamma}_j)|\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j), \rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\}) = \sum_{\overset{\star}{n}_i(\vec{\gamma}_{j^+})=\alpha(\overset{\circledast}{\gamma}_j,\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j))}^{\omega(\overset{\circledast}{\gamma}_j,\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j))} P(G(\overset{\circledast}{\gamma}_j), \overset{\star}{n}_i(\vec{\gamma}_{j^+})|\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j), \rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\}) \tag{4.11}$$

where

$$\alpha(\overset{\circledast}{\gamma}_j, \overset{\star}{n}_i(\overset{\circledast}{\gamma}_j)) \equiv \max(\overset{\star}{n}_i^{\min}(\vec{\gamma}_{j^+}), \overset{\star}{n}_i(\overset{\circledast}{\gamma}_j) - \overset{\times}{n}_0(\vec{\gamma}_{j^-}))$$

and

$$\omega(\overset{\circledast}{\gamma}_j, \overset{\star}{n}_i(\overset{\circledast}{\gamma}_j)) \equiv \min(\overset{\times}{n}_0(\vec{\gamma}_{j^+}), \overset{\star}{n}_i(\overset{\circledast}{\gamma}_j) - \overset{\star}{n}_i^{\min}(\vec{\gamma}_{j^-}))$$

because not only are the values of $\overset{\star}{n}_i(\vec{\gamma}_{j^+})$ constrained by their own minimum and maximum values, but they also can not be so small or so great that $\overset{\star}{n}_i(\vec{\gamma}_{j^-})$ does not fall between its minimum and maximum given a value for $\overset{\star}{n}_i(\overset{\circledast}{\gamma}_j)$.

We can then brake each of these probabilities down into the probability that the gene tree would evolve given the number of lineages descended from $\vec{\gamma}_{i^+}$ and $\vec{\gamma}_{i^-}$ at $\overset{\circledast}{\tau}_i$ and the probability that there would be that many lineages descended from $\vec{\gamma}_{i^+}$ and $\vec{\gamma}_{i^-}$ given that

$\dot{n}_i(\overset{\circledast}{\gamma}_j)$ were descended from $\overset{\circledast}{\gamma}_i$.

$$\begin{aligned} &P(G(\overset{\circledast}{\gamma}_j),\dot{n}_i(\vec{\gamma}_{j^+})|\dot{n}_i(\overset{\circledast}{\gamma}_j),\rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\}) \\ &\quad =P(G(\overset{\circledast}{\gamma}_j)|\dot{n}_i(\vec{\gamma}_{j^+}),\dot{n}_i(\vec{\gamma}_{j^-}),\rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\})P(\dot{n}_i(\vec{\gamma}_{j^+})|\dot{n}_i(\overset{\circledast}{\gamma}_j),\rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\}) \end{aligned} \tag{4.12}$$

Since $\dot{n}_i(\overset{\circledast}{\gamma}_j)=\dot{n}_i(\vec{\gamma}_{j^+})+\dot{n}_i(\vec{\gamma}_{j^-})$, knowing $\dot{n}_i(\overset{\circledast}{\gamma}_j)$ and $\dot{n}_i(\vec{\gamma}_{j^+})$ is the same as knowing $\dot{n}_i(\vec{\gamma}_{j^+})$ and $\dot{n}_i(\vec{\gamma}_{j^-})$. We can calculate the second probability in this equation using (4.5).

The probability of the gene tree above node $\overset{\circledast}{\gamma}_j$ which is found on taxon branch $\vec{\tau}_i$ when we know the number of reconstructed lineages descended from $\vec{\gamma}_{j^+}$ and $\vec{\gamma}_{j^-}$ at $\overset{\circledast}{\tau}_i$, is the same as probability that the gene tree above $\vec{\gamma}_{j^+}$ and the gene tree above $\vec{\gamma}_{j^-}$ would evolve given the number of reconstructed lineages they have at $\overset{\circledast}{\tau}_i$. We must also include another factor in this calculation, as the assignment of the two branches descended from $\overset{\circledast}{\gamma}_j$ to $\vec{\gamma}_{j^+}$ and $\vec{\gamma}_{j^-}$ is arbitrary. The tree above $\overset{\circledast}{\gamma}_j$ could have evolved no matter which of its descendant branches evolved into $G(\vec{\gamma}_{j^+})$ so long as the other descendant branch evolved into $G(\vec{\gamma}_{j^-})$. Thus we must multiply this probability by two in order to account for both possibilities. However, if $G(\vec{\gamma}_{j^+})$ and $G(\vec{\gamma}_{j^-})$ are equivalent then we should not multiply by two as both descendant branches have only one tree that they can evolve into. To account for this factor, I will introduce the new term $k(\overset{\circledast}{\gamma}_j)$, where:

$$k(\overset{\circledast}{\gamma}_j) \equiv \begin{cases} 1 & \text{if } G(\vec{\gamma}_{j^+}) \cong G(\vec{\gamma}_{j^-}) \\ 2 & \text{if } G(\vec{\gamma}_{j^+}) \ncong G(\vec{\gamma}_{j^-}) \end{cases}$$

We can now use this to calculate the probability.

$$\begin{aligned} &P(G(\overset{\circledast}{\gamma}_j)|\dot{n}_i(\vec{\gamma}_{j^+}),\dot{n}_i(\vec{\gamma}_{j^-}),\rho(\overset{\circledast}{\gamma}_j)=\{\vec{\tau}_i\}) \\ &\quad = k(\overset{\circledast}{\gamma}_j)P(G(\vec{\gamma}_{j^+})|\dot{n}_i(\vec{\gamma}_{j^+}),\vec{\tau}_i\in\rho(\vec{\gamma}_{j^+}))P(G(\vec{\gamma}_{j^-})|\dot{n}_i(\vec{\gamma}_{j^-}),\vec{\tau}_i\in\rho(\vec{\gamma}_{j^-})) \end{aligned} \tag{4.13}$$

These two probabilities can be solved using (4.10) creating a loop that will allow us to solve for all the probabilities of gene nodes on a single branch of the taxon tree up to the node at the end of that branch.

Probability of A Gene Tree Starting at the End of a Branch of the Taxon Tree

The probability of a gene tree starting at the end of a branch of the taxon tree depends on what type of element is at the end of the taxon branch. If there is a node at the end of the branch, then the gene lineage will survive on into at least one of the descendant lineages of the taxon tree. On the other hand if there is a tip at the end of the branch of the taxon tree, then the gene lineage has reached the present and must be a gene found in the taxon that is represented by that tip of the taxon tree.

$$P(G(\vec{\gamma}_j)|\{\dot{\tau}_i,\vec{\tau}_i\}\subseteq\rho(\vec{\gamma}_j)) = \begin{cases} P(G(\vec{\gamma}_j)|\{\overset{\circledast}{\tau}_i,\vec{\tau}_i\}\subseteq\rho(\vec{\gamma}_j)) & \text{if } \dot{\tau}_i\in\overset{\circledast}{T}_0 \\ P(G(\vec{\gamma}_j)|\{\overset{\times}{\tau}_i,\vec{\tau}_i\}\subseteq\rho(\vec{\gamma}_j)) & \text{if } \dot{\tau}_i\in\overset{\times}{T}_0 \end{cases} \tag{4.14}$$

Calculating the probability of a gene subtree that has reached a terminal is trivial. If the gene subtree consists of a single terminal lineage that ends with its tip in that tip of the taxon tree, then we are certain that it would evolve as such. On the other hand, if a gene tree has multiple tips or finishes in another tip of the taxon tree, then there is no way that the gene tree could have evolved once we have reached the present. Therefor:

$$P(G(\vec{\gamma}_j)|\overset{\times}{\tau}_i \in \rho(\vec{\gamma}_j)) = \begin{cases} 1 & \text{if } \overset{\times}{\tau}_i \in \rho(\overset{\times}{\gamma}_j)) \\ 0 & \text{if } \overset{\times}{\tau}_i \notin \rho(\overset{\times}{\gamma}_j)) \end{cases} \tag{4.15}$$

Thus we complete our calculation of a gene subtree evolving on the taxon tree.

Once a reconstructed gene lineage reaches a node of the taxon tree it will split into two lineages, one in each of the descendant branches of the taxon tree. Then one of two things can happen: both gene lineages can survive leaving an orthologous node in the gene tree; or the gene lineage can go extinct in one of the descendant lineages of the taxon tree, so that the original branch of the gene tree will continue on in the other descendant lineage of the taxon tree. For any branch of an actual gene tree that could be reconciled to a given node of the taxon tree only one of those options could have occurred, and we can determine which one by using the rules established in section 4.2.1 (Figure 4.4). We can calculate the probability of a gene tree given that the gene lineage at the base of the gene tree is reconciled to the base of a node of the taxon tree by substituting in the probability that corresponds to a possible reconciliation of the gene tree.

$$P(G(\vec{\gamma}_j)|\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) = \begin{cases} P(G(\overset{\circledast}{\gamma}_j), \overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma}_j)|\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) & \text{if } \overset{\circledast}{\tau}_i \in R(\overset{\circledast}{\gamma}_j) \\ P(G(\vec{\gamma}_j), \vec{\tau}_{i^+} \in \rho(\vec{\gamma}_j)|\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) & \text{if } \vec{\tau}_{i^+} \in R(\vec{\gamma}_j) \\ P(G(\vec{\gamma}_j), \vec{\tau}_{i^-} \in \rho(\vec{\gamma}_j)|\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) & \text{if } \vec{\tau}_{i^-} \in R(\vec{\gamma}_j) \end{cases} \tag{4.16}$$

These options will cover all the possible ways that $\vec{\gamma}_j$ could be reconciled to $\overset{\circledast}{\tau}_i$ and $\vec{\tau}_i$ (Figure 4.4).

In order to calculate the probability that a lineage present at the base of $\overset{\circledast}{\tau}_i$ will leave a reconstructed orthologous gene divergence at $\overset{\circledast}{\tau}_i$ and the gene subtree above $\overset{\circledast}{\gamma}_j$ we should separate it into the probability of the gene subtree above $\overset{\circledast}{\gamma}_j$ given that $\overset{\circledast}{\gamma}_j$ is orthologous in $\overset{\circledast}{\tau}_i$ and the probability that $\overset{\circledast}{\gamma}_j$ is orthologous in $\overset{\circledast}{\tau}_i$ given that $\vec{\gamma}_j$ is present at he base of $\overset{\circledast}{\tau}_i$.

$$P(G(\overset{\circledast}{\gamma}_j), \overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma}_j)|\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) = P(G(\overset{\circledast}{\gamma}_j)|\overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma}_j))P(\overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma}_j)|\{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) \tag{4.17}$$

The probability of the gene lineage forming a reconstructed orthologous node in the node of the taxon tree can be calculated using (4.6). We can subdivide the probability of a gene tree given that it starts as an orthologous node into the probabilities of the two subtrees descended from the reconstructed node of the gene tree evolving in their respective clades of the taxon tree.

$$P(G(\overset{\circledast}{\gamma}_j)|\overset{\circledast}{\tau}_i \in \rho(\overset{\circledast}{\gamma}_j)) = P(G(\vec{\gamma}_{j^+})|\{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_{j^+}))P(G(\vec{\gamma}_{j^-})|\{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^-}\} \subseteq \rho(\vec{\gamma}_{j^-})) \tag{4.18}$$

Here I just arbitrarily defined $G(\vec{\gamma}_{j^+})$ as the lineage of the gene tree that evolved $T(\vec{\tau}_{i^+})$ and $G(\vec{\gamma}_{j^-})$ as the lineage of the gene tree that evolved $T(\vec{\tau}_{i^-})$; their names are not relevant to the calculation. Both of these probabilities can be solved using (4.8).

The probability a gene lineage at the base of a taxon tree node will pass through into one of its descendant nodes and evolve into a given gene subtree can be broken down into the probability of the subtree evolving in the one taxon lineage and the probability that a gene lineage reconciled at the base of a taxon node would pass through into that branch of the taxon tree.

$$\begin{aligned} &P(G(\vec{\gamma}_j), \vec{\tau}_{i^+} \in \rho(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) \\ &\qquad = P(G(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_j)) P(\vec{\tau}_{i^+} \in \rho(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_i\} \subseteq \rho(\vec{\gamma}_j)) \end{aligned} \tag{4.19}$$

These two probabilities can be solved for using (4.8) and (4.7). Thus completing the circle and allowing us to pass on to another branch of the taxon tree.

Simplified Calculation

All of this reduces to a much simpler set of equations, as a number of probabilities cancel out. Here I will show how to calculate $\Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i)$ the probability of the gene tree above $\vec{\gamma}_j$ evolving, if we assume that there was one lineage at the beginning of $\vec{\tau}_i$, but we do not assume that lineage survives to the present. We can see that this value is closely related to the probability of the gene tree if that lineage is reconstructed, which we calculated in the previous two sections.

$$\begin{aligned} \Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i) =& P(G(\vec{\gamma}_j) | \mathring{N}_i = 1) \\ =& P(G(\vec{\gamma}_j) | \mathring{n}_i = 1) P(\mathring{n}_i = 1 | \mathring{N}_i = 1) \\ =& P(G(\vec{\gamma}_j) | \{\overset{\circledast}{\tau}_i, \vec{\tau}_{i^+}\} \subseteq \rho(\vec{\gamma}_j))(1 - E_0(\mathring{t}_i)) \end{aligned} \tag{4.20}$$

I will now lay out an efficient method to calculate Ψ based on the derivations from the previous two sections and assuming that the values of the birth-death parameters for the evolution of the gene tree are constant on each branch of the taxon tree, even if they vary between branches. Under this model we will call $\vec{\lambda}_i$ the birth rate on $\vec{\tau}_i$ and $\vec{\mu}_i$ the death rate on $\vec{\tau}_i$. We will calculate $\Psi(\overset{\star}{\gamma}, \vec{\tau})$ for every $\overset{\star}{\gamma} \in \overset{\star}{G}_0$, and for every $\vec{\tau} \in \vec{T}_0 \cap R(\overset{\star}{\gamma})$.

$$\Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i) = \frac{(1 - \vec{u}_i)(1 - \vec{a}_i \vec{u}_i)}{(1 - \vec{u}_i E_0(\overset{\star}{t}_i))^2} \sum_{m = \overset{\star}{n}_i^{\min}(\overset{\star}{\gamma}_j)}^{\overset{\times}{n}_0(\overset{\star}{\gamma}_j)} \Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i, m) \tag{4.21}$$

Where $\vec{a}_i = \vec{\mu}_i / \vec{\lambda}_i$ and $\vec{u}_i$ is the probability of a single reconstructed lineage at the beginning of $\vec{\tau}_i$ leaving more than one reconstructed lineage at the end of that branch, and can be calculated using (2.21) or (2.22).

Next we calculate $\Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i, m)$ for each value of m between $\overset{\star}{n}{}_i^{\min}(\overset{\star}{\gamma}_j)$ and $\overset{\star}{n}_0(\overset{\star}{\gamma}_j)$.

$$\Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i, m) = \begin{cases} 1 & \text{if } m = 1 \text{ and } \overset{\times}{\tau}_i \in R(\overset{\times}{\gamma}_j) \\ \Psi(\overset{\star}{\gamma}_j, \vec{\tau}_{i^+}) E(\mathring{t}_{i^-}) & \text{if } m = 1 \text{ and } \vec{\tau}_{i^+} \in R(\overset{\star}{\gamma}_j) \\ \Psi(\overset{\star}{\gamma}_j, \vec{\tau}_{i^-}) E(\mathring{t}_{i^+}) & \text{if } m = 1 \text{ and } \vec{\tau}_{i^-} \in R(\overset{\star}{\gamma}_j) \\ \Psi(\overset{\star}{\gamma}_{j^+}, \vec{\tau}_{i^+}) \Psi(\overset{\star}{\gamma}_{j^-}, \vec{\tau}_{i^-}) & \text{if } m = 1 \text{ and } \overset{\circledast}{\tau}_i \in R(\overset{\star}{\gamma}_j) \\ \psi(\overset{\circledast}{\gamma}_j, \vec{\tau}_i, m) & \text{if } m > 1 \end{cases} \tag{4.22}$$

And of course for every node of the gene tree we will calculate $\psi(\overset{\circledast}{\gamma}_j, \vec{\tau}_i, n)$ for every possible value of n.

$$\psi(\overset{\circledast}{\gamma}_j, \vec{\tau}_i, n) = \frac{k(\overset{\circledast}{\gamma}_j)}{n-1} \frac{\vec{u}_i}{1 - \vec{u}_i E_0(\overset{\star}{t}_i)} \sum_{m=\alpha(\overset{\star}{\gamma}_j, n)}^{\omega(\overset{\star}{\gamma}_j, n)} \Psi(\overset{\star}{\gamma}_{j^+}, \vec{\tau}_i, m) \Psi(\overset{\star}{\gamma}_{j^-}, \vec{\tau}_i, n-m) \tag{4.23}$$

Where $\alpha(\overset{\star}{\gamma}_j, n)$ and $\omega(\overset{\star}{\gamma}_j, n)$ are calculated as in section 4.2.4.

This entire calculation is made most efficiently by doing a down pass of every branch of the gene tree, $\vec{\gamma}_j \in \vec{G}_0$, such that we start with the terminal branches and finish with the roots and the calculations for $\vec{\gamma}_{j^+}$ and $\vec{\gamma}_{j^-}$ always precede $\vec{\gamma}_j$. For every $\vec{\gamma}_j$ we make a series of calculations for every branch of the taxon tree on which it could be reconstructed, $\vec{\tau}_i \in \vec{T}_0 \cap R(\vec{\gamma}_j)$, starting with the most recent branch and proceeding down to the root. For every combination of $\vec{\gamma}_j$ and $\vec{\tau}_i$ we calculate $\Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i, \overset{\star}{n}_i(\overset{\star}{\gamma}_j))$ for every possible value of $\overset{\star}{n}_i(\overset{\star}{\gamma}_j)$ and then $\Psi(\overset{\star}{\gamma}_j, \vec{\tau}_i)$. It will be possible to calculate $\psi(\overset{\star}{\gamma}_j, \vec{\tau}_i, \overset{\star}{n}_i(\overset{\star}{\gamma}_j))$ because we have already calculated every possible $\Psi(\overset{\star}{\gamma}_{j^+}, \vec{\tau}_i, \overset{\star}{n}_i(\overset{\star}{\gamma}_{j^+}))$ and $\Psi(\overset{\star}{\gamma}_{j^-}, \vec{\tau}_i, \overset{\star}{n}_i(\overset{\star}{\gamma}_{j^-}))$.

Probability of a Gene Tree at the Root of a Taxon Tree

So, I have shown how to calculate the probability of a gene tree once we have established that it was rooted at some point on a taxon tree. However, in reality we only have a gene tree and a taxon tree without any a priori information about where the gene tree actually began evolving on the taxon tree. Therefore we must choose some distribution that describes the relationship between the root of the gene tree and the root of the taxon tree.

The approach taken by Arvestad et al. (2003, 2004) was to assume that the node at the root of the gene tree was orthologous on the node at the root of the taxon tree. Thus each of the two basal clades of the gene tree start at the base of the two basal clades of the taxon tree. This approach has two major problems. The first is that it does not allow us to analyze gene trees for which the basal node could not be orthologous on the gene tree. These include gene trees in which the genes were found only in members of one of the two basal clades of taxa, and trees in which the minimum number of gene lineages at the the root node of the taxon tree is greater than one. Furthermore, this approach violates the general principle of the analysis in which we sum over all possible reconciliations and do not assume that just because a reconstructed gene node can be orthologous it is orthologous.

Therefore we should allow nodes of the gene tree to be paralogous on the root of the taxon tree. However, what the distribution of such nodes should be is not entirely clear. On all the other branches of the taxon tree we were able to use the branch length and the birth-death parameters to determine the probability of a single gene lineage producing a given number of lineages along a branch of the taxon tree. However, the root of the taxon tree is theoretically infinitely long and thus we can not determine when exactly the gene tree started evolving, and thus we can not calculate the probability of a given number of reconstructed lineages at the root.

One approach would be to gather a great deal of data about the position of the gene tree root. This could include apparently orthologous genes from out group taxa, and the sister group of paralogous genes from these same taxa. Genes from a number of nested out group taxa that appeared to be orthologous would allow us to greatly increase our confidence that any gene duplications in our gene tree occurred after our taxon tree diverged from its closest living relative, and thus we would be able to put a maximum age for the root of the gene tree. Similarly if our gene tree had a number of paralogous sister clades that were not present in any out group taxa, then that would also increase our confidence that any gene duplications in our gene tree occurred after our taxon tree diverged from its closest living relative. However, such data is rarely available, and so we must turn to some distribution.

Åkerborg et al. (2009) assumed that the prior position for the root node of the gene tree was uniformly distributed between the root node of the taxon tree and the infinite past. This approach is similar to the one that I used for the distribution of reconstructed lineages in subsection 2.3.3. However it is not practical for the model I use here, as I assume that the birth-death parameters differ between the branches of the taxon tree, and so it is not clear what parameters should be used on the root. The obvious approach is to assume that those parameters are drawn from the prior distribution, but as there is no corroborating data one would have to be very confident in their prior distribution. I am not.

Instead I used a flat prior for the distribution of number of reconstructed gene lineages in the basal node of the taxon tree. Thus the topology of the gene tree on the root of the taxon tree will affect the probability of the gene tree, but the number of gene lineages at the basal node will not, except in their influence on the subsequent evolution of the gene tree. In other words, if $\vec{\gamma}_r$ is the root of the gene tree and $\vec{\tau}_r$ is the root of the taxon tree, then $P(\dot{n}_r(\vec{\gamma}_r)) = C$, where C is some constant. The probability of the gene tree topology given the number of reconstructed lineages at the root of the taxon tree, $P(G_0|\dot{n}_r(\vec{\gamma}_r), \vec{\tau}_r \in \rho(\vec{\gamma}_r))$, can be calculated using (4.10). It would be trivial to analytically sum over all the different reconstructed numbers of gene lineages at the base of the taxon tree, but I treated it as a free parameter in order to better investigate the distribution of this value and to study its effects on the probability of the gene tree. It should be noted that although I placed a flat prior on the number of genes at the root of the taxon tree, this value will be biased towards its minimum by the rest of the likelihood calculation, because reconciliations in which $\rho(\overset{\circledast}{\gamma}) \ni \rho_{\mathrm{MRC}}(\overset{\circledast}{\gamma})$ for any $\overset{\circledast}{\gamma} \in \overset{\circledast}{G}_0$, will tend to have a higher likelihood than a reconciliation in

which $\rho(\overset{\circledast}{\gamma}) \neq \rho_{\mathrm{MRC}}(\overset{\circledast}{\gamma})$ (see section 4.2.1).

Reconstructing Gene Lineages from the Root of the Taxon Tree

I have shown how to calculate the probability of a gene tree from a single gene lineage at the root of a taxon tree which we know will survive to the present. Thus, we assume that at least one descendant of this gene lineage will be found in at least one tip of the taxon tree. In this sense, the gene lineage is reconstructed and this is certainly a proper assumption as this gene tree has been drawn from the set of all gene trees that have members in the taxa we are studying. However, this assumption may not be sufficiently restrictive to capture the distribution of the actual set of gene trees from which our gene tree may be drawn. In particular two more restrictive assumptions may often be appropriate when analyzing a gene tree: we may assume that at least one descendant of all the reconstructed gene lineages at the base of the taxon tree survived in at least one member of each of the basal taxon clades; or even more restrictively, we may assume that at least one descendant of all the reconstructed gene lineages at the base of the taxon tree survived in every taxon we studied. Here I will show how to calculate the probability of these assumptions and how to use them to correct the probability of the gene tree.

Let $\overset{\times}{T}_i$ be some subset of all the tips in T_0, so that $\overset{\times}{T}_i \subseteq \overset{\times}{T}_0$. Furthermore let $\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma)$ be the number of tips in $G(\gamma)$ found in $\overset{\times}{T}_i$, so that $\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma_j) = \left|\{\overset{\times}{\gamma} : \overset{\times}{\gamma} \in \overset{\times}{G}(\gamma_j), \rho_{\mathrm{MRC}}(\overset{\times}{\gamma}) \in \overset{\times}{T}_i\}\right|$. I will also define $E_{\overset{\times}{T}_i}(t)$ as the probability that a single gene lineage alive at time t on the taxon tree has no descendants in any of the tips in $\overset{\times}{T}_i$, so that $E_{\overset{\times}{T}_i}(t) = P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma_j) = 0 | N_t(\gamma_j) = 1)$, and $E_{\overset{\times}{T}_0}(t) = E_0(t)$. In that case it is easy to solve for $E_{\overset{\times}{T}_i}(t)$; prune all the tips not found in $\overset{\times}{T}_i$ and calculate $E_0(t)$ for this reduced tree, as you would for the whole tree.

This calculation creates a situation that is unfamiliar to us so far. A gene lineage can be reconstructed, meaning that it survives into some members of $\overset{\times}{T}_0$ even if it has been lost in every member of $\overset{\times}{T}_i$. Therefore we can calculate the probability of a reconstructed gene lineage being lost in some set of tips. We can calculate the probability that a gene lineage alive at the root of a taxon tree, which has survived to the present, has no descendants in $\overset{\times}{T}_i$ from probabilities that we have already established as follows:

$$\begin{aligned} P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma) = 0 | \overset{\star}{n}_r(\gamma) = 1) &= 1 - P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma) > 0 | \overset{\times}{n}_0(\gamma) > 0, \overset{\star}{n}_r(\gamma) = 1) \\ &= 1 - \frac{P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma) > 0 | \overset{\star}{N}_r(\gamma) = 1)}{P(\overset{\times}{n}_0(\gamma) > 0 | \overset{\star}{N}_r(\gamma) = 1)} \\ &= \frac{E_{\overset{\times}{T}_i}(\overset{\star}{t}_r) - E_0(\overset{\star}{t}_r)}{1 - E_0(\overset{\star}{t}_r)} \end{aligned} \tag{4.24}$$

I have already asserted that there can be more than one reconstructed gene lineage in the root. Therefore in order for there to be no genes in some subset of the tips of the taxon tree,

every gene lineage in the root must leave no descendants in those tips. It is easy to calculate the probability that $\overset{\star}{n}_r(\gamma)$ reconstructed lineages alive at the end of the root $\vec{\tau}_r$ of the taxon tree will leave no descendants in $\overset{\times}{T}_i$, because every one of those initial gene lineages must do the same thing, leave some descendants in $\overset{\times}{T}_0$, but none in $\overset{\times}{T}_i$.

$$P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma)=0|\overset{\star}{n}_r(\gamma)=m) = [P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma)=0|\overset{\star}{n}_r(\gamma)=1)]^m \tag{4.25}$$

We now have enough information to calculate the probability that $\overset{\star}{N}_r$ lineages alive at the basal node of the taxon tree all survive to the present and that at least one survives in each of the basal clades of the taxon tree, $T(\vec{\tau}_{r^+})$ and $T(\vec{\tau}_{r^-})$. It is the probability that all those lineages survive to the present and the probability that if those lineages survive to the present they survive in both basal taxon clades.

$$\begin{aligned}
&P(\overset{\times}{n}_{\overset{\times}{T}(\vec{\tau}_{r^+})}(\gamma)>0, \overset{\times}{n}_{\overset{\times}{T}(\vec{\tau}_{r^+})}(\gamma)>0, \overset{\star}{n}_r(\gamma)=\overset{\star}{N}_r(\gamma)|\overset{\star}{N}_r(\gamma)) \\
&\quad = [1 - \sum_{\vec{\tau}_i \in \{\vec{\tau}_{r^+}, \vec{\tau}_{r^-}\}} P(\overset{\times}{n}_{\overset{\times}{T}(\vec{\tau}_i)}(\gamma)=0|\overset{\star}{n}_r(\gamma))]P(\overset{\star}{n}_r(\gamma)=\overset{\star}{N}_r(\gamma)|\overset{\star}{N}_r(\gamma)) \\
&\quad = (1 - E_0(\overset{\star}{t}_r))^{\overset{\star}{n}_r(\gamma)} - \sum_{\vec{\tau}_i \in \{\vec{\tau}_{r^+}, \vec{\tau}_{r^-}\}} (E_{\overset{\times}{T}(\vec{\tau}_i)}(\overset{\star}{t}_r) - E_0(\overset{\star}{\tau}_r))^{\overset{\star}{n}_r(\gamma)}
\end{aligned} \tag{4.26}$$

We do not have to consider the probability that the lineages do not survive in either clade given that they survive to the present, because that is impossible.

We can also calculate the the probability that all the gene lineages alive at the taxon root node survived to the present and that at least one lineage survived into every tip of the taxon tree as the probability that all those lineages survived and the probability that there were not any taxon tips without at least one gene lineage given that all those gene lineages at the root survived to the present. We can calculate the probability that any set of tips of the taxon tree had no genes, because I have already shown how to calculate that no genes survived into any given set of taxon tips.

$$\begin{aligned}
&P(\bigcap_{\overset{\times}{\tau}_i \in \overset{\times}{T}_0} \overset{\times}{n}_i(\gamma)>0, \overset{\star}{n}_r(\gamma)=\overset{\star}{N}_r(\gamma)|\overset{\star}{N}_r(\gamma)) \\
&\qquad = [1 - P(\bigcup_{\overset{\times}{\tau}_i \in \overset{\times}{T}_0} \overset{\times}{n}_i(\gamma)=0|\overset{\star}{n}_r(\gamma))]P(\overset{\star}{n}_r(\gamma)=\overset{\star}{N}_r(\gamma)|\overset{\star}{N}_r(\gamma)) \\
&\qquad = [1 + \sum_{\overset{\times}{T}_i \in \wp_1(\overset{\times}{T}_0)} (-1)^{\left|\overset{\times}{T}_i\right|} P(\overset{\times}{n}_{\overset{\times}{T}_i}(\gamma)=0|\overset{\star}{n}_r(\gamma))](1 - E_0(\overset{\star}{t}_r))^{\overset{\star}{n}_r(\gamma)} \\
&\qquad = \sum_{\overset{\times}{T}_i \in \wp(\overset{\times}{T}_0)} (-1)^{\left|\overset{\times}{T}_i\right|} (E_{\overset{\times}{T}_i}(\overset{\star}{t}_r) - E_0(\overset{\star}{t}_r))^{\overset{\star}{n}_r(\gamma)}
\end{aligned} \tag{4.27}$$

It is important to note that the power set of $\overset{\times}{T}_0$ includes both the empty set and a set of all the tips in $\overset{\times}{T}_0$. The probability of a gene lineage leaving no descendants in the empty set is

always one, $E_{\varnothing}(t) = 1$, as you can not find any genes if you do not look in any taxa. On the other hand the probability that no genes survive in $\overset{\star}{T}_0$ is $E_0(t)$, so that the last member in this sum will be zero and we could exclude $\overset{\star}{T}_0$ from the sum entirely. However, the equation looks better if we leave it there.

I have now shown how to calculate the probability for three different assumptions about how gene lineages alive at the base of a taxon tree survive to the present. Under the first assumption, all the gene lineages survive to the present; under the second assumption they all survive to the present and at least one survives in each of the basal lineages of the taxon tree; and under the third assumption all the lineages survive and at least one survives in every sampled taxon. We can calculate the probability of a gene tree conditioned on any of these assumptions as follows:

$$P(G_0|\overset{\star}{n}_r, \Theta) = \frac{P(G_0, \overset{\star}{n}_r|\overset{\star}{N}_r = \overset{\star}{n}_r)}{P(\Theta, \overset{\star}{n}_r|\overset{\star}{N}_r = \overset{\star}{n}_r)}$$

where Θ is the assumption about the survival of the gene lineages into the present. It goes without saying that we should only use assumptions that hold true for our gene tree.

4.3 Bayesian Inference

The goal of this paper is to determine if the gene tree evolves at a different rate on different branches of the taxon tree, and if so, how do the rates differ between branches. Our model has three different types of parameters, gene duplication rates, gene loss rates, and the number of reconstructed lineages at the base of the tree. We want to compare models in which the assignment of rates to the various branches of the taxon tree differs. I will compare the rate assignments using Bayesian inference in which I will estimate a likelihood for each assignment that has been marginalized over all the possible values for every parameter of that model (see Ellison 2004; Huelsenbeck et al. 2004). I will then compare the ratios of these marginalized likelihoods, or Bayes factors, in order to determine the relative fit of each rate assignment to the data (see Kass and Raftery 1995). As a byproduct, I will also estimate the posterior distributions of the rate parameters on the different branches of the taxon tree by marginalizing over all the other parameter values and rate assignments. The number of reconstructed lineages at the base of the tree will be strictly a nuisance parameter, although its posterior distribution will tell us something about how the model fits the data.

I will use a reversible-jump Markov Chain Monte Carlo method implementation of the Metropolis-Hastings algorithm to estimate the posterior distribution of these values (Green 1995). Under this method each step in the chain consists of a rate assignment and a set of parameter values. A new step is generated through a random modification of the previous step, and a proposal ratio is calculated. A uniform random number is generated between zero and one, and the new proposal is accepted if the proposal ratio is greater than that

random number. If the proposal ratio is less, then the proposal is rejected and we return to the model and parameter values from the previous step. The proposal ratio is calculated, such that at stationarity the rate assignments and their parameters will be sampled from their posterior probability distribution.

Let A be any assignment of birth-death rates to the branches of the taxon tree, so that $A(\lambda)$ is a function that returns the set of taxon branches which have gene duplication rate λ under that model, and $A(\mu)$ returns the set of taxon branches with μ for a gene loss rate. We will call the original assignment before a proposal A^0, and the original set of parameter values z^0. We generate a second proposal assignment A' with a set of parameter values z'. z^0 is transformed to z' using a set of random values y^0. In contrast z' can be transformed into z^0 using the set of values y'. In that case our MCMC will sample from the posterior probability of assignments when our proposal ratio is:

$$\frac{P(G_0|A',z')P(A')P(z')f(A^0,z^0|A',z')}{P(G_0|A^0,z^0)P(A^0)P(z^0)f(A',z'|A^0,z^0)}$$

where $P(G_0|A^0,z^0)$ is the probability of the gene tree given the parameter values described by A^0 and z^0, which I described how to calculate in section 4.2; $P(A^0)$ and $P(z^0)$ are the prior probabilities of A^0 and z^0; and $f(A^0,z^0|A',z')$ is the probability density for proposing A^0 and z^0 given that you start with A' and z' (see Waagepetersen and Sorensen 2001). We can calculate the ratio between the proposal densities as follows:

$$\frac{f(A^0,z^0|A',z')}{f(A',z'|A^0,z^0)} = \frac{P(A^0|A')q(y'|A',z',A^0)}{P(A'|A^0)q(y^0|A^0,z^0,A')}\left|\frac{\partial(z',y')}{\partial(z^0,y^0)}\right|$$

Where $P(A'|A^0)$ is the probability of generating model A' if you start in A^0; $q(y^0|A^0,z^0,A')$ is the density of y^0 given that you start in model A^0 with parameters z^0 and you propose model A'; and $|\partial(z',y')/\partial(z^0,y^0)|$ is the absolute value of the Jacobian of a vector of all the parameters in z' and all the random values in y' with respect to a vector of z^0 and y^0. These vectors must be of the same length, even though z' and z^0 may be of different sizes.

In the remainder of this section I will describe how new proposals were generated for each of the parameters and the models, and I will show how the proposal ratio was calculated. I have implemented these calculations in the Tree Reconciliation Using Likelihood (TRUL) software package, using C++.

4.3.1 Root Proposals

The simplest proposal and hence the simplest acceptance ratio is for the number of reconstructed lineages at the root of the tree. At the beginning of each step we decide to vary $\dot{n}_r$ with probability c_l. We will call the original number of lineages at the root $\dot{n}_r^0$ and the proposed number of lineages $\dot{n}_r'$. There is some maximum number of lineages by which the number of lineages at the root can change, $\Delta\dot{n}_r$, which is set by the user. Of

course $\dot{n}_r$ could not change by so many lineages that it fell outside the range of possible values. Therefore if $\Delta_D \dot{n}_r^0$ is the maximum amount by which the number of lineages before we generate a new proposal could decrease and $\Delta_I \dot{n}_r^0$ is the maximum amount by which they could increase, $\Delta_D \dot{n}_r^0 = \min(\Delta \dot{n}_r, \dot{n}_r^0 - \dot{n}_r^{\min})$ and $\Delta_I \dot{n}_r^0 = \min(\Delta \dot{n}_r, \check{n}_0 - \dot{n}_r^0)$.

In order to generate a proposal value for the number of lineages at the root we choose an integer at random between $\dot{n}_r^0 - \Delta_D \dot{n}_r^0$ and $\dot{n}_r^0 + \Delta_I \dot{n}_r^0$ excluding $\dot{n}_r^0$, and use that integer as $\dot{n}_r'$. Therefore, there are $\Delta_D \dot{n}_r^0 + \Delta_I \dot{n}_r^0$ possible values for $\dot{n}_r'$, and each of them is equally likely to be chosen. In this case $f(M', z'|M^0, z^0)$ is actually a probability mass and equal to $c_l/(\Delta_D \dot{n}_r^0 + \Delta_I \dot{n}_r^0)$. No other parameters are changed, the assignments of rates to branches remain the same, and the prior is flat for the number of lineages at the root (see section 4.2.4), so the proposal ratio for a change from $\dot{n}_r^0$ to $\dot{n}_r'$ is:

$$\frac{P(G_0|\dot{n}_r')(\Delta_D \dot{n}_r' + \Delta_I \dot{n}_r')}{P(G_0|\dot{n}_r^0)(\Delta_D \dot{n}_r^0 + \Delta_I \dot{n}_r^0)}$$

4.3.2 Rate Proposals

Proposals that just change a single rate work the same no matter which rate we choose to modify. There can be as many different λs as there are branches on the taxon tree, or there can be as few as one λ that is the same for every branch. There can be just as many or few μs. We will call Λ^0 the set of all the λs assigned to branches of the taxon tree under assignment A^0, and M^0 the set of all the μs, so that $\Lambda^0 = \{\lambda : A^0(\lambda) \neq \varnothing\}$ and $M^0 = \{\mu : A^0(\mu) \neq \varnothing\}$. We will choose to modify a rate with probability c_r and we choose one rate to modify from $\Lambda^0 \cup M^0$ at random so that the probability of any rate being picked is $1/(|\Lambda^0| + |M^0|)$. The assignment of rates to branches will not change between the original distribution and the proposal distribution, so both of these values will remain the same and will cancel out in the proposal ratio.

Once we have decided to modify a specific rate, all rates are modified in the same way. Here I will describe how a λ is modified, but the same procedure applies to any μ. Let λ^0 be the original λ, and λ' be the λ for the proposal distribution. There is some number chosen by the user, which I will call Ξ. For each proposal we generate a random number, ξ^0, from the uniform distribution $(-\Xi/2, \Xi/2)$, and use this number to transform λ^0, so that $\lambda' = \lambda^0 e^{\xi^0}$. Thus the distribution of ξ^0 given λ^0 is $\xi^0/\Xi + 0.5$, and we can calculate the density of ξ^0 by taking the derivative of this distribution, so that:

$$q(\xi^0|\lambda^0) = \frac{1}{\Xi}$$

To calculate the Jacobian we must be able to describe λ' and ξ' in terms of λ^0 and ξ^0, where ξ' is the value of ξ necessary to transform λ' into λ^0. We already know how to calculate λ' in terms of λ^0 and ξ^0. We must choose a ξ', such that $\lambda^0 = \lambda' e^{\xi'}$. We can see that this will be accomplished, when $\xi' = -\xi^0$ by substituting our equation for λ' into our last equation.

Thus we can calculate the Jacobian as so:

$$\frac{\partial(\lambda', \xi')}{\partial(\lambda^0, \xi^0)} = \begin{vmatrix} \frac{\partial\lambda'}{\partial\lambda^0} & \frac{\partial\lambda'}{\partial\xi^0} \\ \frac{\partial\xi'}{\partial\lambda^0} & \frac{\partial\xi'}{\partial\xi^0} \end{vmatrix} = \begin{vmatrix} e^{\xi^0} & \lambda' \\ 0 & -1 \end{vmatrix} = -e^{\xi^0}$$

The assignment and the other parameters do not change, so all that remains for us to calculate the proposal ratio is the prior distribution of λ. I assumed an exponential prior with expectation $\hat{\lambda}$ for all rates, so that the prior probability density of any rate λ is:

$$P(\lambda) = \frac{\exp(-\lambda/\hat{\lambda})}{\hat{\lambda}}$$

With this we can now calculate the proposal ratio for a proposal in which λ^0 is replaced with λ':

$$\frac{P(G_0|\lambda')}{P(G_0|\lambda^0)} \exp\left(\xi^0 + \frac{\lambda^0 - \lambda'}{\hat{\lambda}}\right)$$

4.3.3 Assignment Proposals

The last type of proposal is one in which the assignment of rates to the branches of the taxon tree changes. There are two types of possible changes: two sets of branches with a rate assigned to each set can be combined, so that together they only have one rate; or one set of branches with one rate can be split into two. I will call call an assignment in which there are x rates assigned to the various branches of the tree A_x, so that $x = |\Lambda| + |M|$. Here we will consider a case in which an assignment with x rates, A^0_x is transformed into an assignment with $x+1$ rates, A'_{x+1}. I will call the rate that is modified by the proposal λ^0 and the two new rates created by the proposal λ'_1 and λ'_2, $A^0_x(\lambda^0) = A'_{x+1}(\lambda'_1) \cup A'_{x+1}(\lambda'_2)$ and $A'_{x+1}(\lambda^0) = A^0_x(\lambda'_1) = A^0_x(\lambda'_1) = \varnothing$, where $\lambda'_1 < \lambda'_2$. All rates other than these three are assigned to the same branches under both assignments, $A^0(\lambda) = A'(\lambda)$ for all $\lambda \notin \{\lambda^0, \lambda'_1, \lambda'_2\}$.

There is a critical value Ξ_A which is set by the users and is used to transform the rates. Once we have chosen a rate, λ^0, to reassign and distributed its branches to the rates λ'_1 and λ'_2, we must give those rates values. A random number, ξ^0_A is generated from the uniform distribution $(0, \Xi_A)$, so that the probability density for generating ξ^0_A, when you are switching from model A^0_x to A'_{x+1} is:

$$q(\xi^0_A | A^0_x, \lambda^0, A'_{x+1}) = \frac{1}{\Xi_A}$$

ξ^0_A is then used to generate the values for the two new rates, such that $\lambda'_1 = e^{-\xi_A}\lambda^0$ and $\lambda'_2 = e^{\xi_A}\lambda^0$. On the other hand, when we make a proposal that replaces model A'_{x+1} with A^0_x, we take the harmonic mean of λ'_1 and λ'_2 as our new value for λ^0. Thus we do not need to generate a random value for the reverse proposal and the total number of random numbers and transformed rates in each model is two. This allows us to calculate the Jacobian like so:

$$\frac{\partial(\lambda'_1, \lambda'_2)}{\partial(\lambda^0, \xi^0_A)} = \begin{vmatrix} \frac{\partial\lambda'_1}{\partial\lambda^0} & \frac{\partial\lambda'_1}{\partial\xi^0_A} \\ \frac{\partial\lambda'_2}{\partial\lambda^0} & \frac{\partial\lambda'_2}{\partial\xi^0_A} \end{vmatrix} = \begin{vmatrix} e^{-\xi_A} & -e^{-\xi_A}\lambda^0 \\ e^{\xi_A} & e^{\xi_A}\lambda^0 \end{vmatrix} = 2\lambda^0$$

In order to determine which rates to reassign we must first identify which rates can in fact have their taxon branches either divided or combined with those of another rate. Only rates which are assigned to more than one branch can have the taxon branches to which they are assigned divided into two sets. Let L^0 and L' be the number of rates that are assigned to more than one branch under assignments A^0 and A' respectively, so that $L^0 = \{\lambda | |A^0(\lambda)| > 1\} \cup \{\mu | |A^0(\mu)| > 1\}$. Any rate found in L^0 could have its taxon branches subdivided. On the other hand, we can only combine the taxon branches assigned to two rates in assignment A'_{x+1} if the absolute value of the log of the ratio between those two rates is less than $2\Xi_A$, because if it were any larger, then the assignment generated by combining their taxa could not generate A'_{x+1} in a single step and so the probability density for the reverse proposal would be zero. I will define D' and D^0 as the set of all the ordered pairs of rates that could have their assigned taxon branches combined under assignments A' and A^0 respectively, so that

$$D' = \{(\lambda'_1, \lambda'_2) | 0 < \lambda'_2/\lambda'_1 < \Xi_A, \{\lambda'_1, \lambda'_2\} \subseteq \Lambda\} \cup \{(\mu'_1, \mu'_2) | 0 < \mu'_2/\mu'_1 < \Xi_A, \{\mu'_1, \mu'_2\} \subseteq M\}$$

A proposal modifies the rate assignments with probability $c_A = 1 - c_1 - c_r$. If there are no rates assigned to multiple branches, $L^0 = \varnothing$, then it combines rates; if there are no rates which can have their branches combined, $D^0 = \varnothing$, then it divides the branches associated with a rate; and if neither L^0 or D^0 is empty then we divide branches or combine branches with a 50% probability. If we decide to divide the branches associated with a rate, then we choose a rate, λ^0, at random from L^0. The branches associated with λ^0 are assigned at random to either λ'_1 or λ'_2, such that at least one branch is assigned to each rate. Therefore there are $2^{|A^0_x(\lambda^0)|} - 2$ possible ways that the branches could be distributed, and we can calculate the probability of generating assignment A'_{x+1} from A^0_x as:

$$P(A'_{x+1} | A^0_x) = \frac{c_A \phi(D^0)}{|L^0|(2^{|A^0_x(\lambda^0)|} - 2)}$$

Where

$$\phi(X) = \begin{cases} 1 & \text{if } X = \varnothing \\ 0.5 & \text{if } X \neq \varnothing \end{cases}$$

for any set X. On the other hand, if the branches associated with two rates are to be combined, then a pair of rates is chosen at random from D' and their branches are combined, so that the probability of generating model A^0_x from A'_{x+1} is:

$$P(A^0_x | A'_{x+1}) = \frac{c_A \phi(L')}{|D'|}$$

The priors for all the models are the same and the prior for the rates are exponential, as discussed in the previous section, so we now have enough information to calculate the

proposal ratio for a proposal that changed A_x^0 into A'_{x+1}.

$$\frac{P(G_0|A'_{x+1},\lambda'_1,\lambda'_2)}{P(G_0|A_x^0,\lambda^0)}\frac{\lambda^0\Xi_A}{\hat{\lambda}}\frac{|L^0|(2^{|A_x^0(\lambda^0)|+1}-4)}{|D'|}\frac{\phi(L')}{\phi(D^0)}\exp\left(\frac{\lambda^0-\lambda'_1-\lambda'_2}{\hat{\lambda}}\right)$$

The inverse of this value is the proposal ratio for the reverse proposal in which A'_{x+1} is transformed into A_x^0.

4.4 Comparison to Gene Count Model

Comparing the gene tree to the taxon tree should provide us with information about the evolutionary processes that produced the gene tree. However, much of that information is also contained in simple counts of genes in the extant taxa. Hahn et al. (2005) used gene counts alone to deduce changes in the rate of gene gain and loss on branches of the taxon tree. Although, computationally the analysis of a full gene tree is not much more cumbersome than an analysis of gene counts alone, the additional burden of deducing the phylogeny of the genes makes a full gene tree analysis much more time consuming. Yet this extra effort may be worthwhile as the full gene tree may allow for both greater accuracy and precision in the inference of the actual process that produced the gene tree. In particular the full gene tree would allow us to better deduce trends in the evolution of gene numbers and to distinguish changes in the gene loss rate from changes in the gene gain rate on branches of the phylogeny.

In order to see how well the gene tree model performs in the estimation of birth-death parameters and the detection of changes in those parameters, I simulated a number of gene trees on a simple taxon tree using a variety of parameter values. I then analyzed those trees using both the full gene tree model and a model that relies only on the gene counts in each taxon. I used these results to compare models and to investigate the effectiveness of each model on its own. Under the gene count model, the number of gene lineages at the base of the tree is unknown, and treated as a nuisance parameter, as it is in the gene tree model. However, the gene tree itself provides information about the number of gene lineages at the root of the taxon tree. Thus the gene count model would be incapable of detecting an increase in the number of chromosomes that occurred on every branch of the tree, while the gene tree model could infer a trend. Therefore, I analyzed the gene count data in two ways: first with a flat prior on the number of genes at the base of the tree and second assuming that the number of genes at the root was the number used in the simulation. Even though one could not know the actual number of genes at the base of the taxon tree without constructing a gene tree, using both methods allowed me to separate the effects of gene number trend detection from those caused by other differences between the models.

4.4.1 Gene Count Model

The model I used for the analysis of gene counts is essentially the same as the one I used in chapter 3 for the analysis of chromosome numbers with only a few modifications. I assumed that δ was zero, as we are not studying whole genome duplications, and I allowed the birth-death parameters to vary between branches.

I also changed the way that the calculation dealt with the fact that these lineages are only from gene families that survived to the present. In chapter 3 I accounted for the fact that every taxon had to have one chromosome by dividing the transition probability on each branch by the probability that at least one chromosome at the beginning of that branch survived to the end of that branch. Here I did not make that assumption, and I used $p_i(\overset{\star}{N}_i|\overset{\circ}{N}_i) = P(\overset{\star}{N}_i|\overset{\circ}{N}_i, \vec{\lambda}_i, \vec{\mu}_i)$ (3.2) to calculate the transition probability instead of $p_i(\overset{\star}{N}_i|\overset{\circ}{N}_i) = P(\overset{\star}{N}_i|\overset{\circ}{N}_i, \overset{\star}{N}_i > 0, \vec{\lambda}_i, \vec{\mu}_i)$. To account for the fact that we are only investigating gene trees for which we have found members in the studied taxa, I calculated similar probabilities to those calculated in section 4.2.4 except I did not assume that all the lineages at the root are reconstructed, as that is not an assumption of the gene count model.

The probability that at least one gene survives to the present is simply the probability that the genes present at the root node are not all lost.

$$P(\overset{\times}{n}_0 > 1|\overset{\star}{N}_r) = 1-(E_0(\overset{\star}{t}_r))^{\overset{\star}{N}_r} \tag{4.28}$$

The probability that at least one gene survives in each of the basal taxon clades is the probability that the genes present at the root node are not all lost in both those clades.

$$P(\overset{\times}{n}_{\overset{\times}{T}(\vec{\tau}_{r+})} > 0, \overset{\times}{n}_{\overset{\times}{T}(\vec{\tau}_{r-})} > 0|\overset{\star}{N}_r) = (1-(E_{\overset{\times}{T}(\vec{\tau}_{r+})}(\overset{\star}{t}_r))^{\overset{\star}{N}_r})(1-(E_{\overset{\times}{T}(\vec{\tau}_{r-})}(\overset{\star}{t}_r))^{\overset{\star}{N}_r}) \tag{4.29}$$

Finally the probability that one gene survives in all the taxa is the probability that there were not any taxon tips that the genes did not survive into.

$$P(\bigcap_{\overset{\times}{\tau}_i \in \overset{\times}{T}_0} \overset{\times}{n}_{\overset{\times}{\tau}_i} > 0|\overset{\star}{N}_r) = \sum_{\overset{\times}{T}_i \in \wp(\overset{\times}{T}_0)} (-1)^{|\overset{\times}{T}_i|}(E_{\overset{\times}{T}_i}(\overset{\star}{\tau}_r))^{\overset{\star}{N}_r} \tag{4.30}$$

Any of these assumptions can be applied to the probability of a set of gene counts by dividing the raw probability by the appropriate equation.

$$P(\rho(\overset{\times}{G}_0)|\overset{\star}{N}_r, \Theta) = \frac{P(\rho(\overset{\times}{G}_0)|\overset{\star}{N}_r)}{P(\Theta|\overset{\star}{N}_r)}$$

where Θ is an assumption about the genes surviving to the present.

In order to make the results of the gene count model directly comparable to those of the gene tree model, I analyzed the gene tree model using a reversible-jump MCMC. The gene count model has essentially the same parameters as the gene tree model, so I used

the same priors and the same proposal distributions. However, there is one fundamental difference between these two models; the lineages at the base of the gene tree model are all reconstructed, while the lineages at the base of the gene count model could be lost before the present. I gave a flat prior to both, which could potentially bias the analysis. However, we see that if the ratio between the priors for any two non-reconstructed gene counts is one, $P(\dot{N}_r = a)/P(\dot{N}_r = b) = 1$, that the same holds true for two different numbers of reconstructed lineages.

$$\frac{P(\dot{n}_r = a)}{P(\dot{n}_r = b)} = \frac{\sum_{i=a}^{\infty} P(\dot{n}_r = a|\dot{N}_r = i)P(\dot{N}_r = i)}{\sum_{j=b}^{\infty} P(\dot{n}_r = b|\dot{N}_r = j)P(\dot{N}_r = j)} = \frac{\sum_{i=a}^{\infty} P(\dot{n}_r = a|\dot{N}_r = i)}{\sum_{j=b}^{\infty} P(\dot{n}_r = b|\dot{N}_r = j)} = 1$$

Therefore the two priors are in essence the same and should not affect the posterior distribution.

This model has much in common with the one used by Hahn et al. (2005). In particular the likelihoods calculated for a set of birth-death parameters would be the same with the following exceptions: they assumed that $\lambda = \mu$; they assumed that those rates were the same on every branch of the tree; and they did not account for the fact that the gene families were sampled from the set of all gene families that have not been lost. Furthermore they identified branches with large differences in birth-death rates in a different way. They estimated a single value for λ for all gene families, and used this value to calculate the likelihood for each gene families given the distribution of gene counts on the tips of the tree. For those gene families which were found to have significantly low likelihoods for their size, they individually varied the value of λ for each branch in the taxon tree, and determined which branches lead to a significant improvement in the likelihood. These methods obviously differ substantially from my own despite their similarities. I made the changes that I did, so that the results could be compared directly to the gene tree model. However, as a consequence, I can not be certain that the method of Hahn et al. (2005) would not perform better than the one I used here.

4.4.2 Simulations

In order to investigate how these models analyze a gene family, I wanted to study a simple case, so that it would be easy to distinguish what effect the different elements of the process had on the analysis. So, I simulated a large number of gene trees under different sets of parameter values, on the simplest possible taxon tree, and analyzed those simulations using both the gene tree model and the gene count model. The simplest tree possible in which there is a difference between birth-death rates on the different branches of the tree is a tree with two tips and two branches connected at a root node. I arbitrarily chose a branch length of 1.0 for both branches.

Simulations started with some number of genes at the root node. Every gene lineage split into two lineages automatically at the root node, one lineage in each branch. Each lineage was then simulated independently, such that a random number h_1 was generated on the

uniform distribution (0,1). A time t was then calculated as $t = t_0 + \log(h_1)/(\vec{\lambda} + \vec{\mu})$, where t_0 was the time at which the lineage diverged from its sister lineage. If t was greater than one, then the lineage ended in the tip of the taxon tree and became a tip of the reconstructed gene tree. If t was less than one, then a second random number h_2 was generated on the uniform distribution (0,1). If h_2 was less than $\vec{\mu}/(\vec{\lambda}+\vec{\mu})$, then the lineage was lost; if it was larger, then the lineage split into two, and the process was repeated for each of its daughter lineages starting at time t. Trees in which at least one gene lineage did not survive in each branch of the taxon tree were discarded. Gene lineages at the root of the tree that survived to the present were randomly combined in order to complete the base of the gene tree.

I simulated trees under a number of different birth-death parameters. I varied the number of genes at the roots at the start of the simulation and the expected number of genes in one terminal of the taxon tree, while I held the expectation in the other tip constant at twenty genes. I will call the branch with the the expectation of twenty genes at the end $\vec{\tau}_1$ and the other branch $\vec{\tau}_2$. I held either $\vec{\lambda}$ or $\vec{\mu}$ as the same on both branches, and held the death rate on $\vec{\tau}_1$, $\vec{\mu}_1$, as either zero or 0.2, so that there were four possible rate assignments for every set of expectations: $S_{\lambda 0}$, in which $\vec{\lambda}_1 > \vec{\lambda}_2$ and $\vec{\mu}_1 = \vec{\mu}_2 = 0$; $S_{\mu 0}$, in which $\vec{\lambda}_1 = \vec{\lambda}_2$, $\vec{\mu}_1 = 0$ and $\vec{\mu}_2 > 0$; $S_{\lambda 2}$, in which $\vec{\lambda}_1 > \vec{\lambda}_2$ and $\vec{\mu}_1 = \vec{\mu}_2 = 0.2$; and $S_{\mu 2}$, in which $\vec{\lambda}_1 = \vec{\lambda}_2$, $\vec{\mu}_1 = 0.2$ and $\vec{\mu}_2 > 0.2$. The expected number of genes in a tip of the taxon tree can be calculated as:

$$\hat{\hat{n}}_{\vec{\tau}}^{\times} = \frac{\exp(r)}{1 - (E_{\vec{\tau}}^{\times}(\overset{\star}{t}_r))^{\overset{\star}{N}_r}}$$

I used the nlminb function from the R statistical programming language, first to calculate the appropriate value for $\vec{\lambda}_1$ to achieve an expectation of twenty genes, given the number of genes at the root and the appropriate value for $\vec{\mu}_1$. I then used the same method to calculate the free parameter for $\vec{\tau}_2$. Table 4.1 shows all the combinations of genes at the root, expected number of genes at the tip and the parameter values used to achieve those expectations. It was impossible to generate an expectation lower than the number of genes at the root, if the value of $\vec{\mu}_2$ was fixed too low, and so these simulations were excluded.

I simulated 100 trees under each set of parameter values.

4.4.3 Bayesian Analysis

Likelihood Models

I analyzed each simulated tree using the methodology described in section 4.3 and subsection 4.4.1 under three different models of gene evolution: the gene tree model, the gene count model, and the gene count model with the number of genes at the root fixed at the actual number used for the simulation. For all three models I assumed that at least one gene survived in each tip of the taxon tree, as that was a requirement of the simulation procedure. I assumed that the prior for all the rates, $\hat{\lambda}$, was 0.5 events/lineage/branch. Each analysis was started at a set of parameter values drawn randomly from the priors.

Table 4.1: The parameters used for the gene tree simulations.

$\overset{\star}{N}_r$	$\hat{\overset{\star}{n}}_1$	$\hat{\overset{\star}{n}}_2$	Assignment	$\vec{\lambda}_1$	$\vec{\mu}_1$	$\vec{\lambda}_2$	$\vec{\mu}_2$
5	20	10	$S_{\lambda 0}$	1.3863	0.0000	0.6931	0.0000
5	20	10	$S_{\mu 0}$	1.3863	0.0000	1.3863	0.6974
5	20	10	$S_{\lambda 2}$	1.5863	0.2000	0.8931	0.2000
5	20	10	$S_{\mu 2}$	1.5863	0.2000	1.5863	0.9029
5	20	15	$S_{\lambda 0}$	1.3863	0.0000	1.0986	0.0000
5	20	15	$S_{\mu 0}$	1.3863	0.0000	1.3863	0.2878
5	20	15	$S_{\lambda 2}$	1.5863	0.2000	1.2986	0.2000
5	20	15	$S_{\mu 2}$	1.5863	0.2000	1.5863	0.4883
5	20	18	$S_{\lambda 0}$	1.3863	0.0000	1.2809	0.0000
5	20	18	$S_{\mu 0}$	1.3863	0.0000	1.3863	0.1054
5	20	18	$S_{\lambda 2}$	1.5863	0.2000	1.4809	0.2000
5	20	18	$S_{\mu 2}$	1.5863	0.2000	1.5863	0.3054
10	20	10	$S_{\lambda 0}$	0.6931	0.0000	0.0000	0.0000
10	20	10	$S_{\mu 0}$	0.6931	0.0000	0.6931	0.6933
10	20	10	$S_{\lambda 2}$	0.8931	0.2000	0.2000	0.2000
10	20	10	$S_{\mu 2}$	0.8931	0.2000	0.8931	0.8937
10	20	15	$S_{\lambda 0}$	0.6931	0.0000	0.4055	0.0000
10	20	15	$S_{\mu 0}$	0.6931	0.0000	0.6931	0.2877
10	20	15	$S_{\lambda 2}$	0.8931	0.2000	0.6055	0.2000
10	20	15	$S_{\mu 2}$	0.8931	0.2000	0.8931	0.4877
10	20	18	$S_{\lambda 0}$	0.6931	0.0000	0.5878	0.0000
10	20	18	$S_{\mu 0}$	0.6931	0.0000	0.6931	0.1054
10	20	18	$S_{\lambda 2}$	0.8931	0.2000	0.7878	0.2000
10	20	18	$S_{\mu 2}$	0.8931	0.2000	0.8931	0.3054
15	20	10	$S_{\mu 0}$	0.2877	0.0000	0.2877	0.6932
15	20	10	$S_{\mu 2}$	0.4877	0.2000	0.4877	0.8932
15	20	15	$S_{\lambda 0}$	0.2877	0.0000	0.0000	0.0000
15	20	15	$S_{\mu 0}$	0.2877	0.0000	0.2877	0.2877
15	20	15	$S_{\lambda 2}$	0.4877	0.2000	0.2000	0.2000
15	20	15	$S_{\mu 2}$	0.4877	0.2000	0.4877	0.4877
15	20	18	$S_{\lambda 0}$	0.2877	0.0000	0.1823	0.0000
15	20	18	$S_{\mu 0}$	0.2877	0.0000	0.2877	0.1054
15	20	18	$S_{\lambda 2}$	0.4877	0.2000	0.3823	0.2000
15	20	18	$S_{\mu 2}$	0.4877	0.2000	0.4877	0.3054
15	20	20	S_0	0.2877	0.0000	0.2877	0.0000
15	20	20	S_2	0.4877	0.2000	0.4877	0.2000
20	20	10	$S_{\mu 0}$	0.0000	0.0000	0.0000	0.6931
20	20	10	$S_{\mu 2}$	0.2000	0.2000	0.2000	0.8932
20	20	15	$S_{\mu 0}$	0.0000	0.0000	0.0000	0.2877
20	20	15	$S_{\mu 2}$	0.2000	0.2000	0.2000	0.4877
20	20	18	$S_{\mu 0}$	0.0000	0.0000	0.0000	0.1054
20	20	18	$S_{\lambda 2}$	0.2000	0.2000	0.0946	0.2000
20	20	18	$S_{\mu 2}$	0.2000	0.2000	0.2000	0.3054
20	20	20	S_0	0.0000	0.0000	0.0000	0.0000
20	20	20	S_2	0.2000	0.2000	0.2000	0.2000

MCMC

The settings for the MCMC were chosen, such that the acceptance rate was close to 50% for all proposals. Most of the settings were the same for all three models tested. The frequency with which each type of proposal was attempted were $c_r = 0.1$ for modifications of the number of lineages at the root, $c_\lambda = 0.8$ for modifications of the rate, and $c_A = 0.1$ for modifications of the model. Ξ_λ was set to $\log(16)$ and Ξ_A was set to $\log(2)$. The one parameter that did vary between the different models was the maximum change in the number of lineages at the root, $\Delta\dot{n}_r$. This value was obviously set to zero for the gene count model in which the number of lineages at the base was held constant. $\Delta\dot{n}_r$ was set to 1 for the gene tree model and to 10 for the gene count model, as the variance on the posterior distribution of $\dot{N}_r$ was much higher for the gene count model than the gene tree model.

I ran two MCMC chains independently for each simulation and each model. The chains were sampled from every 100 proposals and 1000 samples were taken from each chain. In order to determine an appropriate value for burn in, the likelihoods of the samples were plotted against sample number for the outputs from a large number of analyses covering a range of initial parameters and model comparisons. All analyses appeared to enter stationarity fairly quickly and so 100 samples were removed as a conservative burn-in. In order to determine if the chains had indeed achieved stationarity the two independent chains run for each tree were compared. The primary goal of this analysis was to determine how well each assignment of birth-death parameters to the branches of the taxon tree fit the data. Therefore for each pair of MCMC analyses I determined the fraction of samples for which each chain was in each of the four possible assignments and constructed a contingency table comparing the two different analyses. A χ^2 value was calculated from the contingency table, and for each set of 100 simulations and each model the calculated χ^2 values were plotted against the expected quantiles of χ^2. Analyses that did not fit the χ^2 distribution were continued for another 1000 samples.

Interpreting Results

The point of these analyses was to compare how well the different models inferred the assignments of the birth-death rates to the branches of the taxon tree. To do so I also had to evaluate five nuisance parameters, $\dot{n}_r$, $\vec{\lambda}_1$, $\vec{\lambda}_2$, $\vec{\mu}_1$ and $\vec{\mu}_2$. The proper evaluation of these parameters was critical to the evaluation of the rate assignments. I took the samples of these parameters in stationarity to be samples from the posterior distribution of these parameters, and used those values to calculate summary statistics using the R statistical programming language (R Development Core Team 2010). In this way the density of every particular parameter value is proportional to the product of its prior density and its likelihood under the model weighted by the posterior distributions of the other parameters in the model.

In order to evaluate how well the data supports each assignment of birth-death rates to the branches of the taxon tree I calculated Bayes factors (Kass and Raftery 1995). The

Bayes factor is a way of comparing how well two different hypotheses fit the data, and is defined as the ratio between the probability of the data under the two different hypotheses:

$$BF(H_1, H_2) = \frac{P(D|H_1)}{P(D|H_2)}$$

where D is the data and H_1 and H_2 are two alternative hypotheses. Values larger than one support H_1 and values less than one support H_2. Bayes factors are often reported as $\log_{10}$ Bayes factors, so that $BF_{10} = \log_{10}(BF)$, and I will do so here. In this case values greater than zero support H_1 and those less than zero support H_2. BF_{10} with an absolute value less than 0.5 represent very week support for the appropriate model, between 0.5 and 1 represent substantial support, greater than 1 is strong support and greater than 2 conclusive.

The fraction of times that a particular hypothethis is sampled during stationarity in a reversible-jump MCMC is an estimate of the posterior probability of that hypothesis given the data and the priors. In order to calculate the Bayes factor we must transform the posterior probability of the hypothesis to the likelihood of the hypothesis by dividing by its prior probability, so that:

$$BF(H_1, H_2) = \frac{P(D|H_1)}{P(D|H_2)} = \frac{P(H_1|D)P(H_2)}{P(H_2|D)P(H_1)}$$

In this case I am interested in comparing hypotheses in which the same birth-death rates are assigned to the two branches of the taxon tree to those in which different birth death rates are assigned, so that I want to calculate $BF_{10}(\vec{\lambda}_1 \neq \vec{\lambda}_2, \vec{\lambda}_1 = \vec{\lambda}_2)$ and $BF_{10}(\vec{\mu}_1 \neq \vec{\mu}_2, \vec{\mu}_1 = \vec{\mu}_2)$. The prior probabilities of each of these branch assignments are equal, so I calculated the Bayes factor as the ratio between the number of times that each of these rate assignments was sampled from the reversible-jump MCMC.

4.4.4 Results

MCMC Convergence

Comparisons of groups of 100 simulations to the χ^2 distribution had three different results: 1) the entire group of simulations appeared to fit the distribution well; 2) almost all the simulations fit the distribution well with a few outliers that had very large χ^2 values; 3) the entire set of simulations seemed to have unexpectedly high χ^2 values. Investigation of the outlier simulations indicated that these analyses failed to achieve convergence until after the 100 sample burn-in. Both chains for each of these outliers were run for an additional thousand samples and the last nine hundred samples were taken from each MCMC, leading to convergence of independent chains for each sample.

All analyses under the gene count probability with variable numbers of gene lineages at the root failed to converge in general, while all other analyses either appeared to converge completely or all converged except one or two outliers. This general failure of the gene count

analyses to converge seamed to be from both long burn ins and from too few samples taken, as the flatness of the posterior distribution of $\overset{\star}{N}_r$ lead to excessive autocorrelation between samples. The gene count analyses of the simulations starting with five genes had to be run for 4000 samples in order to achieve convergence with a burn-in of 1500. All the other gene count analyses had to be run for 2000 samples with a burn-in of 500.

Parameter Estimation

The Gene Tree model does a much better job estimating the number of gene lineages at the root of the taxon tree than the gene count model does (Figure 4.6). The posterior mean of the number of reconstructed lineages at the root under the gene tree model is very close to the actual number of gene lineages at the root for the vast majority of simulations. The only exception is when the expected number of genes in one of the tips is less than the number of gene lineages at the root; the posterior mean of $\overset{\star}{n}_r$ for these simulations is slightly less than the actual number of lineages used. This may be a correct estimate, because the number of reconstructed gene lineages may actually be less than the total number number of gene lineages as a consequence of gene loss. It is also possible that the model is underestimating the number of reconstructed lineages as a consequence of gene lineages that are lost in only one taxon lineage obscuring the number of actual gene lineages at the root of the taxon tree (see Figure 4.2c).

On the other hand, the gene count model is completely incapable of estimating the number of lineages at the root, as it relies only on the number of gene lineages in the terminal nodes and thus can not infer a trend in the number of gene lineages (Figure 4.6). The posterior mean of the number of genes at the root for the gene count model is generally slightly larger than the mean of the expected number of genes in the two tips. The among simulation median of the posterior means for $\overset{\star}{N}_r$ is unaffected by the actual number of genes in the root, but increases as the expected number of genes in $\overset{\times}{\tau}_2$ increases. The among simulation variance of the posterior mean for $\overset{\star}{N}_r$ does increase as the number of genes at the root decreases. This is to be expected, as I increased the birth-death parameters in order to generate the same values for $\overset{\hat{\times}}{n}_2$ while decreasing $\overset{\star}{N}_r$, thus increasing the variance in $\overset{\times}{n}_2$. The assignment of birth-death parameters used had little effect on the the estimation of the number of genes at the root, so long as the expected number of genes at the end of the process was held constant.

The estimate of the number of gene lineages at the root of the taxon tree is not only more accurate under the gene tree model than under the gene count model, but it is also more precise. The size of any credibility interval is much larger for the gene count model. This could be a consequence of the gene count model generally estimating larger values for the number of gene lineages at the root. However the two-tailed 50% credibility interval for the log of the number of gene lineages at he root is also much larger for the gene count model (Figure 4.7). With no information to constrain the root, a large range of values are reasonable under the gene count model.

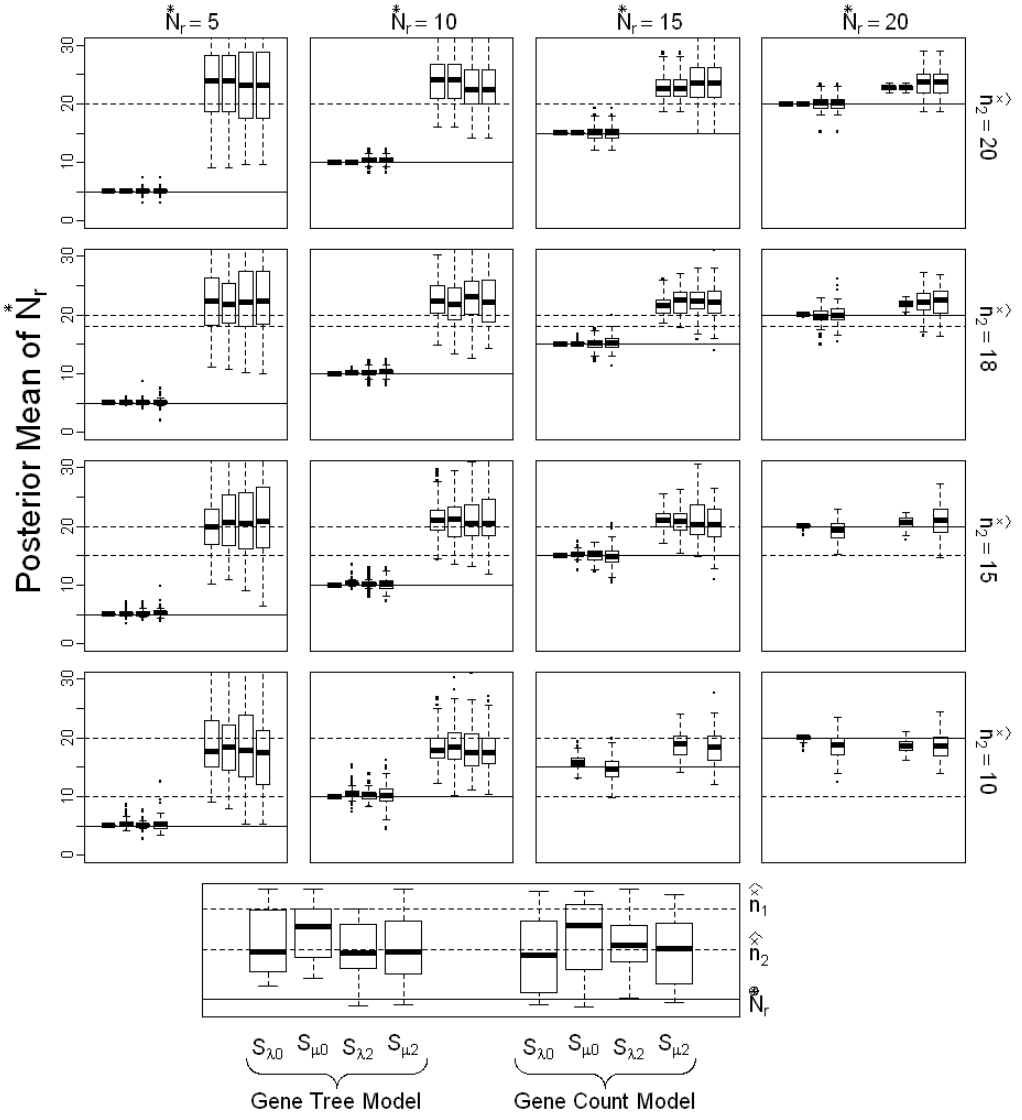

Figure 4.6: Posterior mean of the number of gene lineages at the root of the taxon tree for each simulation as analyzed under both the gene tree model and the gene count model. Each plot shows all the results for every simulation with a given number of gene lineages at the root (solid line) and a given expectation for the number of lineages at the tips of the tree (dashed lines). Each bar and whisker shows the distribution of the posterior mean for 100 simulations with a particular rate assignment and analyzed under a given model as described in the legend. The edge of the whiskers represents the most extreme value within 1.5 times the interquartile distance of either quartile.

and the dashed lines show the expected number of lineages at each of the tips of the taxon tree.

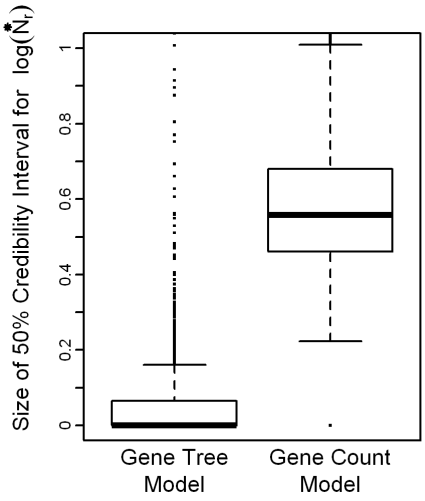

Figure 4.7: The size of the two-tailed 50% credibility interval for the log of the number of gene lineages at the root of the taxon tree for each simulation as analyzed under both the Gene Tree Model and the Gene Count Model. The edge of the whiskers represents the most extreme value within 1.5 times the interquartile distance of either quartile.

The gene tree model and the gene count model with the root fixed both do a fairly good job of estimating the birth-death parameters, while the regular gene count model fails miserably (Figure 4.8). Both the gene tree model and the gene count model with the true root do a good job of estimating $\vec{\lambda}$ over a wide range of values. As the the true value of $\vec{\lambda}$ increases, both models tend to underestimate the value, but are still fairly close to the simulation value. This underestimation may be from the pull of the prior. On the other hand, the gene count model is apparently incapable of distinguishing $\vec{\lambda}$ from the prior when it is not provided with the actual number of gene lineages at the root of the taxon tree. This is not surprising, as I have already shown that this model does a poor job of estimating the root and thus would be incapable of detecting any trends in the number of gene lineages. The gene tree model does a fair job of estimating $\vec{\mu}$, although the influence of the prior is fairly strong as values of $\vec{\mu}$ below the prior tend to be overestimated and those greater than the prior tend to be underestimated. The low estimates of the higher $\vec{\mu}_2$ values may also be a consequence of the prior mean being marginalized over all the assignments of branch rates, as the highest $\vec{\mu}_2$ values tend to be found in simulations where that value is much larger than $\vec{\mu}_1$, and thus would be brought down by the portion of samples in which those

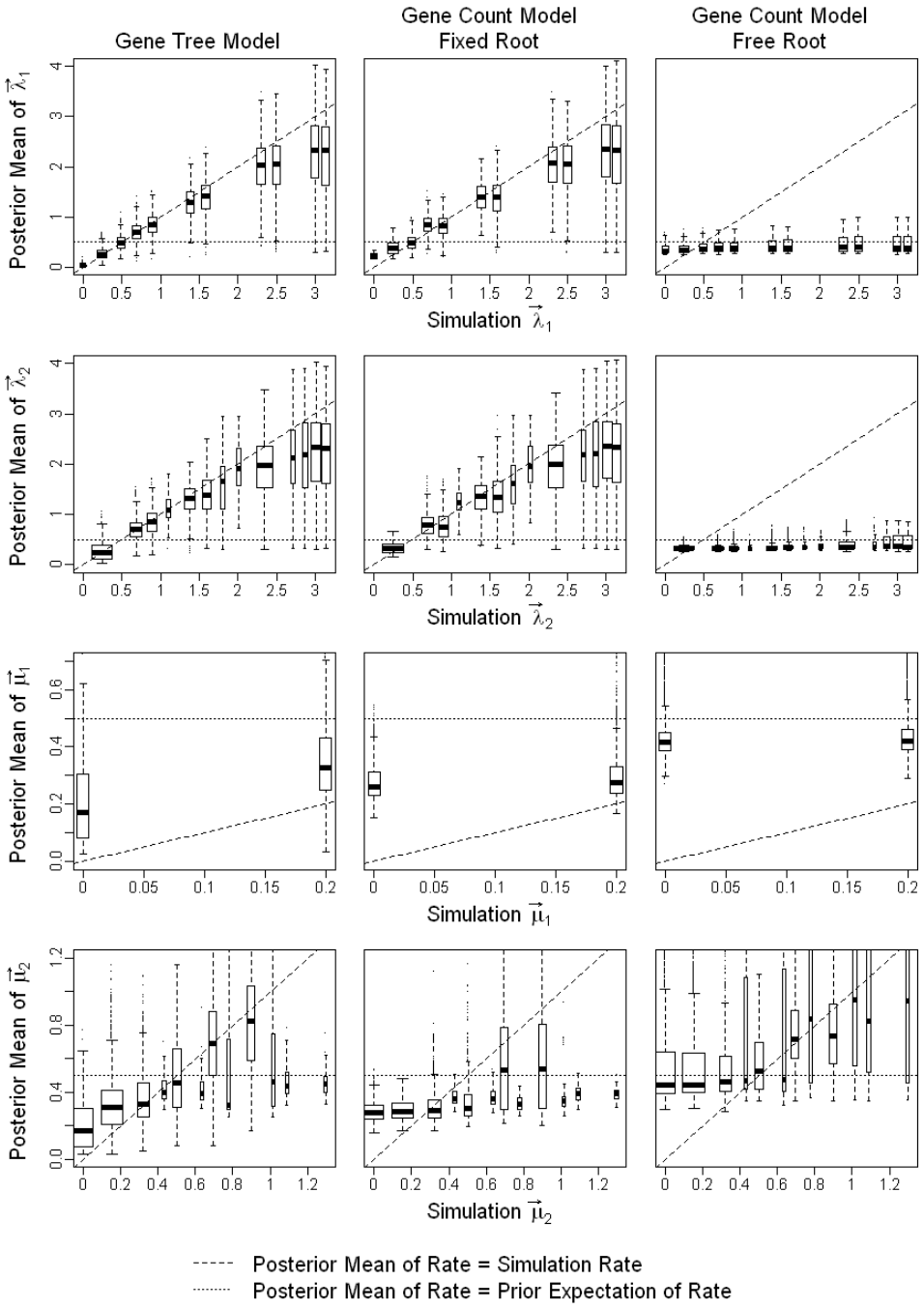

Figure 4.8: The posterior means of the four birth-death rates for the various simulations analyzed under each of the three gene evolution models, as a function of the actual rates used in the simulations. Each plot shows all the posterior means for a given rate as estimated under

a specific model. Each box shows the distribution of the posterior means from a number of different simulations that used the same or similar rates. Values for both λs are combined into bins of size 0.1, and values for μ_2 are combined into bins of size 0.05, in order to improve visualization. The relative widths of the boxes within each plot are proportional to the number of simulations that were used to construct that box. The diagonal dashed line has slope one, and thus is where a well performing analysis should estimate the rate. The horizontal dotted line is at the prior expectation for the rate.

two rates are equal. Neither of the gene count models do a reasonable job of estimating μ, although the model with a fixed root appears to do a slightly better job. Both appear to be strongly influenced by the prior, and they both have a fairly large between simulation variation in the estimate of large values for $\vec{\mu}_2$.

Comparison of Rate Assignments

The gene tree model can detect when the birth-death parameters differ between the two branches of the taxon tree and distinguish which of the two birth-death parameters it is that differs especially when the difference is large (Figures 4.9 and 4.10). Furthermore it rarely suggests that the rates do differ when they are in fact the same. In contrast the power of the gene count model to detect when the rates differed between branches was much smaller, and it was totally incapable of distinguishing which birth-death parameter differed. Providing the gene count model with information about the number of gene lineages at the root of the taxon tree increased its power to detect a difference in rates somewhat, but still left it impotent to determine which rate differed.

In order to see how well each assignment of rates to the branches of the taxon tree was supported under each analysis of each simulation, I calculated $\log_{10}$ Bayes factors comparing those assignment of rates in which the values of λ did vary between branches to those in which they did not, $BF_\lambda = BF_{10}(\vec{\lambda}_1 \neq \vec{\lambda}_2, \vec{\lambda}_1 = \vec{\lambda}_2)$, and comparing those assignment of rates in which the values of μ did vary between branches to those in which they did not, $BF_\mu = BF_{10}(\vec{\mu}_1 \neq \vec{\mu}_2, \vec{\mu}_1 = \vec{\mu}_2)$. $\hat{\hat{n}}_1$ is 20 for all these simulations, so that when $\hat{\hat{n}}_2$ is 20 all the rates are the same on both branches, and as $\hat{\hat{n}}_2$ decreases below 20 the magnitude of the difference between rates increases. Therefore, ideally BF_λ should be less than zero, when $\hat{\hat{n}}_2$ is 20 and for all simulations in which λ was the same throughout the process (eg: $S_{\mu 0}$ and $S_{\mu 2}$), and should be greater than zero when λ differed between branches especially when that difference was large. Furthermore, BF_μ should be less than zero, when $\hat{\hat{n}}_2$ is 20 and for all simulations in which μ was the same throughout the process (eg: $S_{\lambda 0}$ and $S_{\lambda 2}$), and should be greater than zero when μ differed between branches especially when that difference was large.

The results for the gene tree model come close to the ideal. A large majority of all

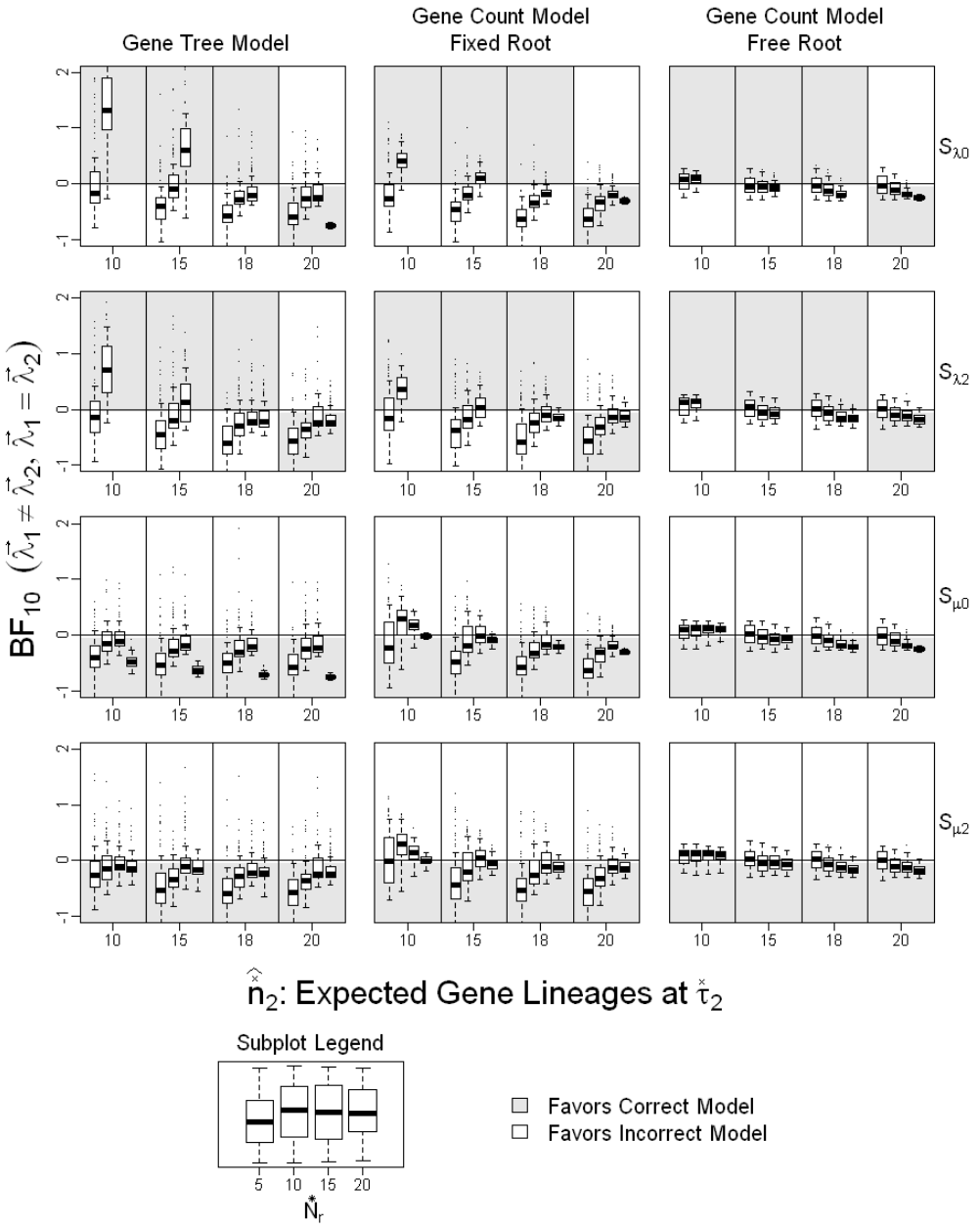

Figure 4.9: Log_{10} Bayes factor comparing assignments in which the two branches of the taxon tree have different values for λ to those assignments in which λ is the same for both branches, as calculated under each of the three gene evolution models. The rows of plots show results for simulations done with different rate assignments used in the simulations; and the columns show analyses done with different models of gene evolution. The simulations analyzed in each subplot

within a single plot differ in the number of lineages expected at the end of $\overset{\star}{\tau}_2$. Each bar shows the results for all the simulations done with the same parameter values, so that bars in a subplot differ only in the number of gene lineages at the root as described in the subplot legend. For simulations in which both branches had the same λ value the part of the plot with $\log_{10}$ Bayes factors below zero is colored gray; for those simulations in which the two values of λ were different the part of the plot with $\log_{10}$ Bayes factors greater than zero is colored gray.

simulations for sets of parameters in which λ was the same on both branches of the taxon tree had a negative BF_λ (Figure 4.9), as we would hope; and most simulations in which μ was the same on both branches had a negative BF_μ (Figure 4.10). However, neither of these hypothethes had strong support as both Bayes factors rarely fell below -1. The vast majority of simulations in which $\hat{\overset{\star}{n}}_2$ was 18 also had a negative BF_{10} under the gene tree model. This is to be expected as the difference between rates in these simulations is actually very small and thus hard to detect. However, as the value of $\hat{\overset{\star}{n}}_2$ fell to 15 and 10 the value of the appropriate Bayes factor rose drastically. For simulations using rate assignments $S_{\lambda 0}$ and $S_{\lambda 2}$ the values of BF_λ calculated under the gene tree model were much greater when the value of $\hat{\overset{\star}{n}}_2$ falls to 15 and 10, especially when $\overset{\star}{N}_r$ is large. The support for rate assignments with different values for λ on the different branches can be very strong with BF_λ exceeding 1 for at least 75% of the simulations under $S_{\lambda 0}$ with 10 expected genes in $\overset{\star}{\tau}_2$ and 10 gene lineages at the root of the tree. Furthermore, while the BF_λ under the gene tree model are larger for simulations with larger differences between λ on the two branches the values of BF_μ are essentially unchanged, as we would hope. For simulations using rate assignments $S_{\mu 0}$ and $S_{\mu 2}$, in which μ differs between branches, the situation is reversed; the values of BF_μ are much greater when the value of $\hat{\overset{\star}{n}}_2$ falls to 15 and 10, especially when $\overset{\star}{N}_r$ is large, while the values of BF_λ are essentially unchanged.

The values of BF_λ and BF_μ under the gene tree model are not only affected by the values of $\hat{\overset{\star}{n}}_2$ and the assignment of rates to the branches of the taxon tree. In general BF_λ appears to be higher for larger values of $\overset{\star}{N}_r$, except when $\vec{\mu}_1$ and $\vec{\mu}_2$ are both zero, in which case BF_λ is quite small (Figure 4.9). Nonetheless the increase in BF_λ as $\hat{\overset{\star}{n}}_2$ decreases is greater for higher values of $\overset{\star}{N}_r$. On the other hand increasing the value of $\overset{\star}{N}_r$ appears to always shift BF_μ towards the correct assignment of μs to the branches of the taxon tree (Figure 4.10). Therefore simulations with larger $\overset{\star}{N}_r$ have greater power to detect differences between μ on the branches of the taxon tree. Thus, the gene tree model can more easily detect differences in both λ and μ when $\overset{\star}{N}_r$ is higher; this may be because simulations with lower values for $\overset{\star}{N}_r$ have higher values for the birth-death parameters, which will lead to greater variation in the evolution of the gene tree, and thus may obscure the effects of the differences between the branches. Simulations using the higher value of $\vec{\mu}_1$ have larger between simulation variance in both BF_λ and BF_μ and both those values tend to be closer to zero, than they are for

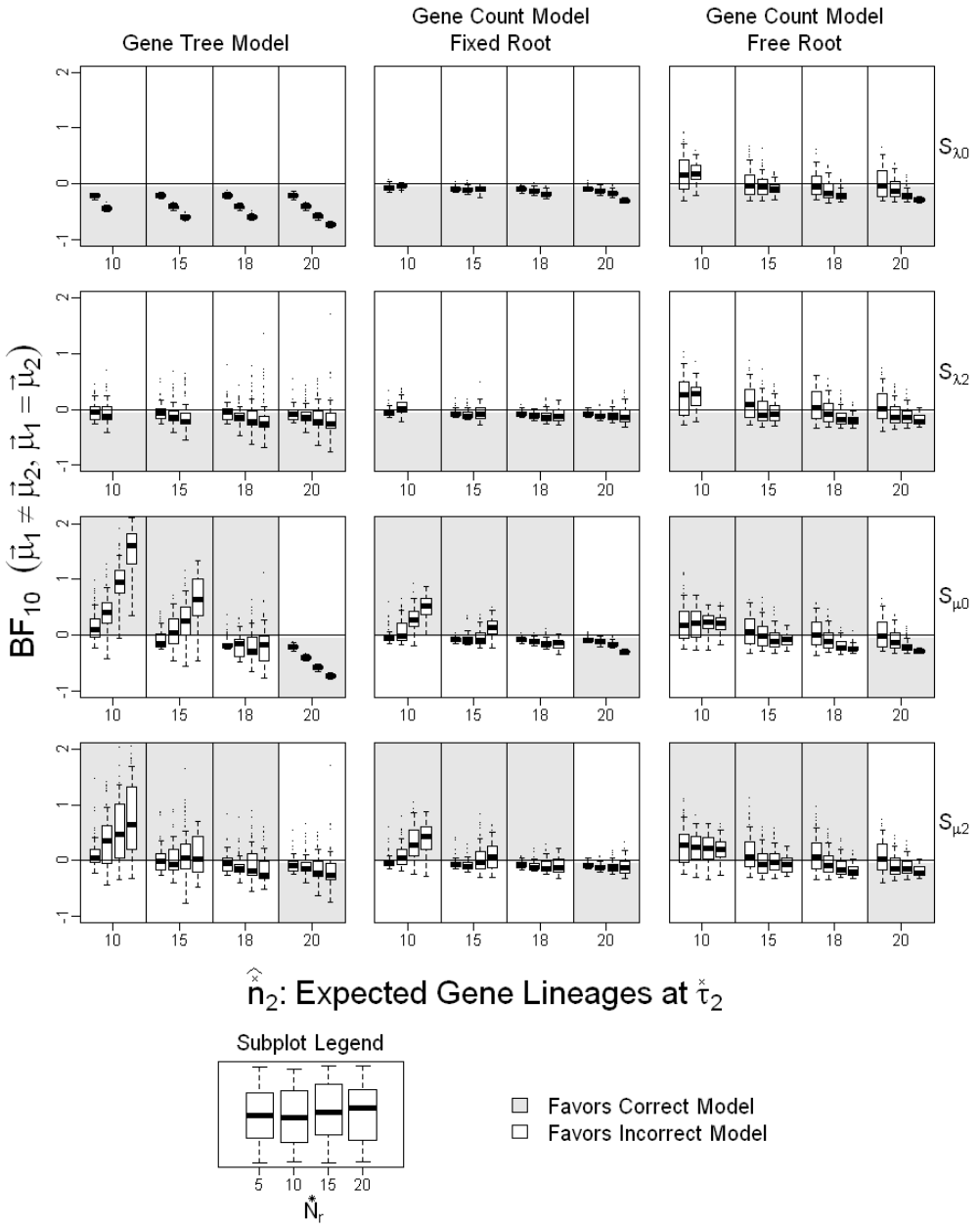

Figure 4.10: Log_{10} Bayes factor comparing assignments in which the two branches of the taxon tree have different values for μ to those assignments in which μ is the same for both branches, as calculated under each of the three gene evolution models. The rows of plots show results for simulations done with different rate assignments used in the simulations; and the columns show analyses done with different models of gene evolution. The simulations analyzed in each subplot

within a single plot differ in the number of lineages expected at the end of $\overset{\star}{\tau}_2$. Each bar shows the results for all the simulations done with the same parameter values, so that bars in a subplot differ only in the number of gene lineages at the root as described in the subplot legend. For simulations in which both branches had the same μ value the part of the plot with $\log_{10}$ Bayes factors below zero is colored gray; for those simulations in which the two values of μ were different the part of the plot with $\log_{10}$ Bayes factors greater than zero is colored gray.

simulations in which $\vec{\mu}_1$ is zero. This also leads to a decrease in power to detect a difference between branches of the taxon tree, and is also probably a consequence of the higher rates for the birth-death parameters obscuring the process of gene tree evolution.

On the other hand the gene count model has a difficult time distinguishing between different assignments of rates to the branches of the taxon tree (Figures 4.9 and 4.10). Both BF_λ and BF_μ do increase as $\hat{\overset{\star}{n}}_2$ decreases. However, these increases are very small, and the magnitude of the increase depends only on $\hat{\overset{\star}{n}}_2$ and is insensitive as to whether the difference in the number of genes in the two taxa is a consequence of the birth rate or the death rate. The performance of the gene count model is improved by supplying it with the true value of N_r. The values of BF_λ and BF_μ do increase by a noticeable amount as $\hat{\overset{\star}{n}}_2$ decreases, although neither one is very large. The values of BF_λ may be slightly higher for those simulations in which λ differs between branches and BF_μ may be larger for those simulations in which μ differs between branches even when the values of $\hat{\overset{\star}{n}}_2$ and $\overset{\star}{N}_r$ are the same. However, the differences are small and not nearly as large as those calculated under the gene tree model.

4.5 Reconciliation Analysis of Real Gene Families

In order to further study the gene tree model, I used it to compare phylogenies for two clades of metazoan genes to phylogenies for the animals in which those genes are found. Each clade of genes had previously been identified as monophyletic; nevertheless, for each gene clade I used MrBayes 3.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) to reconstruct their phylogeny based on gene sequences downloaded from Genbank in order to confirm the monophyly of the clade. Up until now I have assumed that the true gene tree was known. However, for both gene clades studied here there was much uncertainty in the gene tree, as there is for all gene family phylogenies. Therefore, I summed over the uncertainty in the gene topology by including a search for the gene tree based on the gene sequences in the MCMC, such that the estimates of the birth-death parameters and their assignments to the branches of the taxon tree are weighted by the posterior probabilities of the different possible gene trees under the model of nucleotide evolution. Each taxon tree was already well established in the literature.

I studied a clade of 46 cytoplasmic protein tyrosine kinase (PTK) genes found in

Drosophila melanogaster, *Caenorhabditis elegans* and *Homo sapiens*. Protein tyrosine kinase is a large gene family with 478 members in the human genome alone (Manning et al. 2002). They code for cellular enzymes that catalyze the phosphorylation of tyrosine residues, and thus play an important role in intracellular signaling (Hunter and Cooper 1985). The particular clade that I studied here was identified as monophyletic on the NCBI Clusters of Orthologous Groups (COGs) website (Tatusov et al. 2003) and code for a group of cytoplasmic PTKs. I chose this group, because they were found in all three of these genomes and showed a large difference in gene counts between genomes while the total number of genes remained reasonable.

Hox genes are a group of paralogous genes found in clusters within the genomes of all bilaterians as well as cnidarians (Ferrier and Holland 2001). They code for helix-turn-helix transcription factors and are expressed along the anterior-posterior axis during bilaterian development, where they play a critical role in pattern formation (Ruddle et al. 1994). They are arranged collinearly: the order of the genes within the cluster is the same as the order of their expression along the anterior-posterior axis. While there is only a single cluster found in most bilaterians, craniates have from three to eight clusters as a consequence of whole cluster - possibly whole genome - duplications (Wagner et al. 2003; Irvine et al. 2002). The hox genes form three monophyletic clades, the anterior hox, the medial hox and the posterior hox, each of which have independently expanded during bilaterian evolution (De Rosa et al. 1999; Kourakis and Martindale 2000), and it is widely believed that the complement of these genes was produced by gene duplication with only minimal gene loss playing a role. While there are only one or two posterior hox genes in each protostome genome, there are from three to six posterior hox genes in each deuterostome hox cluster, and several authors have suggested that this has lead to increased complexity of deuterostome posterior development (Izpisua-Belmonte et al. 1991; Holland 1992; Tabin 1992). To test this hypothethis I analyzed the evolution of posterior hox genes on a phylogeny of nine Bilateria taxa. If the expansion of posterior hox genes is critical for the evolution of deuterostomes, then we would expect the rate of duplication to be higher in the deuterostomes than it is in the protostomes. Furthermore we may ask whether the rates differed only early in deuterostome evolution or if they have continued to be elevated throughout the history of the Deuterostomia. Lastly we can determine the relative role of gene duplication as opposed to gene loss in the diversification of this gene family.

4.5.1 Protein Tyrosine Kinase

For the first analysis I used a group of cytoplasmic protein tyrosine kinase (PTK) genes which had been identified as monophyletic on the NCBI COGs website (Tatusov et al. 2003). This clade is labeled as KOG0194 on the COGs website, and includes two *D. melanogaster* genes, 42 *C. elegans* genes and two *H. sapiens* genes (Table 4.2). In order to root the gene tree and confirm that this clade was in fact monophyletic, I included eight closely related genes in the analysis. The COGs website identified a number of genes with high BLAST scores

Table 4.2: Protein Tyrosine Kinase Genes

Label	Genbank Protein Accession Number	Genbank Gene Accession Number	Gene Source
In Group			
Drosophila melanogaster			
Dm1	AAF54366	AE014297.2	mRNA
Dm2	AAF54367	AE003682.2	mRNA
Caenorhabditis elegans			
Ce1	NP496201	NM063800.2	mRNA
Ce2	NP498912	NM066511.1	mRNA
Ce3	NP502160	NM069759.1	mRNA
Ce4	NP501818	NM069417.2	mRNA
Ce5	NP501826	NM069425.1	mRNA
Ce6	NP494971	NM062570.1	mRNA
Ce7	NP501307	NM068906.1	mRNA
Ce8	NP501309	NM068908.3	mRNA
Ce9	NP494994	NM062593.1	mRNA
Ce10	NP492004	NM059603.5	mRNA
Ce11	NP501761	NM069360.1	mRNA
Ce12	NP501793	NM069392.2	mRNA
Ce13	NP501994	NM069593.2	mRNA
Ce14	NP501993	NM069592.1	mRNA
Ce15	NP501758	NM069357.1	mRNA
Ce16	NP502037	NM069636.1	mRNA
Ce17	NP500846	NM068445.3	mRNA
Ce18	NP501081	NM068680.1	mRNA
Ce19	NP501934	NM069533.1	mRNA
Ce20	NP492594	NM060193.1	mRNA
Ce21	NP491620	NM059219.1	mRNA
Ce22	NP491966	NM059565.1	mRNA
Ce23	NP500739	NM068338.1	mRNA
Ce24	NP499953	NM067552.1	mRNA
Ce25	NP492826	NM060425.1	mRNA
Ce26	NP492827	NM060426.2	mRNA
Ce27	NP496009	NM063608.1	mRNA
Ce28	NP501907	NM069506.1	mRNA
Ce29	NP500644	NM068243.1	mRNA
Ce30	NP502591	NM070190.1	mRNA
Ce31	NP490975	NM058574.1	mRNA
Ce32	NP493812	NM061411.1	mRNA
Ce33	NP506484	NM074083.1	mRNA
Ce34	NP502563	NM070162.1	mRNA
Ce35	NP502040	NM069639.3	mRNA
Ce36	NP503024	NM070623.1	mRNA
Ce37	NP503039	NM070638.1	mRNA
Ce38	NP500812	NM068411.1	mRNA
Ce39	NP500813	NM068412.1	mRNA
Ce40	NP491913	NM059512.1	mRNA
Ce41	NP490680	NM058279.1	mRNA
Ce42	NP498511	NM066110.3	mRNA
Homo sapiens			
Hs1	NP001996	NM002005.3	mRNA
Hs2	NP005237	NM005246.1	mRNA

Table 4.2: continued

Label	Genbank Protein Accession Number	Genbank Gene Accession Number	Gene Source
Out Group			
Drosophila melanogaster			
Out1	AAF49431.1	AE003526.2	mRNA
Caenorhabditis elegans			
Out2	NP494807	NM062406.2	mRNA
Homo sapiens			
Out3	NP002022	NM002031.2	mRNA
Out4	NP003206	NM003215.2	mRNA
Out5	NP004374	NM004383.2	mRNA
Out6	NP005148	NM005157.4	mRNA
Out7	NP009297	NM007313.2	mRNA
Out8	NP005224	NM005233.5	mRNA

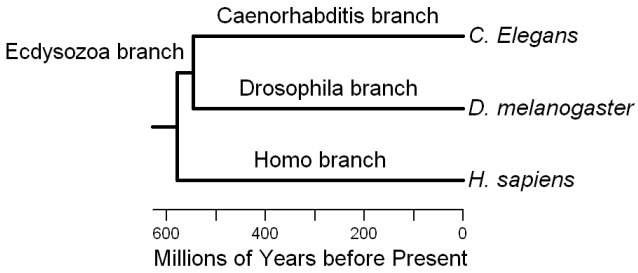

Figure 4.11: Taxon phylogeny on which the Protein Tyrosine Kinase gene tree was analyzed. The topology is from Aguinaldo et al. (1997) and the branch lengths are from Peterson and Butterfield (2005).

relative to the KOG0194 consensus sequence that were not found in KOG0194. I chose the eight genes from the genomes of these three taxa that had the highest BLAST scores, and included them in the analysis; this consisted of one gene each from *D. melanogaster* and *C. elegans* and six genes from *H. sapiens* (Table 4.2).

For the phylogeny of these three taxa, I used the generally accepted topology in which *C. elegans* and *D. melanogaster* are more closely related to each other than to *H. sapiens* (Figure 4.11) (Aguinaldo et al. 1997; Eernisse and Peterson 2004). The smallest clade containing both *C. elegans* and *D. melanogaster* is the Ecdysozoa, and thus I will refer to the internal branch of this phylogeny as the Ecdysozoa branch. I used the minimum evolution (ME) estimates of divergence times from Peterson and Butterfield (2005) for the branch lengths. This analysis was based on 1,747 nucleotides from the 18S rDNA gene sequence and fossil calibration of 12 nodes that occurred in the last 530 million years. The authors used r8s (Sanderson 2003) to estimate the divergence dates for the other nodes and confirmed those dates by comparing them to the acritarch fossil record. I will call this tree T_{HDC}.

Table 4.3: Hox Genes used in analysis.

Label	Genbank Protein Accession Number	Genbank Gene Accession Number	Gene Source(Nucleotides used)
In Group			
Nereis virens			
Ne-Post1	AAD46175	AF151672	DNA(100-379)
Ne-Post2	AAD46176	AF151673	DNA(305-622)
Caenorhabditis elegans			
Ce-php3	NP499573	NM067172	mRNA
Ce-nob1	NP001022941	NM001027770	mRNA
Drosophila melanogaster			
Dm-AbdB	NP996220	NM206498	mRNA
Strongylocentrotus purpuratus			
Sp-Hox9/10	*	AC165428	DNA(472551-804)
Sp-Hox11/13a	*	AC165428	DNA(416163-490)
Sp-Hox11/13b	*	AC165428	DNA(330122-29638)
Sp-Hox11/13c	*	AC165428	DNA(274889-5237)
Ciona intestinalis			
Ci-Hox10	BAE06496	AB210491	mRNA
Ci-Hox12	BAE06497	AB210492	mRNA
Ci-Hox13	BAE06498	AB210493	mRNA
Branchiostoma floridae			
Bf-Hox9	CAA84521	Z35149	DNA
Bf-Hox10	CAA84522	Z35150	DNA
Bf-Hox11	AAF81909	AF276811	DNA
		AF276812	DNA
Bf-Hox12	AAF81903	AF276813	DNA
		AF276814	DNA
Bf-Hox13	AAF81904	AF276815	DNA
Bf-Hox14	AAF81905	AF276816	DNA
		AF276817	DNA
Heterodontus francisci			
Hf-HoxD9	AAF44633	AF224263	DNA(74337-892,75444-685)
Hf-HoxD10	AAF44634	AF224263	DNA(67375-8113,69319-590)
Hf-HoxD11	AAF44635	AF224263	DNA(57461-8025,58571-803)
Hf-HoxD12	AAF44636	AF224263	DNA(49096-675,50437-675)
Hf-HoxD13	AAF44637	AF224263	DNA(42409-3012,43355-605)
Takifugu rubripes			
Tr-HoxD9a	ABF22465	DQ481668	DNA(159791-60367,160730-974)
Tr-HoxD10a	ABF22464	DQ481668	DNA(156468-7203,157727-8001)
Tr-HoxD11a	ABF22463	DQ481668	DNA(152005-566,153088-323)
Tr-HoxD12a	ABF22462	DQ481668	DNA(146557-7103,147397-635)
Mus musculus			
Mm-HoxD9	NP038583	NM013555	mRNA
Mm-HoxD10	NP038582	NM013554	mRNA
Mm-HoxD11	NP032299	NM008273	mRNA
Mm-HoxD12	NP032300	NM008274	mRNA
Mm-HoxD13	NP032301	NM008275	mRNA

4.5.2 Posterior Hox

The posterior hox genes form a widely recognized clade that is present throughout the Metazoa (Ruddle et al. 1994; Kourakis and Martindale 2000). In order to study the diversification of these genes in the deuterostomes I analyzed the phylogeny of the posterior hox genes from three protostomes and six deuterostomes with a large phylogenetic

Table 4.3: continued

Label	Genbank Protein Accession Number	Genbank Gene Accession Number	Gene Source(Nucleotides used)
Out Group			
Symsagittifera roscoffensis			
Sr-Hox4/5	AAN11405	AY117548	DNA(162-331,855-921)
Sr-post	AAN11406	AY117549	DNA(412-706)
Nereis virens			
Ne-Lox5	AAD46174	AF151671	DNA(269-669)
Caenorhabditis elegans			
Ce-mab5	NP498695	NM066294	mRNA
Ce-egl5	NP001021166	NM001025995	mRNA
Drosophila melanogaster			
Dm-AbdA	NP476693	NM057345	mRNA
Branchiostoma floridae			
Bf-Hox8	CAA84520	Z35148	DNA
Mus musculus			
Mm-HoxD8	NP032302	NM008276	mRNA

* The protein sequence does not have a separate Genbank entry. The translation was found in the gene sequence entry.

distribution (Table 4.3). I used complete samples of posterior hox genes from the three Protostome genomes: the polychaete, *Nereis virens*, has two posterior hox genes; the Nematode, *C. elegans* has two; and the fruit fly, *D. melanogaster*, has one. I also used complete samples of the posterior hox genes from three invertebrate Deuterostomes: the sea urchin, *Strongylocentrotus purpuratus*, has four posterior hox genes; the tunicate, *Ciona intestinalis*, has three; and the lancelet, *Branchiostoma floridae*, has six.

The hox cluster appears in multiple copies in vertebrate genomes. This is likely a consequence of multiple duplications that happened in a common vertebrate ancestor and in a teleost ancestor, and the homology of the individual clusters has been established (Irvine et al. 2002; Wagner et al. 2003). I wanted to avoid the signal of gene duplication left by these hox cluster duplications as the mechanism likely involved multiple whole genome duplications and thus differs from the mechanisms that has lead to an expansion in the number of posterior hox genes in all deuterostomes. Therefore, I only used hox genes from the D clusters of the three Vertebrate genomes that I analyzed, and only the Da cluster of the puffer fish. I chose the D cluster, because it appears to have retained a nearly full set of what are believed to be the ancestral craniate posterior hox genes in the three vertebrate genomes that I studied. The horn shark, *Heterodontus francisci*, and the mouse, *Mus musculus*, both have five posterior hox genes in their D cluster and the puffer fish, *Takifugu rubripes*, has four posterior hox genes in its hox Da cluster.

In order to root the posterior hox gene tree I included several medial hox genes from both Protostome and Deuterostome genomes that were used in this analysis (Table 4.3). I used one medial hox gene from each of the genomes of *N. virens*, *D. melanogaster*, *B. floridae* and *M. musculus*; and two from *C. elegans*. I also included one medial and the lone posterior hox gene from the acoel flatworm, *Symsagittifera roscoffensis*. The acoel flatworms are believed to be basal to the other Bilateria (Ruiz-Trillo et al. 1999), and thus, if there has been no gene

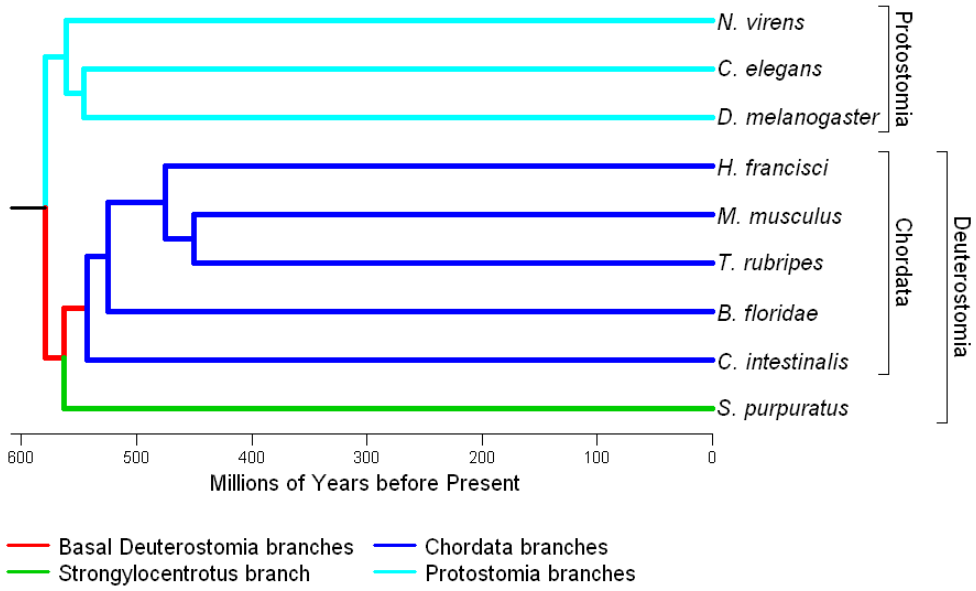

Figure 4.12: Taxon phylogeny of Bilateria on which the posterior hox genes were analyzed. The basal topology is based on Halanych (2004) and the internal Chordate topology is from Rowe (2004). Branching times are from ME estimates in Peterson and Butterfield (2005) and from Kishino et al. (2001). The branch colors indicate groups of branches that were assumed to have the same birth-death parameters. The names of these groups of branches are given in the legend.

loss, then Sr-post, the only posterior acoel hox gene, should also be basal to the posterior hox genes of the other Bilateria. However, even if it is found within the posterior hox genes of the other Bilateria, it still does not violate the monophyly of this gene clade, as no acoel genes are included in this analysis. Thus Sr-post is not a proper member of the outgroup, but instead is included as a test of the hypothethis that the common ancestor of all Bilateria had only one posterior hox gene.

For the taxon tree I use the modern view of Bilaterian relationships in which the deuterostomes and protostomes are both monophyletic and the protostomes are broken into two large clades, the Ecdysozoa, which includes both arthropods and nematodes, and the Lophotrochozoa, which includes annelids (Halanych 2004) (Figure 4.12). The deuterostomes contain both a monophyletic Echinodermata and Chordata as well as several other groups. The large scale phylogeny of Chordata has been stable for quite some time and I used this classic set of relationships in my analysis (Rowe 2004). Here urochordates are sister to the Euchordata, which contains the Cephalochordata and the Craniata. Within the Craniata the Osteichthyes are monophyletic and sister to the Chondrichthyes. For the dates of the nodes within the protostomes and for the protostome-deuterostome split I used the dates

from Peterson and Butterfield (2005), as I did for the PTK analysis. However, this paper did not have estimates of the relevant node dates within the deuterostomes. Therefore, for the intra-deuterostome divergence times I used the estimates from Blair and Hedges (2005) (Figure 4.12). These authors estimated the divergence times from a concatenated dataset of 325 proteins for 15 deuterostome taxa using a Bayesian local clock analysis (Kishino et al. 2001) with minimum constraints for 13 of 14 internal nodes and maximum constraints for 3 internal nodes derived from the fossil record. Both sets of divergence times were completely compatible with each other. I will call this tree T_{Bil}.

This taxon phylogeny has 16 branches that could potentially have different birth-death parameters under my model. This leads to more than 10^{20} possible rate assignments. However, we are only concerned with the relative rates in certain clades. In particular we want to know if the birth-death rates differ between the Deuterostomia and the Protostomia and if the birth-death rates were different early in deuterostome history. Thus in analyzing the posterior hox data, I assumed that the birth-death rates were equal in four groups of branch lengths (Figure 4.12). I assumed that the birth-death rates were the same within the Protostomia including the branch at the base, and I called these branches the Protostomia branches. I wanted to see if the birth-death process operated differently early in deuterostome history, so I defined the Basal Deuterostome branches as the branch at the base of Deuterostomia and the branch at the base of Chordata, and I assumed that these branches had the same rate. Defining these branches broke the rest of the deuterostome branches into two groups: the Strongylocentrotus branch, the terminal branch leading to *S. purpuratus*; and the Chordata branches consisting of all the branches within the chordates, but not the branch at the base.

4.5.3 Alignment and Outgroup Determination

I downloaded the nucleotide sequence of each gene used in this analysis as well as the amino acid sequences of the proteins for which they code from GenBank (Benson et al. 1998) (Tables 4.2 and 4.3). I aligned the amino acid sequences using ClustalW (Thompson et al. 1994; Larkin et al. 2007). I identified conserved domains in each protein using the Conserved Domains and Protein Classification server on the NCBI web site (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), and confirmed that all identified domains were properly aligned. I then used the nucleotide sequences to reverse translate the amino acid alignments. I further aligned these reverse translated nucleotide alignments by eye.

The gene tree reconciliation model used in this paper assumes that all the genes used are monophyletic with respect to all the other genes in the genomes from which they were sampled. Therefore it was necessary that I first confirm an outgroup, to determine that these genes are in fact monophyletic and second to root the gene tree. For each data set I used MrModelTest (Posada and Crandall 1998) run on PAUP* 4.0 (Swofford 2002) to determine the model of DNA substitution. I then used those models to analyze the alignments with

MrBayes 3.1 in order to infer the gene phylogeny (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). For each alignment I made two independent runs with four chains each sampling every 1000 generations using the default priors. In order to assure convergence, I ran the MCMC until the standard deviation of split frequencies for the last 75% of samples fell below 0.01. I then used the compare and split functions in AWTY to insure that the two runs had in fact converged and determine an appropriate burn-in (Nylander et al. 2008). I constructed a posterior phylogram from the distribution of posterior trees using the majority rule consensus tree command in Mesquite (Maddison and Maddison 2007).

4.5.4 Phylogeny Reconciliation Analysis

The data for this analysis is now the taxon tree, T_0, and the gene sequences, D, instead of the taxon tree and the gene tree, G_0. I want to find the posterior distribution of A, the assignment of birth death parameters to the branches of the taxon tree, but to do so, I must now deal with a number of nuisance parameters: Λ and M, the birth-death rates; $\dot{n}_r$, the number of gene lineages at the root of the taxon tree; G_0, the gene tree; and Θ, the nucleotide evolution parameters. I have already shown how to calculate the probability of a gene tree given a taxon tree and a set of birth death parameters, and there is an immense literature on how to calculate the probability of a set of gene sequences given a gene tree and a set of nucleotide evolution parameters. I can calculate the total probability by multiplying these two probabilities together.

$$P(D, G_0|A, \Theta, \Lambda, M, T_0) = P(D|\Theta, G_0)P(G_0|A, \Lambda, M, T_0)$$

I will use the MCMC so that my target distribution is the joint posterior density of G_0, A, Λ, M and Θ given T_0 and D. Thus in estimating a posterior distribution for A, $P(A|D, T_0)$, I sum over the posterior distributions of all these nuisance parameters, and in the process will also estimate posterior distributions for the nuisance parameters. All of these methods were implemented in TRUL.

Phylogeny Search

I calculated the probability of the gene sequences using the Kimura two-parameter model in which transitions and transversions occur at different rates (Kimura 1980). As the branch lengths of the gene phylogeny were also allowed to vary and were not constrained by anything other than the amount of nucleotide change, this model had only one free parameter, κ, the transition-transversion ratio. If we assume that the prior distributions for both the transition rate and the transversion rate are exponential, as we did for the other rates, then the prior density of κ would be

$$f(\kappa) = \frac{\tilde{\kappa}}{(\kappa + \tilde{\kappa})^2}$$

where $\tilde{\kappa}$ is the prior median of κ, and the ratio between the prior means of the transition and transversion rates. I assumed this prior distribution of κ, with a value of 2.0 for $\tilde{\kappa}$. I marginalized the calculation over all possible assignments of characters to the internal nodes Felsenstein (1973, 1981). I also assumed that the prior distributions for the branch lengths of the tree were exponential with a prior mean of 0.1. The prior for G_0 id derived from the model of gene tree evolution.

I used Extending Subtree Pruning and Regrafting (eSPR) for the topology rearrangement proposals (Lakner et al. 2008). My method only differed from Lakner et al. (2008) in that I only moved the subtrees from the non-root end of the chosen branch, as the position of the root is critical for the birth-death portion of the calculation. I used a probability of moving on to another branch, p_e, of 0.5. I modified both κ and the branch lengths in the same manor that I did the birth-death rates (subsection 4.3.2), by multiplying them by some number e^{ξ}, where ξ is chosen from the uniform distribution $(-\Xi/2, \Xi/2)$. Thus there were two new parameters for the MCMC: Ξ_κ, Ξ for the κ proposals, which I set to $\log(1.2)$; and Ξ_{bl}, Ξ for the branch length proposals, which I set to $\log(2)$. I also used Ξ_{bl} for the branch length modifications accompanying the eSPR proposals. The proposal ratio for these modifications can be calculated as in subsection 4.3.2.

Birth-Death Reconciliation Search

In order to compare the rates of gene duplication and loss on the different branches of the taxon tree, I not only calculated the probability of the DNA sequences given each gene tree but also the probability of each gene tree evolving on the appropriate taxon tree using the methodology described in section 4.3. However, as the tips of the gene tree are now characterized by gene sequences two tips must have the same gene sequences to be equivalent, and as the tree itself now has branch lengths two clades have to have not only the same topology but also the same set of branch lengths to be equivalent. This will never happen, so $k(G(\vec{\gamma}))$ is 2 for every branch in the tree. I used a reversible-jump MCMC to compare all possible assignments of birth-death rates to the branches of the taxon tree for the PTK analysis and to the groups of branches of the taxon tree for the posterior hox analysis. I assumed that at least one gene survived into each terminal of the taxon tree. I also assumed a prior expectation for each birth-death rate of 8 events/lineage/billion years, as the average estimate for Bilateria in Lynch and Conery (2003). Ξ_λ was set to $\log(8)$, Ξ_A was set to $\log(1.2)$ and $\Delta\mathring{n}_r$ was set to 1. The settings for the MCMC were chosen, such that the acceptance rate was close to 50% for all proposals.

Run Details

Since I considered the topology of the gene tree as a free parameter in my analysis, I had three additional types of proposals and thus had to determine the frequencies at which those proposal occurred. c_κ is the fraction of proposals for which I modified κ, c_{bl} is the fraction

Table 4.4: The order of branches in T_{HDC} as they appear in rate assignments from left to right.

Homo Branch
Ecdysozoa branch
Drosophila branch
Caenorhabditis branch

of proposals for which I modified a single branch length and c_G is the fraction of proposals for which I modified the gene tree topology, such that $c_\kappa + c_{bl} + c_G + c_r + c_\lambda + c_A = 1$ The frequencies with which each type of proposal was attempted were $c_r = 0.05$, $c_\lambda = 0.2$, $c_A = 0.3$, $c_\kappa = 0.05$, $c_{bl} = 0.1$, and $c_G = 0.3$.

I sampled from the MCMC every 1,000 generations and I took 10,000 samples for the PTK tree and 25,000 for the posterior hox trees. I had two independent runs for each analysis for which the starting parameter values were chosen at random from the prior. I chose the starting tree topology by randomly selecting trees from the samples MrBayes took in stationarity. I took a burn-in of 2000 samples for the PTK analysis and 12,000 samples for the posterior hox analysis. In order to assess whether the independent runs had in fact converged, I compared the two most important parameters, the tree topology and the assignment of rates. I compared the number of times each assignment of rates to the branches appeared in each run by constructing a contingency table and comparing the two independent runs with a χ^2-test using the R statistical programming language (R Development Core Team 2010). I confirmed that the tree topologies had converged and reached stationarity by using the compare and split commands in AWTY (Nylander et al. 2008). I also used AWTY to calculate the split frequencies and then calculated the average standard deviation of split frequencies in R.

Comparing Birth-Death Rate Assignments

In order to describe the different assignments of rates to the branches of the taxon tree, I will use a system in which the branches of the taxon tree are listed in a particular order. The rate for the first branch will be referred to as 1 and the rate for every other branch that is the same as the rate of the first branch will also be referred to as 1. The rate for the most leftward listed branch that is not equal to the first branch will be referred to as 2, as will all the other branches that have that same rate, and so on until all the rates are accounted for. So that for a tree with three branches, if all three branches had the same rate assigned to them, then $A = 111$, if they all had different rates assigned, then $A = 123$, and if the first and third branch had the same rate, while the second differed, then $A = 121$.

In these analyses, I will be dealing with the assignments to two different taxon trees, one consisting of *D. melanogaster*, *C. elegans* and *H. sapiens* used to analyze the PTK genes, T_{HDC}, and another more inclusive tree of Bilateria used to analyze the posterior hox genes,

Table 4.5: The order of branch groups in T_{Bil} as they appear in rate assignments from left to right.

Basal Deuterostomia branches
Strongylocentrotus branch
Chordata branches
Protostomia branches

T_{Bil}. Each will have a set of rate assignments that describes the assignments of the $\vec{\lambda}$s, A^{λ}_{HDC} and A^{λ}_{Bil}, and one that describes the assignments of the $\vec{\mu}$s, A^{μ}_{HDC} and A^{μ}_{Bil}. The full assignments of birth-death rates to the taxon trees will be referred to as A_{HDC} and A_{Bil} and will be represented as the $\vec{\lambda}$ assignment with a λ subscript followed by a slash and the $\vec{\mu}$ assignment with a μ subscript (i.e: $A_{HDC} = 1213_{\lambda}/1221_{\mu}$). The order of branches are shown in Table 4.4 for T_{HDC} and the order of branch groups are shown in Table 4.5 for T_{Bil}.

I compared different assignments of birth-death rates to the branches of the taxon trees using Bayes factors, in the same manor as I did for the simulations (section 4.4.3). However, unlike the rate assignments for the simple two taxon tree that I used for the simulations, many of the rate assignments that I wanted to compare for the more complex taxon trees used in these analyses have different prior probabilities, which must be corrected for in order to make these comparisons. When comparing two fully resolved rate assignments, say 111 to 112, both will have the same prior probability and so the ratio between their posterior probabilities will be equal to the Bayes factor. However, what if we want to compare those rate assignments in which the first rate and the second rate are unequal to those in which they are equal. I will refer to these two assignments as 12N and 11N respectively, where the N refers to indifference about that rate. As we can see 12N is made up of the fully resolved rate assignments 121, 122 and 123, while 11N is made up of 111 and 112. Thus the prior probability of 12N is 1.5 times the prior probability 11N and the appropriate correction should be made to their Bayes factors.

That appears to be sufficient, but under this scheme the distribution of assignments are weighted differently under 11N and 12N. To see why imagine that we take each of the assignments that make up 12N and set the first and the second rate to be equal to each other. In that case 121 and 122 will both be equivalent to 111, while only 123 will be equivalent to 112. Thus the assignments in which the third rate does not equal either of the other two rates are weighted more heavily under 11N than they are under 12N. To compensate for this fact, I adopted a prior scheme in which the prior probabilities are all still equal for all the less constrained fully resolved assignments (i.e. 121, 122 and 123), which make up the less constrained unresolved assignment (i.e. 12N). But for the more constrained unresolved assignment (i.e 11N), I calculate the prior probabilities of the fully resolved assignments that make it up (i.e. 111 and 112), by adding up the prior probabilities of their equivalent less constrained fully resolved assignments. So that for the case I already discussed, if 121, 122 and 123 all have prior probability p, then 112 will also have prior probability p because it

is only equivalent to 123, while 111 will have prior probability 2p as it is equivalent to 121 and 122. This scheme will put greater weight on those assignments in which more rates are equal to each other.

For those cases in which we want to compare the hypothesis that a set branches X all have the same $\vec{\lambda}$, and a set branches Y all have the same $\vec{\lambda}$ but X and Y do not have the same $\vec{\lambda}$, $X \subseteq A(\lambda_1)$ and $Y \subseteq A(\lambda_2)$, to the hypothesis that all members of both groups have the same rate, $\{X \cup Y\} \subseteq A(\lambda_0)$, then we can calculate an appropriate Bayes factor by weighting each fully resolved assignment under the first hypothesis by 1 and under the second hypothesis by $2^{|A(\lambda_0)|-|X|-|Y|}$. This same weighting scheme will work for $\vec{\mu}$. However, if we want to see if two groups have the same $\vec{r}$, meaning that they have the same $\vec{\lambda}$ and $\vec{\mu}$, then we must weight the assignments by $(2^{|A(\lambda_0)|-|X|-|Y|}+1)(2^{|A(\mu_0)|-|X|-|Y|}+1)-1$. For some more complex cases, I had to write code in R that would count assignments in order to generate the appropriate weighting scheme.

4.5.5 Protein Tyrosine Kinase Results

All the PTK genes had two well defined domains: SH2 for identifying phosphorylated kinases and PTKc for phosphorylating kinases. The *D. melanogaster* sequences also had an F-BAR domain for dimerization. All these domains were well aligned by ClustalW. The full reverse translated DNA alignment had 10503 characters. MrModelTest chose a GTR+Γ+I model, which I used to analyze the data set in MrBayes 3.1. After 6,000,000 generations the last 75% of samples from the two independent runs had an average standard deviation of split frequencies of 0.004164. Using AWTY (Nylander et al. 2008) I examined the posterior probabilities of splits over time using the slide command and plotted the splits between the two independent runs using compare; both analyses indicated that a burn-in of 30% was sufficient. Accepting splits with a posterior probability greater than 0.500 lead to a fully resolved tree (Figure 4.13). The monophyly of the in group was strongly supported (posterior probability=0.999), as was the monophyly of all three clades of in group genes from each taxon (posterior probability = 1.000) and many splits within the out group and the clade of *C. elegans* genes. The clade including all the *C. elegans* and *D. melanogaster* in group genes is only supported with a posterior probability of 0.732; the alternative arrangement in which the clade of *D. melanogaster* genes is sister to the clade of *H. sapiens* has a posterior probability of 0.255, and the arrangement in which the *H. sapiens* genes are sister to the *C. elegans* genes has a posterior probability of 0.013. The resolution of this node has a large effect on the birth-death analysis, as the topology with the *C. elegans* and *D. melanogaster* genes sister to each other requires no gene losses, while the topology with a *D. melanogaster* genes- *H. sapiens* genes clade requires a minimum of three gene losses and an additional gene duplication on the root (Figure 4.14). The uncertain topology within the *C. elegans* gene clade may have some small effect on the birth-death analysis, but it is unlikely to be very strong.

Comparisons of branch rate assignments and gene tree topologies indicated that the two

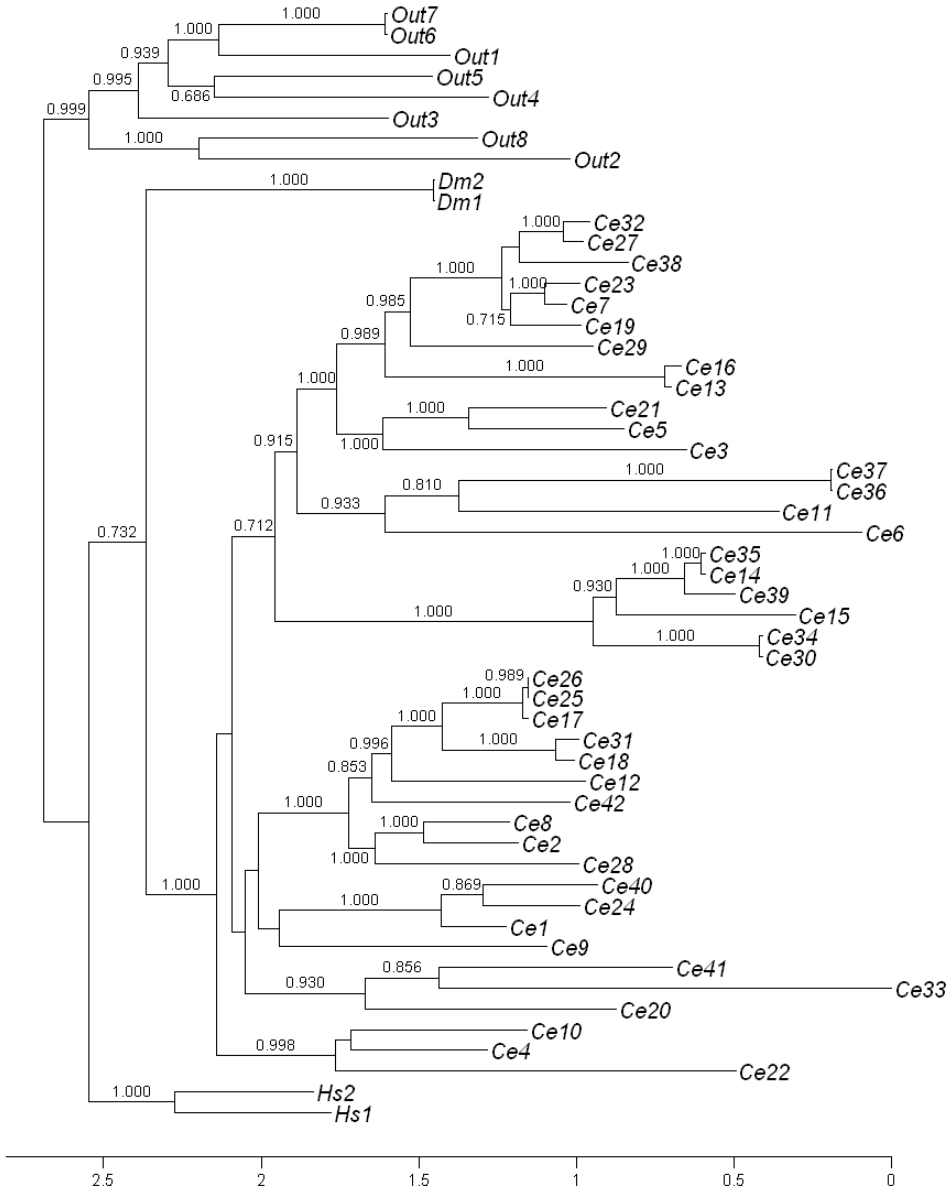

Figure 4.13: Posterior phylogeny of a clade of 46 cytoplasmic protein tyrosine kinase genes found in *D. melanogaster*, *C. elegans* and *H. sapiens*. The tree was reconstructed from an alignment of 10,503 nucleotide characters by MrBayes 3.1 using a GTR+Γ+I model of nucleotide substitution. All splits with a posterior probability greater than 0.500 are shown and the posterior probabilities of splits greater than 0.666 are shown on the appropriate branch.

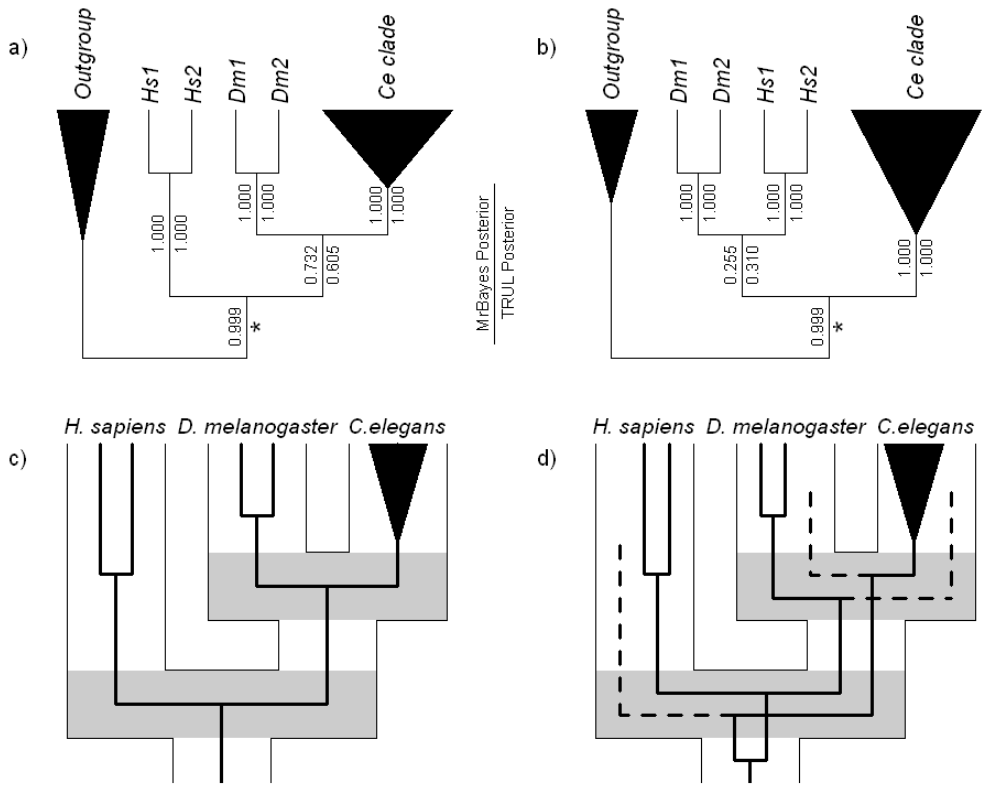

Figure 4.14: Two different possible topologies for the PTK genes and their maximum parsimony reconciliations. (a and b) Two different possible topologies for PTK gene tree. Numbers on the branches show posterior probabilities for each split as inferred by MrBayes 3.1 (to left of branch) and TRUL (to right of branch) using an alignment of 10,503 nucleotides. The asterisk refers to the fact that the in group was assumed to be monophyletic in the TRUL analysis and so that split is not a conclusion of the analysis. c) Maximum parsimony reconciliation of gene tree from (a) with the taxon tree. d) Maximum parsimony reconciliation of gene tree from (b) with the taxon tree. Gene lineages that we can infer must have existed but were lost before the present are shown with dashed lines.

MCMC chains for the TRUL PTK analysis appeared to be converged after removing the burn-in. I completely failed to reject the hypothethis that the distribution of rate assignments from the two chains were from the same distribution by constructing a contingency table and calculating a χ^2-value (χ^2=192.4, p=0.944). I plotted the split frequencies of the two runs against each other and they appeared to be very similar. Furthermore, the average standard deviation of split frequencies was 0.004677.

The phylogeny of PTK genes inferred by the reconciliation analysis is very similar to the in group phylogeny of the MrBayes analysis (Figure 4.15). The in group genes from each taxon were found to be monophyletic in all samples. The topology of the *C. elegans* PTK clade differs somewhat between the two trees, but is overall very similar. Some of these differences require that one dissolves splits that are well supported in at least one of the trees. Of course changes in the topology of the *C. elegans* PTK clade will not have a large affect on the reconciliation analysis. On the other hand the relationship between the gene clades of each taxon will affect the analysis. Although both analyses have their majority of samples from trees in which the *D. melanogaster* genes are most closely related to the *C. elegans* genes, that arrangement is more strongly supported by the MrBayes analysis (posterior probability = 0.732) than by the gene tree reconciliation analysis (posterior probability = 0.605). This is surprising; the gene tree reconciliation analysis also considers the fit of the gene tree to the taxon tree and this arrangement is more parsimonious than either of the other arrangements, so we would expect the reconciliation analysis to support it more strongly (Figure 4.14). The difference in support is likely a consequence of the different models of nucleotide substitution used in the two analyses.

It is apparent that the maximum parsimony reconciliation of the best supported PTK tree to its taxon tree requires no gene losses (Figure 4.14c). Under this hypothethis, the most recent common ancestor (MRCA) of the Bilateria had only one gene in this clade which was passed on to the MRCA of the Ecdysozoa. On the other hand the other two possible arrangements of the tree taxon PTK clades take up almost 40% of the posterior probability and require at least two gene lineages at he root and in the MRCA of the Ecdysozoa (Figure 4.14d). This is reflected in the posterior distribution of $\dot{n}_r$, which is one 46% of the time, two 32% of the time, and is less than 5 96% of the time (Figure 4.16). The MCMC does spend a significant amount of time sampling from reconciliations with more gene lineages at the root other than the maximum parsimony reconciliation, but usually stays pretty close to what we would expect under the maximum parsimony reconciliation.

The single assignment of birth-death rates to the branches of T_{HDC} with the highest posterior probability is one in which all the branches have the same $\vec{\mu}$, and $\vec{\lambda}$ is the same for the Homo branch, the Drosophila branch and the Ecdysozoa Branch, but is different for the Caenorhabditis branch ($BF_{10}(A_{HDC} = 1112_{\lambda}/1111_{\mu}, A_{HDC} \neq 1112_{\lambda}/1111_{\mu}) = 0.6764$) (Figure 4.17). However, varying the assignments of $\vec{\mu}$ and $\vec{\lambda}$ for the Ecdysozoa has little effect on the posterior probability. Assignments in which $\vec{\lambda}$ for the Caenorhabditis branch does not equal $\vec{\lambda}$ for the Homo branch or the Drosophila branch are well supported

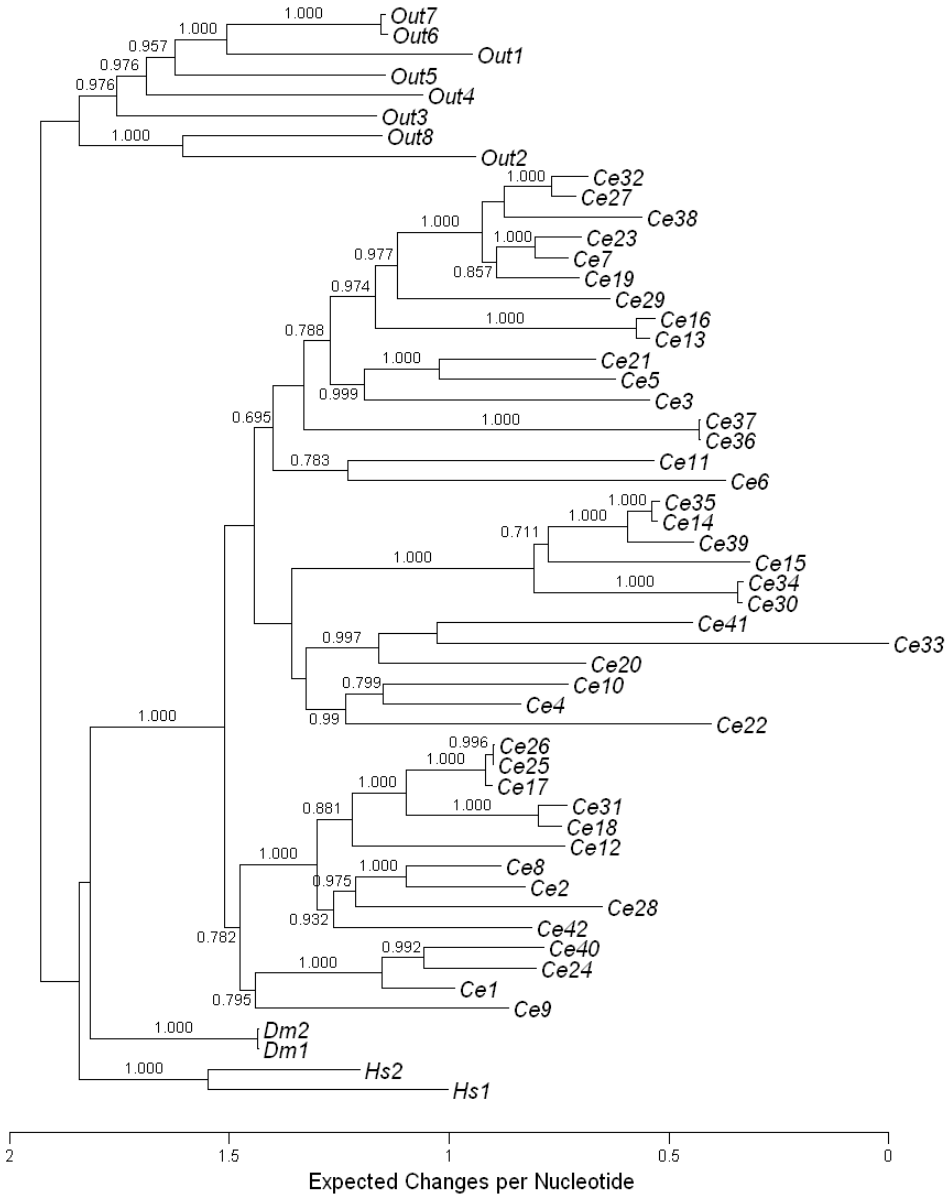

Figure 4.15: Posterior phylogeny of a clade of 46 cytoplasmic protein tyrosine kinase genes found in *D. melanogaster*, *C. elegans* and *H. sapiens*. The tree was reconstructed from an alignment of 10,503 nucleotide characters by TRUL. All splits with a posterior probability greater than 0.500 are shown and the posterior probabilities of splits greater than 0.666 are shown on the appropriate branch.

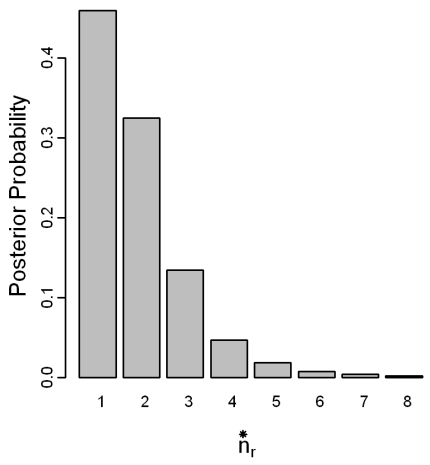

Figure 4.16: Posterior distribution of $\mathring{n}_r$ for ML PTK tree, when compared to its taxon tree.

Table 4.6: Log_{10} Bayes Factor for the hypothethis that $\vec{r}$ for the branch in each row is greater than $\vec{r}$ for the branch in each column vs the hypothethis that those values are equal, $BF_{10}(\vec{r}_1{>}\vec{r}_2, \vec{r}_1{=}\vec{r}_2)$.

		Branch_2			
		Homo	Ecdysozoa	Drosophila	Caenorhabditis
Branch_1	Homo	-	-0.09371	-0.32622	-0.80911
	Ecdysozoa	-0.02818	-	-0.02381	-0.38231
	Drosophila	-0.29924	-0.07678	-	-0.79894
	Caenorhabditis	0.72766	0.34176	0.66616	-

($BF_{10}(A^{\lambda}_{HDC} \in \{1N12, 1N23\}, A^{\lambda}_{HDC} \notin \{1N12, 1N23\}) = 0.5652$), but the support for $\vec{\lambda}$ being equal on the Homo branch and the Drosophila branch is pretty week ($BF_{10}(A^{\lambda}_{HDC} = 1N1N, A^{\lambda}_{HDC} = 1N2N) = 0.0028$). The $\vec{\mu}$ assignments have less effect on the likelihood than the $\vec{\lambda}$ assignments. The assignments in which $\vec{\mu}$ is the same on the Caenorhabditis branch, the Drosophila branch and the Homo branch are the best supported, while those in which all three of those branches have different $\vec{\mu}$s are the worst supported. All the $\vec{\mu}$ assignments in which only two of those branches have the same $\vec{\mu}$ are supported approximately equally.

Comparing the values of $\vec{r}$ between pairs of branches instead of assignments of $\vec{\lambda}$ or $\vec{\mu}$ to all the branches helps to clarify the picture. Table 4.6 shows Bayes Factors comparing the hypothethis that the value of $\vec{r}$ is larger on one branch than it is on another to the hypothethis that $\vec{r}$ is the same for both branches. This is essentially a series of one tailed tests for the relationships between the birth-death processes on the various branches of T_{HDC}.

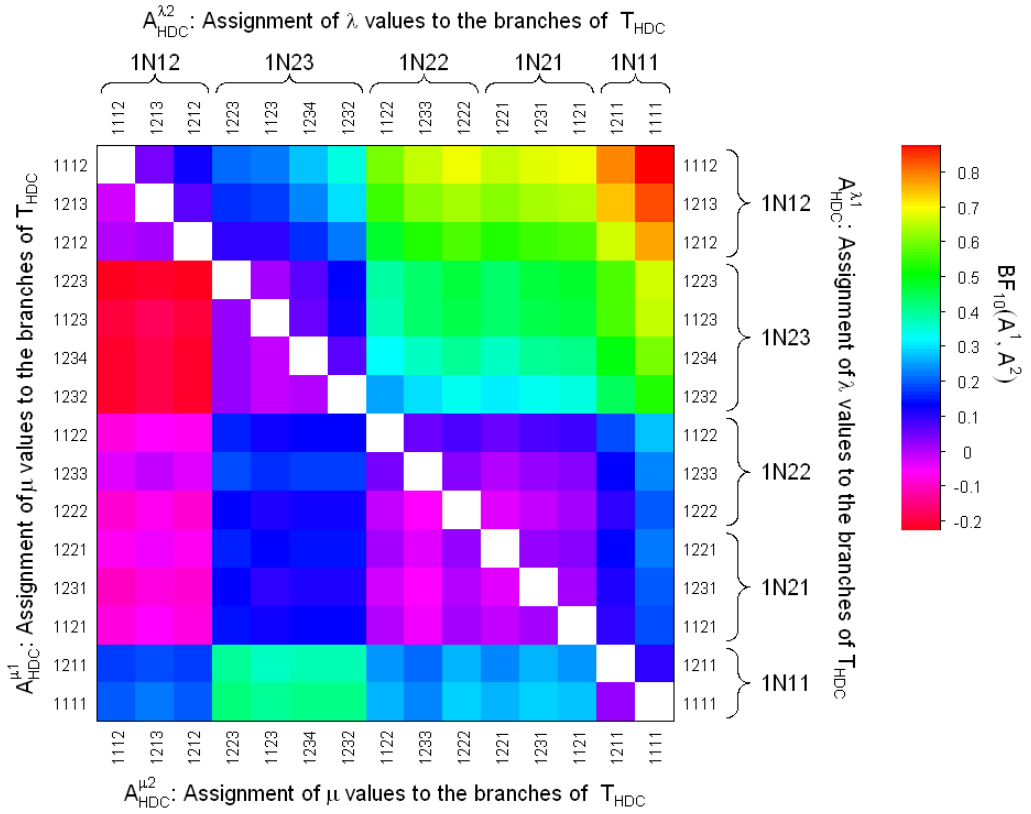

Figure 4.17: Log_10 Bayes Factors comparing assignments of birth-death rates for the PTK gene tree to the branches of the taxon tree. Above the diagonal shows comparisons of different $\vec{\lambda}$ assignments showing support for assignments in the rows versus those in the columns, $BF_{10}(A^{\lambda 1}_{PTK}, A^{\lambda 2}_{PTK})$. These assignments are further subdivided into those that have the same $\vec{\lambda}$ assignments to the terminal branches of the taxon tree. Below the diagonal shows comparisons of different $\vec{\mu}$ assignments showing support for assignments in the rows versus those in the columnss, $BF_{10}(A^{\mu 1}_{PTK}, A^{\mu 2}_{PTK})$.

Table 4.7: Log_{10} Bayes Factors show that the data supports a model in which the difference between $\vec{r}$ for the Caenorhabditis branch and the other branches is a consequence of differences in $\vec{\lambda}$, not $\vec{\mu}$. Each entry shows $BF_{10}(H_1, H_2)$.

	H_1	$\vec{\lambda}_C \neq \vec{\lambda}_1 \cap \vec{\mu}_C = \vec{\mu}_1$	$\vec{\lambda}_C \neq \vec{\lambda}_1 \cap \vec{\mu}_C = \vec{\mu}_1$
	H_2	$\vec{\lambda}_C \neq \vec{\lambda}_1 \cap \vec{\mu}_C \neq \vec{\mu}_1$	$\vec{\lambda}_C = \vec{\lambda}_1 \cap \vec{\mu}_C \neq \vec{\mu}_1$
$Branch_1$	Homo	0.41236	0.64284
	Ecdysozoa	0.26521	0.37216
	Drosophila	0.42332	0.65445

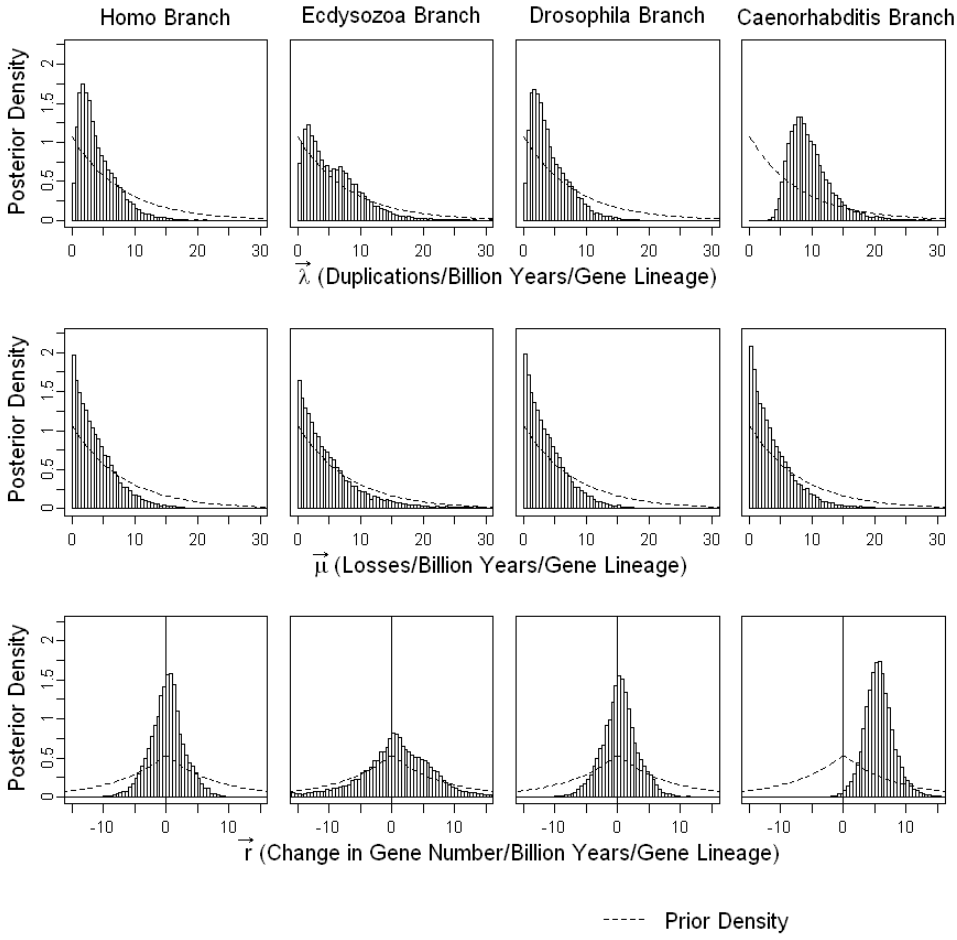

Figure 4.18: Posterior distribution of birth-death rates for the evolution of the PTK tree. Samples from the MCMC for $\vec{\lambda}$, $\vec{\mu}$ and $\vec{r}$ for each branch of T_{HDC} are grouped into bins of 0.5 Events/Gene Lineage/Billion Years. The prior distribution is shown as a dashed line for comparison.

We see here that there is solid support for $\vec{r}$ being larger on the Caenorhabditis branch than it is on any of the other branches, especially the Homo branch and the Drosophila branch. We can also be certain that $\vec{r}$ for the Caenorhabditis branch is not lower than $\vec{r}$ on any of the other branches. The rest of the comparisons do not strongly favor one hypothethis over another, especially those involving the Ecdysozoa branch. Furthermore, there is also a fair amount of support for the hypothethis that the difference in $\vec{r}$ between the Caenorhabditis branch and the other branches is a consequence of differences in $\vec{\lambda}$ rather than $\vec{\mu}$, as the assignments in which $\vec{\lambda}$ and not $\vec{\mu}$ differs between the pair of branches have a much higher posterior probability than those in which both $\vec{\lambda}$ and $\vec{\mu}$ differ, or only $\vec{\mu}$ differs (Table 4.7).

The posterior distribution of the birth-death rates reflects what we have already seen

from comparing different rate assignments (Figure 4.18). $\vec{\lambda}$ for the Homo branch and the Drosophila branch both have strong modes at about 1.75 duplications/lineage/billion years. The $\vec{\lambda}$ for the Caenorhabditis branch also has a strong posterior mode at about 8 duplications/lineage/billion years. On the other hand, $\vec{\lambda}$ for the Ecdysozoa branch has two modes that correspond to the modes of the other three branches and its posterior distribution appears to be a mix of the posterior distribution for the other three branches and the prior, indicating that the data provides little information about what $\vec{\lambda}$ actually is on that branch. The posterior distributions of $\vec{\mu}$ appear to be more strongly influenced by their prior distributions than the posterior distributions of $\vec{\lambda}$ are. There is some information driving all these distributions closer to zero than the prior distribution. This effect is particularly strong for the Caenorhabditis branch and particularly weak for the Ecdysozoa branch, for which the posterior distribution is very close to the prior. The posterior distributions of $\vec{r}$ are essentially what we would expect given the distributions for $\vec{\lambda}$ and $\vec{\mu}$. However, it is informative to see that the distributions of $\vec{r}$ for the Homo branch, the Ecdysozoa branch and the Drosophila branch all have modes near zero, and for the Homo branch and the Drosophila branch, these modes are much stronger than the prior. The Caenorhabditis branch has a very strong mode for $\vec{r}$ near 5.5 changes/lineage/billion years and interestingly the variance for $\vec{r}$ appears to be less than it is for $\vec{\lambda}$.

4.5.6 Posterior Hox Results

All the posterior Hox genes had a single homeodomain for DNA binding, which was well aligned by ClustalW. The full reverse translated DNA alignment had 1,866 characters. MrModelTest chose a GTR+Γ+I model, which I used to analyze the data set in MrBayes 3.1. After 20,000,000 generations the last 75% of samples from the two independent runs in MrBayes had a average sandard deviation of split frequencies of 0.004205. Using AWTY (Nylander et al. 2008) I examined the posterior probabilities of splits over time using the slide command and plotted the splits between the two independent runs using compare; both analyses indicated that a burn-in of 30% was sufficient.

The posterior hox gene tree is not well resolved (Figure 4.19). The monophyly of the in group is strongly supported (posterior probability(pp)=1.000), but only as long as we include Sr-Post, which is nested well inside the in group. This is a common result in phylogenetic analyses of posterior hox genes, in which the posterior hox are clearly resolved as monophyletic, but there is little resolution within the posterior hox (see De Rosa et al. 1999; Kourakis and Martindale 2000). The nodes that are well resolved yielded some unexpected results. All the protostome posterior hox are found nested within the chordate posterior hox, implying that the most recent common ancestor of bilaterians had multiple posterior hox genes, many of which were lost in the protostomes. The analysis recovered a monophyletic gnathostome HoxD10 (pp=1.000), HoxD11 (pp=1.000), HoxD12 (pp=0.999) and HoxD13 (pp=0.999). However, within the HoxD11 (pp=0.965) and HoxD12 (pp=0.805)

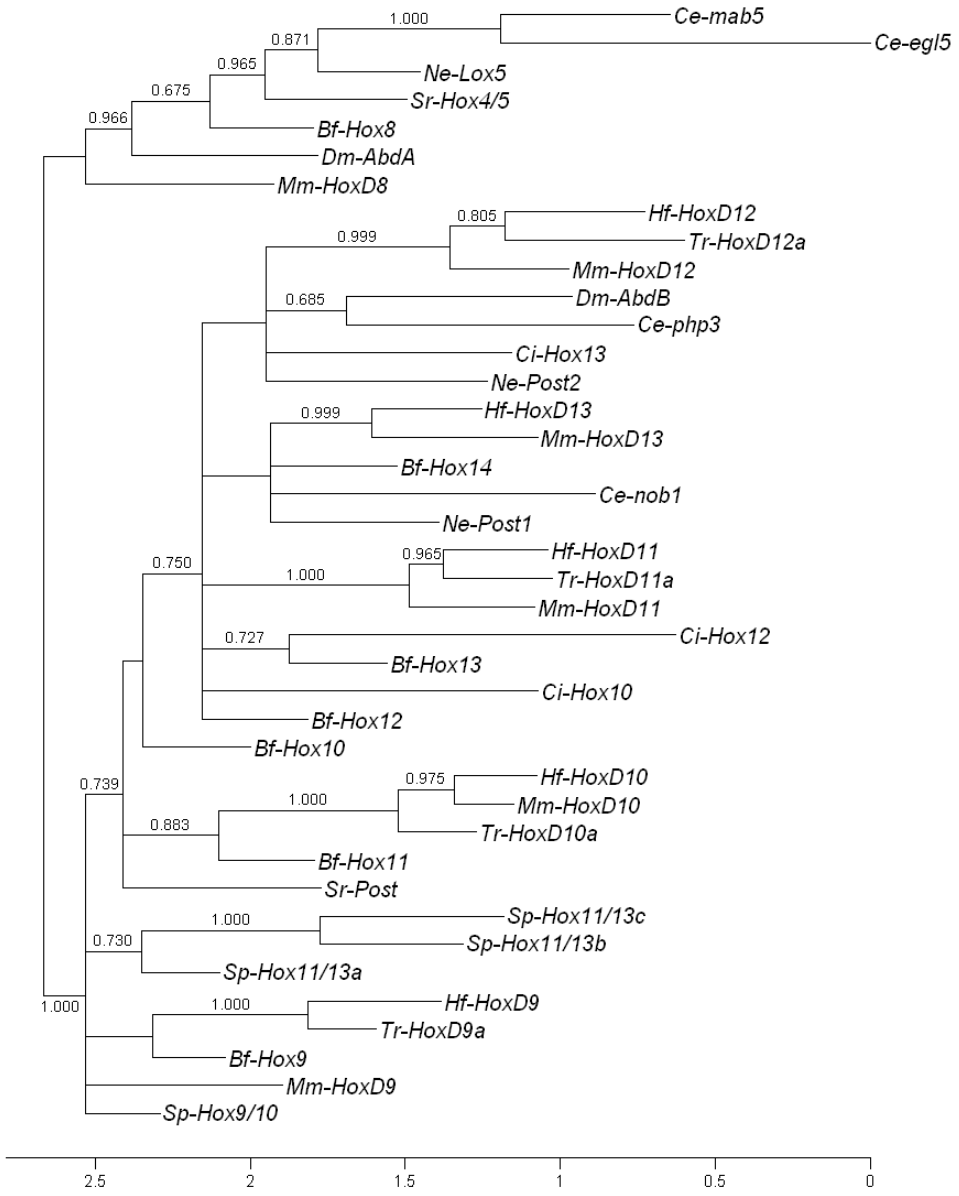

Figure 4.19: Posterior phylogeny of a clade of 32 posterior hox genes found in nine Bilaterian taxa. The tree was reconstructed from an alignment of 1,866 nucleotide characters by MrBayes 3.1 using a GTR+Γ+I model of nucleotide substitution. All splits with a posterior probability greater than 0.500 are shown and the posterior probabilities of splits greater than 0.666 are shown on the appropriate branch.

the *T. rubripes* gene was most closely related to the *H. francisci* and the *M. musculus* and *H. francisci* HoxD10 genes formed a clade (pp=0.975), whereas a maximum parsimony interpretations of the reconciliation would expect monophyletic osteichthyes genes. One interpretation of this result is that the homology of the hox clusters has been incorrectly assigned, but there is no consistent pattern within these results to assign an alternative homology. Furthermore, a monophyletic gnathostome HoxD9 is rejected (pp=0.322), and the *C. intestinalis* and *B. floridae* posterior hox rarely are grouped with the appropriate gnathostome genes. Another feature of note is the monophyletic Sp-Hox11/13 (pp=0.730), which has been found in previous studies (Cameron et al. 2006).

The last 13,000 samples from the two independent TRUL runs appeared to be completely converged. I was extremely far from rejecting the null hypothethis that the distribution of assignments from the two runs were from the same distribution (χ^2=181.7, p=0.980). Furthermore, the average standard deviation of split frequencies was only 0.00104. Visual comparisons of split frequencies and slide analyses also indicated that the two runs had converged.

The posterior gene tree for the TRUL analysis is no better resolved than the posterior gene tree from the MrBayes analysis, but there are substantial differences in topology (Figure 4.20). The vertebrate HoxD10 (pp=0.982), HoxD11 (pp=0.999), HoxD12 (pp=0.999) and HoxD13 (pp=0.866) each form clades, as they do for the MrBayes tree, but now Ci-Hox13 is grouped with the vertebrate HoxD13 genes (pp=0.844). Furthermore, the *C. elegans* (pp=0.917) and *N. virens* (pp=0.584) genes are each monophyletic unlike in the MrBayes trees. The TRUL tree also has a large polytomy consisting of each of the protostome posterior hox clades, the vertebrate hoxD10 through hoxD13 clades, the *C. intestinalis* genes and the *B. floridae genes*, which is at least partially resolved in the MrBayes tree. Finally, the TRUL tree has a completely monophyletic *S. purpuratus* posterior hox.

Three different factors could account for the difference between the tree reconstructions. The two analyses used different models of nucleotide substitution, the TRUL analysis did not include Sr-post and the TRUL analysis took the prior distribution of tree topologies from the gene tree evolution model. It is difficult to determine the effect of the first two factors, but it does appear that the topology prior had a large affect on the posterior distribution of trees. Posterior hox trees sampled from the TRUL analysis were in general more compatible with T_{Bil} than those sampled from MrBayes (Figure 4.21). This is not surprising, because the TRUL analysis took the tree reconciliation into account, and so it should tend to resolve topological uncertainties in favor of topologies that require fewer gene duplications and losses. It should be noted that both analyses produced trees with many more duplications and losses than the five duplications and three losses required by the distribution of gene numbers among taxa.

The results of the reconciliation analysis are largely driven by the incongruence between the gene tree and the species tree. It predicts a large number of reconstructed gene lineages at the root (median = 16, 50% credibility interval = 12 to 28). This implies that there has been a large amount of gene loss in every lineage and that is reflected in the birth-death

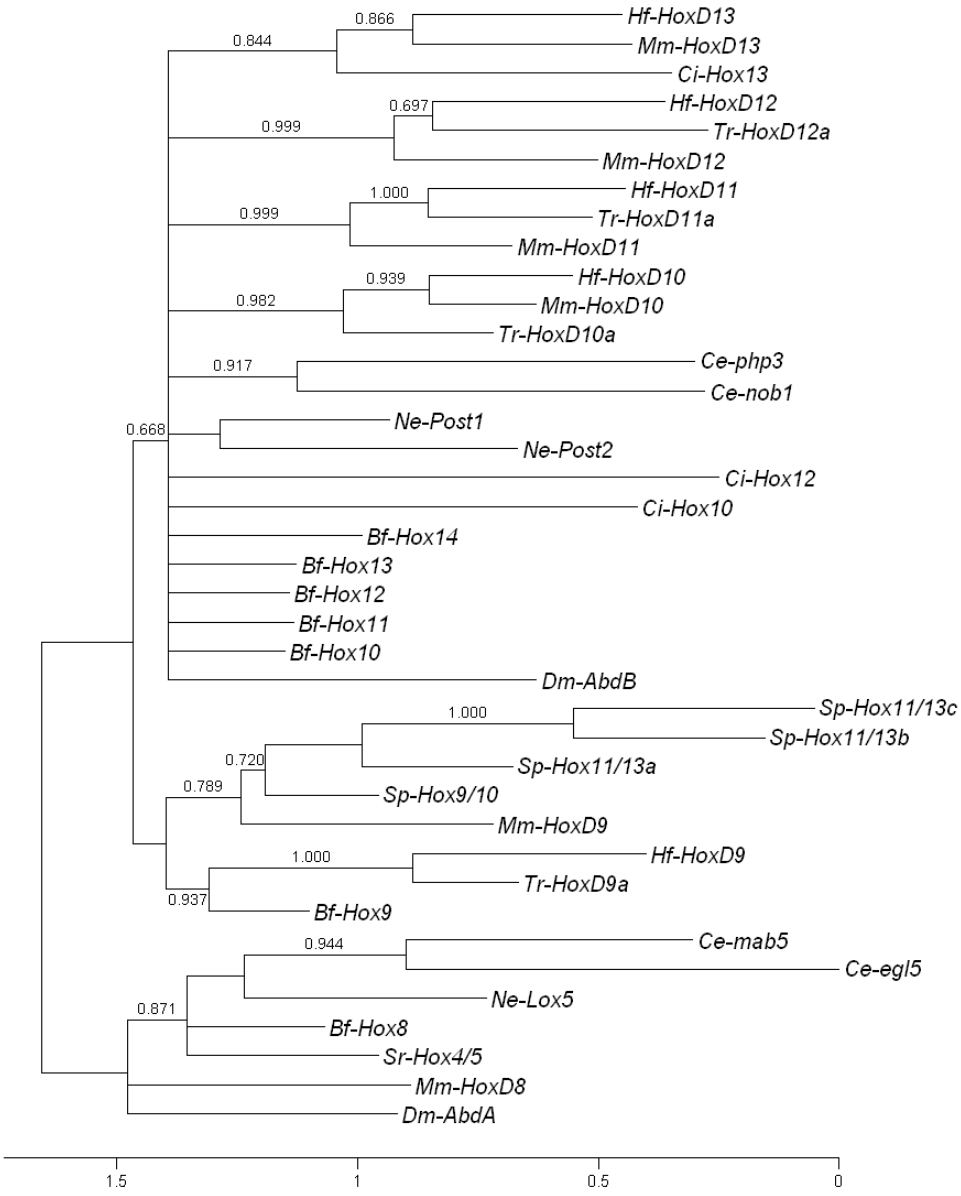

Figure 4.20: Posterior phylogeny of a clade of 31 posterior hox genes found in nine Bilaterian taxa. The tree was reconstructed from an alignment of 1,866 nucleotide characters by TRUL. All splits with a posterior probability greater than 0.500 are shown and the posterior probabilities of splits greater than 0.666 are shown on the appropriate branch.

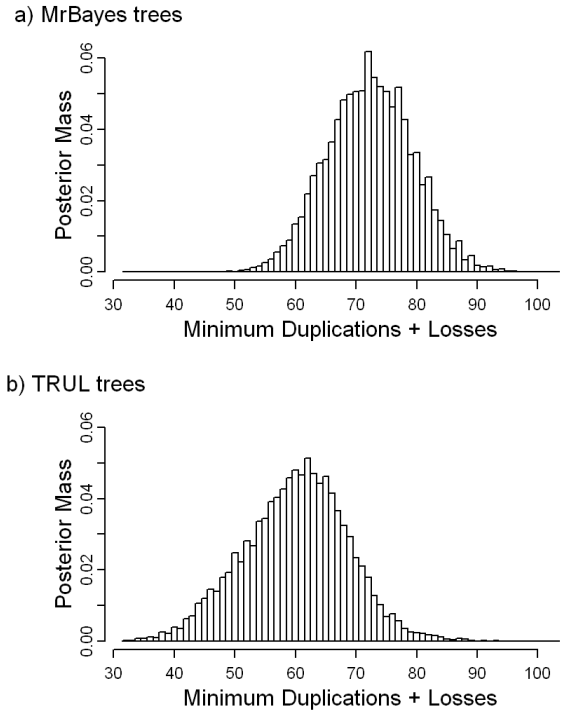

Figure 4.21: The distribution of the minimum number of gene duplications and losses among the posterior hox phylogenies sampled from MrBayes and TRUL. The number of required duplications and losses in the maximum parsimony reconciliation were calculated for each tree sampled from stationarity. The plots show the posterior mass of each value for the trees sampled from a) MrBayes and b) TRUL.

rate estimates (Figure 4.22). There is strong support for every branch having a negative $\vec{r}$. For the Basal Deuterostomia branches and especially the Chordata branches there is a high posterior probability that the duplication rate is less than 1 duplication/Lineage/Billion years, implying that there were very few duplications on these branches. The posterior distributions of all rates are well defined and are very distinct from the prior distribution.

Three rate assignments stand out as better supported than all the rest. In those assignments $\vec{\lambda}$ is the same on the Basal Deuterostomia branches and the Chordata branches and is the same on the Strongylocentrotus branch and the Protostomia branches, but differs between those two sets of branch sets. Similarly $\vec{\mu}$ is the same on the Strongylocentrotus branch and the Protostomia branches, but is different on the Chordata branches, while the assignment of $\vec{\mu}$ to the Basal Deuterostomia branches differs between the three rate assignments. These three assignments are strongly supported when compared to all other assignments ($BF_{10}(A_{Bil} = 1212_{\lambda}/N121_{\mu}, A_{Bil} \neq 1212_{\lambda}/N121_{\mu}) = 1.188$).

Comparison of birth-death rates on pairs of branch sets indicates that there is good

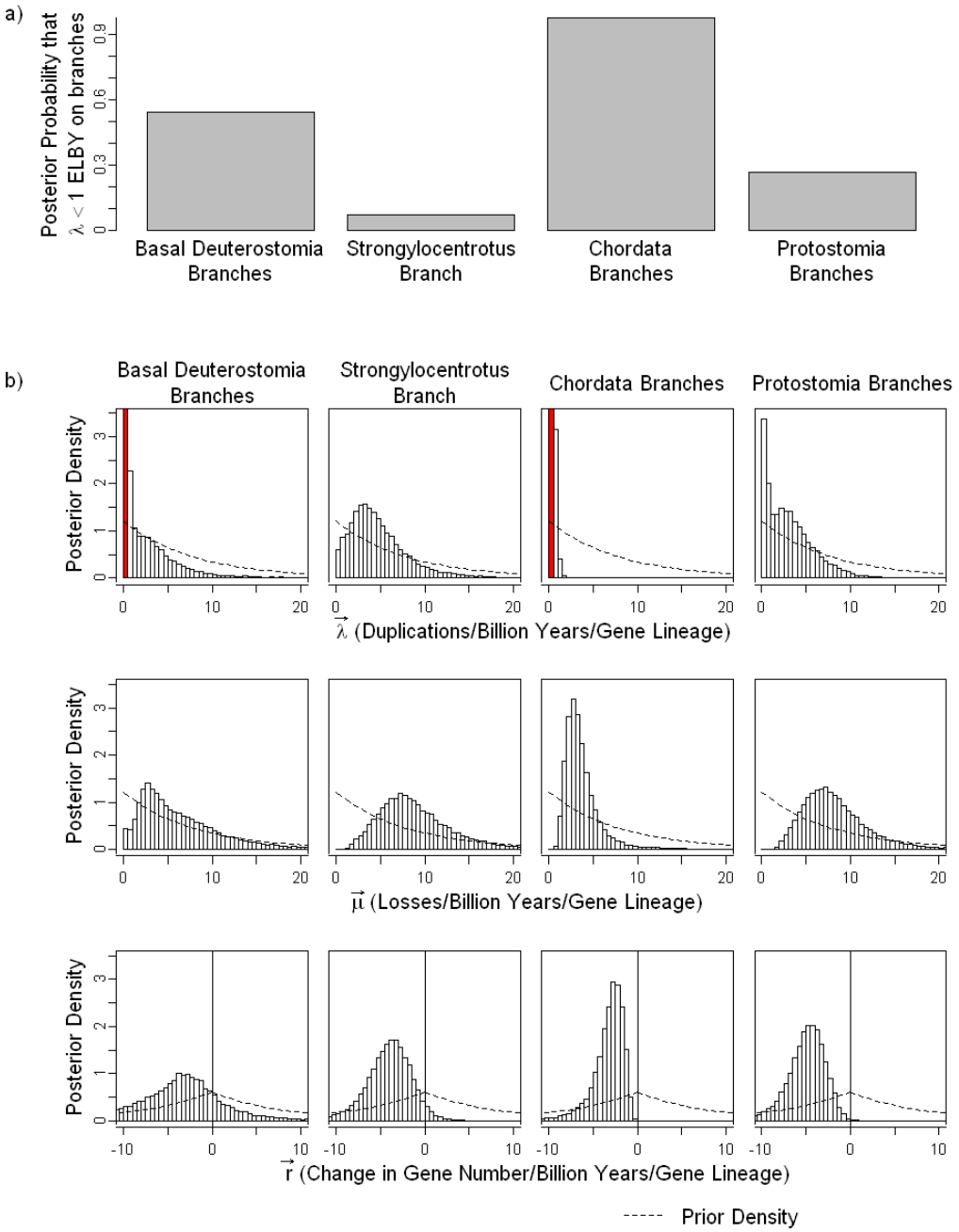

Figure 4.22: Posterior distribution of birth-death rates for the evolution of the posterior hox gene tree. a) The posterior probability that each set of branches had a gene duplication rate less than 1 duplication/lineage/billion years (ELBY). b) Samples from the MCMC for $\vec{\lambda}$, $\vec{\mu}$ and $\vec{r}$ for each set of branches of T_{Bil} are grouped into bins of 0.5 Events/Gene Lineage/Billion Years. The prior distribution is shown as a dashed line for comparison. Bins that exceed the range of the plot are marked red and are described in (a).

Table 4.8: Log_{10} Bayes Factor for the hypothethis that pairs of birth-death rates for the posterior hox gene tree differ between pairs of taxon tree branch sets. Above the diagonal are tests for the gene duplication rate differing between the rows and the columns ($BF_{10}(\vec{\lambda}_1 \neq \vec{\lambda}_2, \vec{\lambda}_1 = \vec{\lambda}_2)$). Below the diagonal are tests for the gene loss rate differing between the rows and the columns ($BF_{10}(\vec{\mu}_1 \neq \vec{\mu}_2, \vec{\mu}_1 = \vec{\mu}_2)$).

		Branch set$_2$			
		Basal Deuterostomia	Strongylocentrotus	Chordata	Protostomia
Branch set$_1$	Basal Deuterostomia	-	0.24954	-0.07484	0.10898
	Strongylocentrotus	0.09711	-	0.84423	-0.08662
	Chordata	0.23759	0.85721	-	0.3244
	Protostomia	0.08528	-0.32637	0.93645	-

support for a difference in $\vec{\lambda}$ between the Chordata branches and the Strongylocentrotus branch and a difference in $\vec{\mu}$ between the Chordata branches and both the Strongylocentrotus branch and the Protostomia branches (Table 4.8). We see that both $\vec{\lambda}$ and $\vec{\mu}$ are lower on the Chordata branches than on either the Strongylocentrotus branch or the Protostomia branches (Figure 4.22). The other comparisons imply that the Protostomia branches and the Strongylocentrotus branch have similar birth-death parameters with a higher turn over rate than the Basal Deuterostomia branches or the Chordata branches. Furthermore, although the different branch sets do differ in both $\vec{\lambda}$ and $\vec{\mu}$, the differences in $\vec{r}$ tend to be much smaller with mean values of 3 to 4 gene losses/ lineage/billion years for all branches (Figure 4.22).

4.6 Discussion

In this paper I introduced a Bayesian method to infer changes in the patterns of gain and loss in a gene family during the evolution of a group of organisms by comparing the phylogeny of those organisms to the phylogeny of the gene family. I used simulations to show that this method could detect differences in the birth-death process between two branches and infer when those differences were from differences in the duplication rate or the loss rate, so long as the true gene tree was known. This method had much more power to infer differences than a model based on gene counts alone. I also showed how to incorporate uncertainty in the gene tree reconstruction by including a search among gene tree topologies based on the gene sequences in my MCMC, and used that method to analyze the evolutionary history of two real gene families.

This is the first likelihood method to detect changes in the process of gene family diversification using the gene family phylogeny. In order to calculate the probability of a gene tree given a taxon tree and a set of birth-death parameters, I used a slightly modified

version of the model introduced by Arvestad et al. (2003, 2009). Although these authors had used this model to infer the birth-death parameters, they did not allow those parameters to vary between the branches of the taxon tree or attempt to compare different rate assignments to the branches of that tree. By using a reversible-jump MCMC, I was able to incorporate uncertainty in the reconciliation, in the gene tree phylogeny and in the parameter values as well as describe the confidence in my result numerically. Most interpretations of gene tree phylogenies are in effect parsimony methods, which assume a maximum parsimony reconciliation with low rates and a single gene tree, and can only describe alternative interpretations qualitatively.

When the true gene tree was known this model could detect a difference in the birth-death process between two branches of a taxon tree using Bayes factors, especially when the difference in parameters was large relative to the magnitude of the rates. A model based on gene counts alone was much less powerful. The gene tree model was in part more powerful than the gene count model because it could readily reconstruct the number of gene lineages at the root of the taxon tree. However, even when the gene count model was provided with the actual number of gene lineages at the root, it did not perform as well as the gene tree model.

Hahn et al. (2005) have used a similar gene count model to detect changes in the diversification of gene families on branches of a taxon tree. The small differences between our likelihood calculations and the rather large differences between how we evaluated those likelihoods, could mean that their method has more power to detect changes in the birth-death process than the gene count model I used here. However, the difference between the gene tree model and the gene count model was so large in my analysis, that I believe it is reasonable to conclude that this gene tree model would outperform any gene count model.

Analysis of simulations showed that the gene tree model also did a very good job of inferring the number of lineages at the root and the actual values of the birth-death parameters. The posterior means of the number of reconstructed lineages at he root of the taxon tree were almost always right on target. The gene count model also did a good job of estimating λ, although the between simulation variance was high and the posterior means tended to be low, when values of λ were very high. This model did not do nearly as well at estimating μ, and the prior appeared to have a much larger effect on the posterior distribution, implying that the interpretation of this data was not very powerful. Nevertheless, the gene tree model was able to make fair estimates of μ and did estimate larger values of μ for those simulations in which larger values of μ were used.

Most impressively, the gene tree model was capable of distinguishing whether differences in the gene diversification process were from differences in the duplication rate or the gene loss rate. Although this model did not always infer that two branches had different birth-death rates when they in fact did, it almost never inferred that two branches with the same λ had different λs or that two branches with the same μ had different μs. Even when two branches had different λs, it rarely inferred that they had different μs, and when they had different μs it rarely inferred that hey had different λs. The gene count model, on the

other hand, was completely incapable of distinguishing differences in the duplication rate from differences in the loss rate, even when it was provided with the actual number of gene lineages at he root of the taxon tree. Analyses of reconstructed taxon trees have also had a difficult time distinguishing changes in the birth rate from changes in the death rate (Nee et al. 1994a; Rabosky 2010). Comparisons of a tree to a tree in which it evolved have a unique ability to detect changes in the death rate that is not found in other methods based on the analysis of only extant lineages. When reconstructing the evolution of one phylogeny within another we can infer that certain lineages must have been lost, when a gene - or parasite or whatever - found in one taxon clade is not found in its sister clade. Thus we can be confidant that a lineage must have been lost, even without observing the distribution of lineages at some time in the past.

My analysis of a clade of protein tyrosine kinase genes found in *D. melanogaster*, *C. elegans* and *H. sapiens* concluded that the diversity of this clade was structured by a large increase in the gene duplication rate in the ancestors of *C. elegans*. Most if not all of the gene duplications appear to have occurred on the terminal branches of the taxon tree, and I generally estimated very low rates of gene loss throughout the history of this family. There was some uncertainty in the relationship among the clades of genes found in each taxon, which left open the possibility that several gene losses may be required (Figure 4.14). This was reflected in the posterior distribution of the number of reconstructed gene lineages at the root of the tree, which was greater than one more than half the time, and probably lead to an increase in the sampled values of gene loss rates.

In contrast my reconstruction of the posterior hox genes in nine bilaterian taxa inferred a history of massive gene loss. There was a lower turn over rate in the chordates and the two branches at the base of the deuterostomes than there was in the protostomes or the branch leading to *S. purpuratus*, but over all rates of gene number change were negative and approximately equal throughout the tree. In fact there was a large posterior probability of the gene duplication rate being essentially zero in the chordates and the Basal Deuterostomia branches. This is in stark contrast to a non-phylogenetic view of posterior hox evolution in which there are several gene duplications but very few losses and overall low turnover rates (Thomas-Chollier et al. 2010). However, my results were based on the large incongruence between the taxon phylogeny and the gene phylogeny, which was inferred from the sequences of the genes themselves. Previous results have shown a lack of resolution within the posterior hox genes (De Rosa et al. 1999; Kourakis and Martindale 2000), but my results are not a consequence of a lack of phylogenetically informative data, but on data that favors an incongruent gene tree. Using the gene tree evolution model as a prior for the the gene tree topology should improve the congruity of the gene tree, and the trees sampled from my analysis did have fewer minimum gene duplications and losses than those sampled from MrBayes, but even still they were highly incongruent. Nevertheless it is difficult to believe that the most recent common bilaterian ancestor had 15 posterior hox genes, which have been paired down to five or less in all extant bilaterian hox clusters, and there may be some misleading sites in the alignment. It is also possible that the bilaterian phylogeny used

here is incorrect. Phylogenetic analysis of whole genomes implies that Euchordata is not monophyletic and that in fact vertebrates are more closely related to *C. intestinalis* than to *B. floridae* (Nicholas et al. 2008). However, this is not consistent with my MrBayes posterior hox tree and would result in a more incongruent gene tree.

It would be possible to include uncertainty in the taxon tree as well as the gene tree in an MCMC. I would not recommend basing ones taxon phylogeny on the phylogeny of a single gene family with a complex evolutionary history. However, if one were to analyze multiple gene families at once, then you could have more confidence in the taxon phylogeny. Alternatively, one could put priors on different taxon topologies based on previous more intensive analyses. Another interesting possibility is to treat each cluster of genes within a genome, such as the hox cluster, as a separate terminal in the container tree. In other words use a hox cluster tree instead of a taxon tree. One could then reasonably judge different possible topologies for the cluster tree based on the gene family phylogeny, as the gene sequences within the hox family are the only basis for inferring the hox cluster phylogeny. I can imagine an analysis in which a gene tree is nested in a cluster tree, which is in turn nested within a taxon tree, although as a practical matter such an analysis seams daunting.

The methods described here would probably be improved by including a mechanism to relate the length of the branches of the gene tree inferred from the nucleotide substitution model to those inferred by the gene tree evolution model. Åkerborg et al. (2009) have used a relaxed clock to relate the amount of time between nodes of the gene tree under a given reconciliation to the amount of nucleotide change that occurred between those nodes. Here I assumed that the rate of nucleotide change was independent of the amount of time that has passed, but that is unreasonable and doubtlessly throws out a great deal of information. We would expect more nucleotide changes to have accumulated since older gene duplications than since more recent ones. Therefore the amount of nucleotide change tells us something about when a given gene lineage split occurred, and which reconciliation is correct, so that the probability of a given reconciliation would be affected not just by the model of gene tree evolution but also by the model of gene sequence evolution. Under different reconciliations, gene duplications and losses occur on different branches of the taxon tree, and so they would imply different assignments of birth-death rates to the branches of the taxon tree.

Here I used the Kimura two-parameter model (Kimura 1980) to calculate the probability of the gene sequences. This is the second simplest likelihood model available for nucleotide changes. It is common to use models in which more than two rates describe the probability of change between the four different nucleotides (see Felsenstein 2004, Chapter 13). The most complicated of these models that is commonly used is the general time reversible model with 10 parameters, but since our gene tree is inherently rooted, we could use a non-reversible model with up to 12 rates. It is also common to use models in which rates of change vary between the different sites such as Γ-distributed rates model (Yang 1994). It has been shown that tree reconstruction is made more inaccurate by use of the incorrect gene evolution model, especially when that model is too simple (Huelsenbeck and Rannala 2004), so incorporation of more complicated nucleotide evolution models would likely improve the

performance of this method. It may also be beneficial to incorporate other types of characters into the analysis, such as features of the secondary structure of the proteins for which the gene encodes, as these characters may be less prone to homoplasy than are individual sites in a gene sequence, and thus provide more information about the gene tree topology.

The dynamics of gene duplication and loss are actually quite complicated. Genes can be duplicated when a genetic locus is duplicated within the genome either as a tandem duplication or as a part of a larger chromosomal rearrangement, or reverse transcribed RNAs can be reinserted into the genome (Betran and Long 2002). In the former case a gene is likely to be copied along with its transcriptional machinery and so be able to produce a transcript and in turn be selected on, while in the latter case the gene sequence will be randomly inserted into the genome and so is unlikely to be transcribed and instead will undergo rapid pseudogenization (Graur et al. 1989). Once a gene has been duplicated, it must become fixed in a population either through drift or a selective sweep. Gene losses may be a result of deletions of large pieces of chromosomes or a two step process involving pseudogenization followed by decay and ultimately loss of the actual gene sequence (Moran 2003). The birth-death process is obviously only a rough summary of these processes, and in fact the actual rate of gene gain and loss should depend on many factors including mutation rate, environmental conditions and population size. However, one rarely has all the data necessary to account for all these factors, and as an approximation, the birth-death process has many features that are relevant to any model of changes in gene family size. The basic null model in which the rates do not vary is critical for testing any hypothethis of varying rates. Furthermore, as opposed to other stochastic processes, under the birth-death process the rate of change in the size of a gene family is proportional to the size of the gene family, a condition that we would expect to be true under any of the above scenarios.

On the other hand, the model used here would not hold up if a polymorphisms were inherited between two nodes in the gene tree. Let us imagine a situation in which two alleles, A and B, at a given locus were inherited in both taxon lineages, X and Y, descended from a speciation and then allele A was duplicated in lineage X. If allele B became fixed in the the other duplicate loci in lineage X, then, no matter which allele were fixed in lineage Y, the gene tree would suggest that a duplication happened in a common ancestor of the two taxa with a gene loss in lineage Y. There are numerous other scenarios in which inheritance of polymorphic loci could confound a gene tree analysis. There is an immense literature on the coalescent process (see Liu et al. 2009; Degnan and Rosenberg 2009), a model used to account for and infer information from the inheritance of polymorphic loci. However, without independent data about population size and population structure, it would be very difficult to untangle the effects of the coalescent and gene duplication and loss. I basically assumed that the branch lengths of my taxon tree were so long that there was sufficient time for all loci to become fixed between the nodes of my gene tree. In this sense, the coalescent is a microevolutionary process and gene duplication and loss is a macroevolutionary process, as they operate at different temporal scales. Nevertheless, if a researcher were using this model to study closely related taxa or a loci, such as MHC, which is known to retain polymorphisms

for a long time (Garrigan and Hedrick 2003) then the coalescent may have an effect on their analysis.

Here I studied models in which the rate of gene gain and loss varied between the branches of a taxon tree for all members of a gene family simultaneously. However, it would also be biologically realistic to investigate a model in which the birth-death rates evolved on the gene tree, such that certain clades within a gene family would have different rates of gene duplication and loss than the other members of the same family in the same genome at the same time. Several methods have been developed to identify exceptionally diverse clades of taxa within a reconstructed taxon phylogeny (e.g. Magallon and Sanderson 2001; McConway and Sims 2004; Moore and Donoghue 2007). It would not be particularly difficult to modify these methods so that they would work on a reconstructed gene tree evolving in a taxon tree. Such a method would have to incorporate the gene tree evolution model described here, in order to deal with the simultaneous formation of orthologous nodes in every lineage of a gene family as a consequence of speciation.

Ideally this method would be used not to merely mine for gene families with phylogenetically variable histories of diversification or to confirm that a given gene family has in fact diversified at different rates at different times; the study of evolution would be better served if these methods were used to detect correlations between the process of gene diversification and other biological or ecological characters. Over the last two decades several methods have been developed to detect correlations between pairs of characters (Pagel 1994; Felsenstein 1985), or between characters and rates of taxon diversification (Maddison et al. 2007; Paradis 2005; FitzJohn 2010). The second group of methods attempt to detect correlations between a biological character and the process of diversification by comparing models in which the birth-death rates are dependent on the state of that biological character to models in which those rates vary independently of that character. We would expect that if the process of gene family diversification is in fact critical for evolution that a given gene family may expand or contract in separate taxon lineages under similar evolutionary conditions. Several studies have suggested a link between changes in gene family size and convergent evolutionary changes (see Demuth and Hahn 2009). For example multiple insect lineages have gained resistance to organophosphate pesticides in parallel by expanding certain esterase gene families (Mouches et al. 1986; Field et al. 1988; Vontas et al. 2000), and it has been suggested that multiple *Drosophila* species have decreased the size of their odorant receptor families in response to increased host plant specificity (McBride et al. 2007). It would not be difficult to expand the methods described here in such a way that the birth-death rates are dependent on the states of other characters, in order to detect correlations between gene family diversification and the evolution of other taxonomic characters.

This method is currently too slow for application on a genomic scale. When the true tree is known, a pair of runs takes a reasonable amount of time (10 to 20 minutes on a single PC for the simulation analyses in this paper). However, the burden of identifying the correct gene tree adds a significant amount of time to the analysis. The method described by Hahn et al. (2005) and implemented in CAFE (De Bie et al. 2006) is substantially faster,

but methods which use the gene tree are much more powerful and give a more accurate view of the process of gene evolution. Thus currently a good approach would be to use the method of Hahn et al. (2005) to scan a large number of gene families, and then use the gene tree method described here to get a more refined view of the gene families that have been identified as having a varying history of gene gain and loss. My model also works better for cases in which a hypothethis about a gene family's evolution has already been developed and needs to be tested. Furthermore, as computing power increases, it will be possible to apply these methods on a genomic scale.

The method described in this paper is the first likelihood method that can detect changes in the duplication and loss rates of genes in a gene family on a taxon phylogeny by using the full gene tree. I couched this method in terms of gene trees and species trees, but it could also be used to describe the relationships between other entities in which one entity evolves inside another, such as domains in genes, genes in chromosomes, cell types in organs or taxa, parasites or symbionts in hosts, and taxa in geographic areas. I used examples involving genes in animals, for which the rates of horizontal transfer are very small, but in many of these other systems, one must account for the possibility of horizontal as well as vertical transmission. Not only is the method I described here useful in its own right, it can also serve as the basis for a large number of additional methods that attempt to identify changes in the process of lineage gain and loss.

Chapter 5

Conclusion

So, what is there left to say?

I described the reconstructed time variable birth-death process in which the rates of lineage gain and loss can vary with time but not between lineages alive at the same time, and we assume that any observed lineage survived to the present. I calculated the distribution of the number of reconstructed lineages at any time, the time between lineage splitting events and all the lineage splitting events in a phylogenetic tree. All of these distributions can be calculated when conditioned on any set of assumptions about how many reconstructed lineages their are at any times. I also showed how to sample from the distribution of branching times in a phylogenetic tree under any time variable birth-death process.

I introduced the discrete time birth-death process, a time variable birth-death process in which time is broken down into several periods during which the birth-death parameters are constant but between which they may vary. This process can be used as a simple and efficient numerical solution to any time variable birth-death process. It is trivial to calculate the inverse of all values under the discrete time birth-death process. Furthermore, it can be used to analytically incorporate sampling and mass extinction into any time variable birth-death process.

I showed how to compare a real tree to any time variable birth-death process using both mathematical and visual methods. One can calculate the maximum likelihood of a set of branching times under any time variable birth-death model in order to deduce the best parameter values. The likelihood or Kolmogorov-Smirnov's D under the maximum likelihood parameters can then be used as statistics to compare the fit of a real data set to a time-variable birth-death model. One could also visually compare the number of reconstructed lineages through time or the time between waiting times to the distribution of those values under a time variable birth-death process. The comparison of waiting times is particularly informative, as it allows you to see both when exactly one's data diverges from a model and how exactly a model fits the data. These visualizations can also be used without a real tree in order to see how different time variable birth-death models and different parameters for those models would affect the shape of phylogenetic trees.

In order to detect whole genome duplications from comparative analysis of chromosome counts, I used a modified birth-death process, in which I used not only the normal birth-death parameters but also a stochastic rate of every lineage duplicating at once. I used the Akaike Information Criterion to compare the maximum likelihoods of a model in which the genome doubling rate was zero to one in which that rate was free to vary. Once I had concluded that the genome duplication model was in fact a better fit, I calculated the posterior probability of a genome duplication on each branch of a taxon tree using the maximum likelihood parameters. I used this model to analyze chromosome counts in 125 molluscan taxa and concluded that there had been three paleopolyploidies: one near the base of the cephalopods; one near the base of the Stylommatophora; and a third at the base of a clade containing both the Capulidae and the neogastropods.

I used a likelihood method to compare a gene phylogeny to a phylogeny of the taxa in which those genes were found in order to detect changes in the process of gene duplication and loss in the history of a clade of genes. The gene evolution model can calculate the probability of a gene tree given a taxon tree and a set of birth-death parameters on the branches of the taxon tree. I implemented this model with a reversible-jump Markov chain Monte Carlo method in order to estimate the joint posterior distribution of the birth-death parameters and different assignments of those parameters to the branches of a taxon tree, given a taxon tree and a gene tree.

I wanted to show that this method had more power to detect changes in the process of gene duplication and loss than one which relied only on gene counts in the terminal taxa. So, I simulated 100 gene trees on a two taxon tree using a wide range of birth-death parameters. I then used a reversible-jump MCMC to analyze those trees both with the gene tree evolution model and a model in which gene counts evolved on the branches of the taxon tree. Bayes factors calculated for both models showed that the gene tree evolution model did a better job than the gene count model of detecting differences in the birth-death parameters between the branches of the taxon tree and in distinguishing whether those differences were in the duplication rate or the loss rate. The gene tree evolution model also did a better job of estimating parameter values.

In order to examine the evolution of real gene trees, it is necessary to account for the uncertainty in the gene phylogeny reconstruction. Therefore, I expanded my model by calculating the probability of the gene sequences and the gene tree, given the taxon tree, a set of birth death parameters and a set of nucleotide evolution parameters, by using the gene tree probability as a prior for the gene tree topology in a standard model of gene sequence evolution. I then used a reversible-jump MCMC to estimate the joint posterior distribution of all those parameters given a gene sequence alignment and a taxon tree. I used this model to analyze a clade of protein tyrosine kinase genes found in three metazoan taxa and all the posterior hox genes found in nine metazoan taxa. The protein tyrosine kinase diversification appeared to be characterized by a very low gene loss rate over all and by an increase in the gene duplication rate on the branch leading to *C. elegans*. On the other hand the posterior hox genes appeared to have had a great deal of gene loss in all lineages since the most recent

common ancestor of the Bilateria, with higher turn over rates in the protostomes and the echinoderms than in the chordates or at the base of the deuterostomes.

So, I described three different methods for detecting changes in the rate of lineage formation and lineage loss of biological entities under a number of different circumstances. I described a diverse set of results for inferring changes through time in macroevolutionary processes using reconstructed branching times in taxon phylogenies. I showed how to distinguish whole genome duplications from the normal process of chromosome duplication and loss by comparing chromosome counts on a phylogeny. Finally, I developed a method to infer changes in the rate of gene duplication and loss on the branches of a taxon tree by comparing a gene phylogeny to a taxon phylogeny. All three methods apply to active areas of biological research and continue a trend of using the birth-death process to analyze the process of biological lineage diversification.

Bibliography

L. Abi-Rached, A. Gilles, T. Shiina, P. Pontarotti, and H. Inoko. Evidence of en bloc duplication in vertebrate genomes. *Nature Genetics*, 31(1):100–105, 2002.

P. M. Agapow and A. Purvis. Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51(6):866–872, 2002.

A. Aguinaldo, J. Turbeville, L. Linford, M. Rivera, J. Garey, R. Raff, and J. Lake. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632): 489–493, 1997.

M. AHMED. Chromosome cytology of marine pelecypod molluscs. *Journal of Science, Karachi*, 4:77–94, 1976.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Ö. Åkerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–5719, 2009.

D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Advances in Applied Probability*, 37(4):1094–1115, 2005.

A. P. Anisimov, N. P. Tokmakova, and O. S. Poveshchenko. Somatic polyploidy in some tissues of the succineid snail *Succinea lauta* (Gastropoda: Pulmonata). *Tsitologiya*, 37(4): 311–330, 1995 1995.

T. Annilo, Z. Chen, S. Shulenin, J. Costantino, L. Thomas, H. Lou, S. Stefanov, and M. Dean. Evolution of the vertebrate ABC gene family: analysis of gene birth and death. *Genomics*, 88(1):1–11, 2006.

W. J. Ansorge. Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4): 195–203, 2009.

L. Arvestad, A. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19(Suppl. 1):i7–i15, 2003.

L. Arvestad, A. Berglund, J. Lagergren, and B. Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the eighth annual international conference on Resaerch in Computational Molecular Biology*, pages 326–335. ACM, 2004.

L. Arvestad, J. Lagergren, and B. Sennblad. The gene evolution model and computing its associated probabilities. *Journal of the ACM*, 56(2):1–44, 2009.

N. T. J. Bailey. *The Elements of Stochastic Processes*. John Wiley & Sons, New York, 1964.

G. M. Barker. Gastropods on land: phylogeny, diversity and adaptive morphology. In G. M. Barker, editor, *Biology of Terrestrial Molluscs*, pages 1–146. CAB International, Wallingford, 2001.

J. Barsiene, G. Tapia, and D. Barsyte. Chromosomes of molluscs inhabiting some mountain springs of eastern Spain. *Journal of Molluscan Studies*, 62(4):539–543, 1996.

A. R. Beaumont and J. E. Fairbrother. Ploidy manipulation in molluscan shellfish: a review. *Journal of Shellfish Research*, 10:1–18, 1991.

D. Benson, M. Boguski, D. Lipman, J. Ostell, and B. Ouellette. GenBank. *Nucleic Acids Research*, 26(1–7):1, 1998.

J. Berger. *Statistical Decision Theory: Foundations, Concepts, and Methods*. Springer Series in Statistics. Springer, New York, 1980.

E. Betran and M. Long. Expansion of genome coding regions by acquisition of new genes. *Genetica*, 115(1):65–80, 2002.

J. Blair and S. Hedges. Molecular phylogeny and divergence times of deuterostome animals. *Molecular Biology and Evolution*, 22(11):2275–2284, 2005.

G. Blanc and K. H. Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16(7):1667–1678, 2004.

M. G. B. Blum and O. Francois. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, 2006.

F. Burbrink and R. Pyron. How does ecological opportunity influence rates of speciation, extinction, and morphological diversification in new world ratsnakes (tribe Lampropeltini)? *Evolution*, 64(4):934–943, 2010.

L. J. M. Butot and B. Kiauta. Cytotaxonomic observation in the stylommatophoran family Helicidae, with considerations on the affinities within the family. *Malacologia*, 9:261–262, 1969.

K. P. Byrne and G. Blanc. Computational analyses of ancient polyploidy. *Current Bioinformatics*, 1(2):131–146, 2006.

R. Cameron, L. Rowen, R. Nesbitt, S. Bloom, J. Rast, K. Berney, C. Arenas-Mena, P. Martinez, S. Lucas, P. Richardson, et al. Unusual gene order and organization of the sea urchin hox cluster. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306(1):45–58, 2006.

S. Chambers. Rates of evolution in chromosome numbers in snails and vertebrates. *Evolution*, 41(1):166–175, 1987.

J. A. Chapman, E. Edsinger-Gonzales, E. Begovic, D. R. Lindberg, and D. S. Rokhsar. The genome of the owl limpet: A model for molluscs? In *International Plant & Animal Genomes Conference*, 2007.

O. Cohen and T. Pupko. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Molecular Biology and Evolution*, 27(3):703–713, 2010.

D. J. Colgan, W. F. Ponder, E. Beacham, and J. Macaranas. Molecular phylogenetics of Caenogastropoda (Gastropoda: Mollusca). *Molecular Phylogenetics and Evolution*, 42(3): 717–737, 2007.

T. M. Collins, K. Frazer, A. R. Palmer, G. J. Vermeij, and W. M. Brown. Evolutionary history of northern hemisphere Nucella (Gastropoda, Muricidae): Molecular, morphological, ecological, and paleontological evidence. *Evolution*, 50(6):2287–2304, 1996.

J. Cotton and R. Page. Rates and patterns of gene duplication and loss in the human genome. *Proceedings of the Royal Society B: Biological Sciences*, 272(1560):277–283, 2005.

J. Cotton and R. Page. The shape of human gene family phylogenies. *BMC Evolutionary Biology*, 6(1):66, 2006.

M. Crisp and L. Cook. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. *Evolution*, 63(9):2257–2265, 2009.

M. Csűrös and I. Miklós. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. In A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, and M. Waterman, editors, *Research in Computational Molecular Biology*, pages 206–220. Springer Berlin, Heidelberg, 2006.

N. Cusimano and S. Renner. Slowdowns in diversification rates from real phylogenies may not be real. *Systematic Biology*, 59(4):458–464, 2010.

T. De Bie, N. Cristianini, J. Demuth, and M. Hahn. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, 22(10):1269–1271, 2006.

R. De Rosa, J. Grenier, T. Andreeva, C. Cook, A. Adoutte, M. Akam, S. Carroll, and G. Balavoine. Hox genes in brachiopods and priapulids and protostome evolution. *Nature*, 399(6738):772–776, 1999.

J. Degnan and N. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340, 2009.

J. Demuth and M. Hahn. The life and death of gene families. *Bioessays*, 31(1):29–39, 2009.

R. Diaz-Uriarte and T. Garland. Effects of branch length errors on the performance of phylogenetically independent contrasts. *Systematic Biology,*, 47(4):654–672, 1998.

G. Dolman and A. Hugall. Combined mitochondrial and nuclear data enhance resolution of a rapid radiation of Australian rainbow skinks (Scincidae: Carlia). *Molecular Phylogenetics and Evolution*, 49(3):782–794, 2008.

A. Drummond, G. Nicholls, A. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, 2002.

A. J. Drummond, S. Y. W. Ho, M. J. Phillis, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLOS Biology*, 4(5):e88, 2006.

D. Eernisse and K. Peterson. The history of animals. In J. Cracraft and M. Donoghue, editors, *Assembling the Tree of Life*, pages 197–208. Oxford University Press, New York, 2004.

A. Egan and K. Crandall. Divergence and diversification in North American Psoraleeae (Fabaceae) due to climate change. *BMC Biology*, 6(1):55, 2008.

A. M. Ellison. Bayesian inference in ecology. *Ecology Letters*, 7(6):509–520, 2004.

W. Feller. Die grundlagen der Volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5(1):11–40, 1939.

J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249, 1973.

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

J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125(1):1–15, 1985.

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

D. Ferrier and P. Holland. Ancient origin of the Hox gene cluster. *Nature Reviews Genetics*, 2(1):33–38, 2001.

L. Field, A. Devonshire, and B. Forde. Molecular evidence that insecticide resistance in peach-potato aphids (*Myzus persicae* Sulz.) results from amplification of an esterase gene. *Biochemical journal*, 251(1):309, 1988.

W. Fitch. Distinguishing homologous from analogous proteins. *Systematic Biology*, 19(2): 99–113, 1970.

R. G. FitzJohn. Quantitative traits and diversification. *Systematic Biology*, 59(6):619–633, 2010.

M. Foote. Discordance and concordance between morphological and taxonomic diversity. *Paleobiology*, 19(2):185–204, 1993.

M. Foote, J. P. Hunter, C. M. Janis, and J. J. Sepkoski Jr. Evolutionary and preservational constraints on origins of biologic groups: Divergence times of eutherian mammals. *Science*, 283(5406):1310–1314, 1999.

R. Freckleton. The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, 22(7):1367–1375, 2009.

D. Garrigan and P. Hedrick. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution*, 57(8):1707–1722, 2003.

T. Gernhard. The conditioned reconstructed process. *Journal of Theoretical Biology*, 253 (4):769–778, 2008a.

T. Gernhard. New analytic results for speciation times in neutral models. *Bulletin of Mathematical Biology*, 70(4):1082–1097, 2008b.

J. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, USA, 1994.

G. Giribet and W. Wheeler. On bivalve phylogeny: a high-level analysis of the Bivalvia (Mollusca) based on combined morphology and DNA sequence data. *Invertebrate Biology*, 121(4):271–324, 2002.

M. A. Goldman, P. T. Loverde, and C. L. Chrisman. Hybrid origin of poly ploidy in fresh water snails of the genus *Bulinus* Mollusca Planorbidae. *Evolution*, 37(3):592–600, 1983.

M. Goodman, J. Czelusniak, G. Moore, A. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28(2):132–163, 1979.

D. Graur, Y. Shuali, and W. Li. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *Journal of Molecular Evolution*, 28(4):279–285, 1989.

P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

X. Gu and H. Zhang. Genome phylogenetic analysis based on extended gene contents. *Molecular Biology and Evolution*, 21(7):1401–1408, 2004.

C. Guyer and J. Slowinski. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution*, 45(2):340–350, 1991.

M. Hahn, T. De Bie, J. Stajich, C. Nguyen, and N. Cristianini. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15(8): 1153–1160, 2005.

K. Halanych. The new view of animal phylogeny. *Annual Review of Ecology, Evolution, and Systematics*, 35:229–256, 2004.

J. B. S. Haldane. *The Causes of Evolution*. Cornell Univ. Press, Ithaca, NY, 1932.

L. Harmon, J. Schulte, A. Larson, and J. Losos. Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, 301(5635):961–964, 2003.

L. Harmon, J. Weir, C. Brock, R. Glor, and W. Challenger. GEIGER: Investigating evolutionary radiations. *Bioinformatics*, 24(1):129–131, 2008.

K. Hartmann, D. Wong, and T. Stadler. Sampling trees from evolutionary models. *Systematic Biology*, 59(4):465–476, 2010.

G. Haszprunar. Is the Aplacophora monophyletic? A cladistic point of view. *American Malacological Bulletin*, 15:115–130, 2000.

S. Heard. Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, 50(6):2141–2148, 1996.

R. Highton and A. Larson. The genetic relationships of the salamanders of the genus Plethodon. *Systematic Biology*, 28(4):579–599, 1979.

D. Hillis, C. Moritz, B. Mable, and R. Olmstead. *Molecular Systematics*. Sinauer Associates, Sunderland, MA, 1996.

P. Holland. Problems and paradigms: Hoemeobox genes in vertebrate evolution. *BioEssays*, 14(4):267–273, 1992.

P. W. Holland, J. Garcia-FernÃ ndez, N. A. Williams, and A. Sidow. Gene duplications and the origins of vertebrate development. *Development*, Supplement:125–133, 1994.

J. Huelsenbeck and B. Rannala. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53 (6):904–913, 2004.

J. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

J. Huelsenbeck, B. Larget, R. Miller, and F. Ronquist. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51(5):673–688, 2002.

J. Huelsenbeck, B. Larget, and M. Alfaro. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, 21(6):1123–1133, 2004.

A. Hugall and M. Lee. Molecular claims of Gondwanan age for Australian agamid lizards are untenable. *Molecular Biology and Evolution*, 21(11):2102–2110, 2004.

T. Hunter and J. Cooper. Protein-tyrosine kinases. *Annual Review of Biochemistry*, 54(1): 897–930, 1985.

S. Q. Irvine, J. L. Carr, W. J. Bailey, K. Kawasaki, N. Shimizu, C. T. Amemiya, and F. H. Ruddle. Genomic analysis of Hox clusters in the sea lamprey *Petromyzon marinus*. *Journal of Experimental Zoology*, 294(1):47–62, 2002.

W. Iwasaki and T. Takagi. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics*, 23(13):i230–i239, 2007.

J. Izpisua-Belmonte, H. Falkenstein, P. Dollé, A. Renucci, and D. Duboule. Murine genes related to the Drosophila AbdB homeotic genes are sequentially expressed during development of the posterior part of the body. *The EMBO Journal*, 10(8):2279–2289, 1991.

J. Kadereit, E. Griebeler, and H. Comes. Quaternary diversification in European alpine plants: pattern and process. *Philosophical Transactions of the Royal Society of London-Series B: Biological Sciences*, 359(1442):265–274, 2004.

G. Karev, Y. Wolf, A. Rzhetsky, F. Berezovskaya, and E. Koonin. Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evolutionary Biology*, 2(1):18, 2002.

G. Karev, Y. Wolf, and E. Koonin. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19(15):1889–1900, 2003.

G. Karev, Y. Wolf, F. Berezovskaya, and E. Koonin. Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology*, 4(1):32, 2004.

M. Kasahara. The 2R hypothesis: an update. *Current Opinion in Immunology*, 19(5): 547–552, OCT 2007 2007.

R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995.

D. G. Kendall. On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15, 1948.

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

H. Kishino, J. Thorne, and W. Bruno. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, 18(3): 352–361, 2001.

M. Kourakis and M. Martindale. Combined-method phylogenetic analysis of Hox and ParaHox genes of the metazoa. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 288(2):175–191, 2000.

D. Krylov, Y. Wolf, I. Rogozin, and E. Koonin. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, 13(10):2229–2235, 2003.

C. Lakner, P. Van Der Mark, J. Huelsenbeck, B. Larget, and F. Ronquist. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic Biology*, 57(1):86–103, 2008.

M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

G. Le Pennec, A. Marhic, J. C. Martinez, D. Moraga, J. Normand, C. Tartu, and M. L. Pennec. Polyploidisation and gametogenesis in a cultivated bivalve mollusc, Ostreid *Crassostrea gigas*. *Bollettino Malacologico*, 43(1-8):87–95, 2007.

O. Lespinet, Y. Wolf, E. Koonin, and L. Aravind. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research*, 12(7):1048–1059, 2002.

A. R. Lindgren, G. Giribet, and M. K. Nishiguchi. A combined approach to the phylogeny of the Cephalopoda (Mollusca). *Cladistics*, 20(5):454–486, 2004.

W. Link and R. Barker. Model weights and the foundations of multimodel inference. *Ecology*, 87(10):2626–2635, 2006.

L. Liu, L. Yu, L. Kubatko, D. Pearl, and S. Edwards. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, 53(1):320–328, 2009.

L. G. Lundin. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, 16(1):1–19, 1993.

M. Lynch and J. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

M. Lynch and J. Conery. The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, 3(1):35–44, 2003.

W. P. Maddison and D. R. Maddison. Mesquite: A modular system for evolutionary analysis. version 2.0, 2007. URL `http://mesquiteproject.org`.

W. P. Maddison, P. E. Midford, and S. P. Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701–710, 2007.

S. Magallon and M. J. Sanderson. Absolute diversification rates in angiosperm clades. *Evolution*, 55(9):1762–1780, 2001.

G. Manning, D. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko. Comparison of site-specific rate-inference methods for protein sequences: Empirical bayesian methods are superior. *Molecular Biology and Evolution*, 21(9):1781–1791, 2004.

I. Mayrose, M. Barker, and S. Otto. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Systematic Biology*, 59(2):132–144, 2010.

C. McBride, J. Arguello, and B. O'Meara. Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics*, 177(3):1395–1416, 2007.

K. J. McConway and H. J. Sims. A likelihood-based method for testing for nonstochastic variation of diversification rates in phylogenies. *Evolution*, 58(1):12–23, 2004.

A. McLysaght, K. Hokamp, and K. H. Wolfe. Extensive genomic duplication during early chordate evolution. *Nature Genetics*, 31(2):200–204, 2002.

B. R. Moore and M. J. Donoghue. Correlates of diversification in the plant clade Dipsacales: Geographic movement and evolutionary innovations. *American Naturalist*, 170(S2):S28–S55, 2007.

B. R. Moore, K. M. A. Chan, and M. J. Donoghue. Detecting diversifcation rate variation in supertrees. In O. Bininda-Emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 487–533. Kluwer Academic, Dordrecht, 2004.

N. Moran. Tracing the evolution of gene loss in obligate bacterial symbionts. *Current Opinion in Microbiology*, 6(5):512–518, 2003.

P. Mordan and C. Wade. Heterobranchia II. the Pulmonata. In W. F. Ponder and D. R. Lindberg, editors, *Phylogeny and Evolution of the Mollusca*, pages 409–426. University of California Press, Berkeley, 2008.

C. Mouches, N. Pasteur, J. Berge, O. Hyrien, M. Raymond, B. de Saint Vincent, M. de Silvestri, and G. Georghiou. Amplification of an esterase gene is responsible for insecticide resistance in a California Culex mosquito. *Science*, 233(4765):778–780, 1986.

H. K. Nakamura. A review of molluscan cytogenetic information based on the CISMOCH-Computerized Index System for molluscan chromosomes. Bivalvia, Polyplacophora and Cephalopoda. *Venus*, 44(3):193–225, 1985.

S. Nee. Inferring speciation rates from phylogenies. *Evolution*, 55(4):661–668, 2001.

S. Nee, A. Mooers, and P. Harvey. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America*, 89(17):8322–8326, 1992.

S. Nee, E. Holmes, R. May, and P. Harvey. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions: Biological Sciences*, 344(1307):77–82, 1994a.

S. Nee, R. May, and P. Harvey. The reconstructed evolutionary process. *Philosophical Transactions: Biological Sciences*, 344(1309):305–311, 1994b.

S. Nee, E. C. Holmes, R. M. May, and P. H. Harvey. Estimating extinction from molecular phylogenies. In J. H. Lawton and R. M. May, editors, *Extinction Rates*, chapter 11, pages 164–182. Oxford University Press, 1995.

M. Nei and A. Rooney. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics*, 39:121–152, 2005.

M. Nei, X. Gu, and T. Sitnikova. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15):7799–7806, 1997.

M. Nei, I. Rogozin, and H. Piontkivska. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):10866–10871, 2000.

H. Nicholas, D. Thomas Butts, F. Rebecca, T. Uffe Hellsten, R. Marc, A. Eiichi Shoguchi Jr, E. Yu, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, 2008.

R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, 1998.

M. K. Nishiguchi and R. H. Mapes. Cephalopoda. In W. F. Ponder and D. R. Lindberg, editors, *Phylogeny and Evolution of the Mollusca*, pages 163–200. University of California Press, Berkeley, 2008.

A. S. Novozhilov, G. P. Karev, and E. V. Koonin. Biological applications of the theory of birth-and-death processes. *Briefings in Bioinformatics*, 7(1):70–85, 2006.

J. Nylander, J. Wilgenbusch, D. Warren, and D. Swofford. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, 24(4):581–583, 2008.

S. Ohno. *Sex Chromosomes and Sex-linked Genes*. Springler-Verlag, Berlin, 1967.

S. Ohno. *Evolution by Gene Duplication*. Springer-Verlag, New York, 1970.

A. Okusu, E. Schwabe, D. J. Eernisse, and G. Giribet. Towards a phylogeny of chitons (Mollusca, Polyplacophora) based on combined analysis of five molecular loci. *Organisms Diversity & Evolution*, 3(4):281–302, 2003.

B. O'Meara, C. Ané, M. Sanderson, and P. Wainwright. Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60(5):922–933, 2006.

T. Ota and M. Nei. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Molecular Biology and Evolution*, 11(3):469–482, 1994.

R. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43(1):58–77, 1994.

R. Page and M. Charleston. From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. *Molecular Phylogenetics and Evolution*, 7(2):231–240, 1997.

M. Pagel. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences*, 255(1342):37–45, 1994.

M. Pagel. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 48(3):612–622, 1999.

E. Paradis. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. *Proceedings of the Royal Society B: Biological Sciences*, 264(1385):1141–1147, 1997.

E. Paradis. Statistical analysis of diversification with species traits. *Evolution*, 59(1):1–12, 2005.

E. Paradis, J. Claude, and K. Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

G. M. Park. Polyploidy in three sphaeriids (Bivalvia : Veneroida) from Korea. *Molluscan Research*, 28(2):133–136, 2008.

C. M. Patterson. *Chromosomes of Molluscs*, volume 2 of *Proceedings of the 2nd Symposium of Mollusca*, pages 635–689. Marine Biological Association of India, Ernakulam, Cochin, India, 1969.

C. M. Patterson and J. B. Burch. Chromosomes of pulmonate molluscs. In V. Fretter and J. Peake, editors, *Systematics, Evolution and Ecology*, volume 2A of *Pulmonates*, pages 171–217. Academic Press, New York, 1978.

K. Peterson and N. Butterfield. Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9547–9552, 2005.

H. Piontkivska, A. Rooney, and M. Nei. Purifying selection and birth-and-death evolution in the histone H4 gene family. *Molecular Biology and Evolution*, 19(5):689–697, 2002.

J. Piskur. Origin of the duplicated regions in the yeast genomes. *Trends in Genetics*, 17(6): 302–303, 2001.

W. F. Ponder and D. R. Lindberg. Towards a phylogeny of gastropod mollusc: An analysis using morphological characters. *Zoological Journal of the Linnean Society*, 119(2):83–265, 1997.

D. Posada and K. Crandall. MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 14(9):817–818, 1998.

A. Purvis, S. Nee, and P. Harvey. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society B: Biological Sciences*, 260(1359):329–333, 1995.

O. Pybus and P. Harvey. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*, 267(1459):2267–2272, 2000.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`.

D. Rabosky. Likelihood methods for detecting temporal shifts in diversification rates. *Evolution*, 60(6):1152–1164, 2006a.

D. Rabosky. LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evolutionary Bioinformatics Online*, 2 (6):1152–1164, 2006b.

D. Rabosky. Extinction rates should not be estimated from molecular phylogenies. *Evolution*, 64(6):1816–1824, 2010.

D. Rabosky and I. Lovette. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution*, 62(8):1866–1875, 2008a.

D. Rabosky and I. Lovette. Density-dependent diversification in North American wood warblers. *Proceedings of the Royal Society B: Biological Sciences*, 275(1649):2363–2371, 2008b.

A. Rambaut. PhyloGen: Phylogenetic tree simulator package, 2002. URL `http://tree.bio.ed.ac.uk/software/phylogen/`.

A. Rambaut and L. Bromham. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution*, 15(4):442–448, 1998.

B. Rannala. Gene genealogy in a population of variable size. *Heredity*, 78(4):417–423, 1997.

B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution*, 43(3):304–311, 1996.

D. Raup, S. Gould, T. Schopf, and D. Simberloff. Stochastic models of phylogeny and the evolution of diversity. *The Journal of Geology*, 81(5):525–542, 1973.

R. H. Ree. Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. *Evolution*, 59(2):257–265, 2005.

W. Reed and B. Hughes. A model explaining the size distribution of gene and protein families. *Mathematical Biosciences*, 189(1):97–102, 2004.

F. Ronquist and J. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

G. Roth, J. Blanke, and D. B. Wake. Cell-size predicts morphological complexity in the brains of frogs and salamanders. *Proceedings of the National Academy of Sciences of the United States of America*, 91(11):4796–4800, 1994.

T. Rowe. Chordate phylogeny and development. In J. Cracraft and M. Donoghue, editors, *Assembling the Tree of Life*, pages 384–409. Oxford University Press, New York, 2004.

L. Rüber and R. Zardoya. Rapid cladogenesis in marine fishes revisited. *Evolution*, 59(5): 1119–1127, 2005.

F. Ruddle, J. Bartels, K. Bentley, C. Kappen, M. Murtha, and J. Pendleton. Evolution of Hox genes. *Annual Review of Genetics*, 28(1):423–442, 1994.

I. Ruiz-Trillo, M. Riutort, D. Littlewood, E. Herniou, and J. Baguñà. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science*, 283(5409): 1919–1923, 1999.

M. Sanderson. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, 2003.

M. Sanderson and G. Bharathan. Does cladistic information affect inferences about branching rates? *Systematic Biology*, 42(1):1–17, 1993.

V. Savolainen, S. Heard, M. Powell, T. Davies, and A. Mooers. Is cladogenesis heritable? *Systematic Biology*, 51(6):835–843, 2002.

M. Semon and K. H. Wolfe. Consequences of genome duplication. *Current Opinion in Genetics & Development*, 17(6):505–512, 2007.

B. Sennblad and J. Lagergren. Probabilistic orthology analysis. *Systematic Biology*, 58(4): 411–424, 2009.

A. Shaw, C. Cox, B. Goffinet, W. Buck, and S. Boles. Phylogenetic evidence of a rapid radiation of pleurocarpous mosses (Bryophyta). *Evolution*, 57(10):2226–2241, 2003.

H. Sims and K. McConway. Nonstochastic variation of species-level diversification rates within angiosperms. *Evolution*, 57(3):460–479, 2003.

M. Slatkin and R. Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562, 1991.

J. B. Slowinski and C. Guyer. Testing whether certain traits have caused amplified diversification - an improved method based on a model of random speciation and extinction. *American Naturalist*, 142(6):1019–1024, 1993.

A. Solem and E. L. Yochelson. North american paleozoic land snails, with a summary of other paleozoic nonmarine snails. *U S Geological Survey Professional Paper*, 1072:1–42, 1979.

J. Spring. Vertebrate evolution by interspecific hybridisation - are we polyploid? *FEBS Letters*, 400(1):2–8, 1997.

T. Stadler. Lineages-through-time plots of neutral models for speciation. *Mathematical Biosciences*, 216(2):163–171, 2008.

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

M. Steeman, M. Hebsgaard, R. Fordyce, S. Ho, D. Rabosky, R. Nielsen, C. Rahbek, H. Glenner, M. Sorensen, and E. Willerslev. Radiation of extant cetaceans driven by restructuring of the oceans. *Systematic Biology*, 58(6):573–585, 2009.

R. Strathmann and M. Slatkin. The improbability of animal phyla with few species. *Paleobiology*, 9(2):97–106, 1983.

D. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, 2002.

E. V. Tabakova, I. A. Kirsanova, and A. P. Anisimov. Morphological variability and ploidy of nuclei in neurons of the central nervous system of bivalves in connection with somatic polyploidy. *Biologiya Morya (Vladivostok)*, 31(5):352–357, 2005.

C. Tabin. Why we have (only) five fingers per hand: Hox genes and the evolution of paired limbs. *Development*, 116(2):289–296, 1992.

R. Tatusov, N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1):41, 2003.

J. S. Taylor and J. Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics*, 38(1):615–643, 2004.

C. Thiriot-Quievreux. Review of the literature on bivalve cytogenetics in the last ten years. *Cahiers De Biologie Marine*, 43(1):17–26, 2002.

C. Thiriot-Quievreux. Advances in chromosomal studies of gastropod molluscs. *Journal of Molluscan Studies*, 69(3):187–201, 2003.

M. Thomas-Chollier, V. Ledent, L. Leyns, and M. Vervoort. A non-tree-based comprehensive study of metazoan hox and parahox genes prompts new insights into their origin and evolution. *BMC Evolutionary Biology*, 10(1):73, 2010. URL `http://www.biomedcentral.com/1471-2148/10/73`.

E. Thompson. *Human Evolutionary Trees*. Cambridge University Press, Cambridge, 1975.

J. Thompson, D. Higgins, and T. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

N. P. Tokmakova, N. A. Galimulina, and A. P. Anisimov. Morphofunctional characterization and ploidy levels of cells of the bivalve digestive gland with special reference to somatic polyploidy. *Biologiya Morya (Vladivostok)*, 32(4):270–276, 2006.

J. Turgeon, R. Stoks, R. Thum, J. Brown, and M. McPeek. Simultaneous Quaternary radiations of three damselfly clades across the Holarctic. *American Naturalist*, 165(4): E78–E107, 2005.

L. Valente, V. Savolainen, and P. Vargas. Unparalleled rates of species diversification in Europe. *Proceedings of the Royal Society B: Biological Sciences*, 277(1687):1489–1496, 2010.

K. Vandepoele, W. D. Vos, J. S. Taylor, A. Meyer, and Y. V. de Peer. Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1638–1643, 2004.

A. E. Vinogradov. Larger genomes for molluskan land pioneers. *Genome*, 43(1):211–212, 2000.

J. Vontas, G. Small, and J. Hemingway. Comparison of esterase gene amplification, gene expression and esterase activity in insecticide susceptible and resistant strains of the brown planthopper, *Nilaparvata lugens* (Stål). *Insect Molecular Biology*, 9(6):655–660, 2000.

R. Waagepetersen and D. Sorensen. A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping. *International Statistical Review*, 69(1):49–61, 2001.

C. M. Wade, P. B. Morton, and F. Naggs. Evolutionary relationships among the pulmonate land snails and slugs (Pulmonata, Stylommatophora). *Biological Journal of the Linnean Society*, 87(4):593–610, 2006.

H. Wägele, Klussmann-Kolb, V. Vonnemann, and M. Medina. Heterobranchia I. the Opisthobrnachia. In W. F. Ponder and D. R. Lindberg, editors, *Phylogeny and Evolution of the Mollusca*, pages 385–408. University of California Press, Berkeley, 2008.

G. Wagner, C. Amemiya, and F. Ruddle. Hox cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25):14603–14606, 2003.

J. F. Wendel. Genome evolution in polyploids. *Plant Molecular Biology*, 42(1):225–249, 2000.

M. J. D. White. *Animal Cytology and Evolution*. Cambridge University Press, Cambridge, 1973.

K. Wolfe. Evolutionary genomics: Yeasts accelerate beyond BLAST. *Current Biology*, 14 (10):R392–R394, 2004.

K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature (London)*, 387(6634):708–713, 1997.

S. Wong, G. Butler, and K. H. Wolfe. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proceedings of the National Academy of Sciences of the United States of America*, 99(14):9272–9277, 2002.

I. G. Woods, C. Wilson, B. Friedlander, P. Chang, D. K. Reyes, R. Nix, P. D. Kelly, F. Chu, J. H. Postlethwait, and W. S. Talbot. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Research*, 15(9):1307–1314, 2005.

H. Yang and X. Guo. Polyploid induction by heat shock-induced meiosis and mitosis inhibition in the dwarf surfclam, *Mulinia lateralis* Say. *Aquaculture*, 252(2-4):171–182, 2006.

Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(306):314, 1994.

Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution*, 14(7):717–724, 1997.

G. Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.

A. Zilch. Euthyneura. In O. Schindewolf, editor, *Handbuch der Paläozoologie*, volume 6, chapter 2. Gebräder Bornträger, Berlin, 1959-1960.

Appendix A

Mollusk Chromosome Counts

Table A.1: Statistics describing the distribution of observed chromosome counts among members of the terminal molluscan clades used in this study.

		No of Species		Range	
Terminal Clade	Mode	Total	At Mode	Low	High
Polyplacophora					
Chitonidae	12	4	3	12	13
Ischnochiton	12	2	2	12	12
Lepidozona	12	3	3	12	12
Mopaliidae	12	4	3	12	16
Acanthochitonidae	8	7	4	8	12
Bivalvia					
Heterodonta					
Sphaeriidae	18	2	1	18	22
Corbiculidae	13	3	1	12	19
Veneroida	19	18	15	15	23
Myoida	17	1	1	17	17
Mactroidea	18	5	2	17	19
Solenoidea	19	4	4	19	19
Tellinoidea	19	5	5	19	19
Lasaeidae	20	2	1	18	20
Pholadoidea	17	3	2	17	19
Cardioidea	19	4	2	12	20
Palaeoheterodonta					
Unionoidea	19	26	24	10	19
Pteriomorpha					
Mytiloidea	14	31	14	11	16
Arcoida	19	10	7	14	19
Pterioidea	14	10	9	13	14
Ostreoidea	10	23	23	10	10
Pectinoidea	19	18	13	13	19
Limoidea	16	1	1	16	16
Anomioidea	7	3	1	6	13
"Protobranchs"					
Nuculanoidea	19	1	1	19	19
Solemyoidea	11	1	1	11	11
Nuculoidea	12	1	1	12	12
Cephalopoda					
Nautiloidea					
Nautilidae	21	2	2	21	21
Coleoidea					
Octopoda	30	6	3	28	30
Sepiidae	46	3	2	46	56
Loliginidae	46	4	3	46	84
Scaphopoda					
Dentaliidae	10	3	3	10	10

Table A.1: continued

Terminal Clade	Mode	No of Species Total	No of Species At Mode	Range Low	Range High
Gastropoda					
"Patellogastropoda"					
Lottiidae	10	11	11	10	10
Patelloididae	10	3	3	10	10
Patellidae	9	9	8	8	9
Vetigastropoda					
Trochidae	18	17	16	18	20
Turbinoidea	18	4	3	8	18
Haliotidae	18	15	9	14	18
Fissurellidae	16	6	3	13	16
Neritimorpha					
Neritidae	12	26	24	11	13
Helicinidae	18	3	3	18	18
"Caenogastropoda"					
"Architaenioglossa"					
Cyclophoroidea	14	7	7	14	14
Diplommatinidae	13	16	16	13	13
Ampullarioidea	14	9	7	13	14
Viviparoidea	9	27	8	7	32
"Littorinimorpha"					
Hydrobiidae	17	10	6	16	18
Rissoidae	16	1	1	16	16
Assimineidae	15	7	3	15	15
Bithyniidae	17	6	5	17	18
Vermetidae	17	1	1	17	17
Hipponicidae	17	1	1	17	17
Atlantidae	15	12	9	14	16
Carinariidae	16	2	1	16	17
Pterotracheidae	16	4	4	16	16
Littorinidae	17	5	4	8	17
Pomatiidae	13	2	2	13	13
Naticidae	17	3	3	17	17
Calyptraeidae	17	1	1	17	17
Strombidae	12	1	1	12	12
Capulidae	31	1	1	31	31
Cypraeidae	36	6	2	26	36
Ranellidae	35	1	1	35	35

Table A.1: continued

Terminal Clade	Mode	No of Species Total	No of Species At Mode	Range Low	Range High
Gastropoda					
"Caenogastropoda"					
Neogastropoda					
Nassariidae	32	6	3	32	34
Mitridae	19	1	1	19	19
Fasciolariidae	35	1	1	35	35
Melongenidae	28	1	1	28	28
Muricidae	35	13	9	18	36
Turbinellidae	33	1	1	33	33
Buccinidae	35	8	5	34	36
Columbellidae	34	1	1	34	34
Conoidea	35	3	1	17	36
Cerithioidea					
Turritellidae	16	1	1	16	16
Pleuroceridae	18	33	17	7	19
Thiaridae	18	7	3	16	36
Pomatiopsidae	17	1	1	17	17
Cerithiidae	18	14	10	7	18
Heterobranchia					
"Lower Heterobranchia"					
Valvatoidea	10	1	1	10	10
Pyramidellidae	17	1	1	17	17
Opisthobranchia					
Nudibranchia	13	56	51	12	16
Notaspidea	13	4	2	12	13
Sacoglossa	17	14	12	14	17
Thecosomata	10	8	4	10	17
Gymnosomata	16	4	4	16	16
Anaspidea	17	11	8	10	17
Cephalaspidea	17	19	11	13	18
"Basommatophora"					
Amphibolidae	18	1	1	18	18
Siphonariidae	16	8	8	16	16
Anclidae	18	11	3	15	60
Chilinidae	18	1	1	18	18
Latiidae	18	1	1	18	18
Planorbidae	18	56	46	18	72
Lymnaeidae	18	35	21	16	19
Physidae	18	7	7	18	18
Ellobiidae	18	10	7	17	18

Table A.1: continued

Terminal Clade	Mode	No of Species Total	No of Species At Mode	Range Low	Range High
Gastropoda					
Heterobranchia					
Systellommatophora					
Veronicellidae	17	3	2	16	17
Onchidiidae	18	3	2	17	18
Stylommatophora					
Succineidae	18	10	32	5	25
Helicidae	26	59	16	21	30
Bradybaenidae	29	31	16	28	30
Camaenidae	29	12	10	27	29
Polygyridae	29	21	16	26	31
Haplotrematidae	30	2	1	29	30
Cerionidae	27	1	1	27	27
Endodontidae	31	3	2	29	31
Bulimulidae	30	3	2	29	30
Arionidae	26	6	2	25	29
Philomycidae	24	2	2	24	24
Testacellidae	32	1	1	32	32
Megalobulimidae	31	3	3	31	31
Milacidae	33	4	3	33	34
Limacidae	30	9	4	20	31
Vitrinidae	28	1	1	28	28
Ariophantidae	27	6	2	25	32
Zonitidae	30	5	2	20	31
Trochomorphidae	28	2	1	28	30
Helicarionidae	28	2	1	24	28
Achatinidae	30	1	1	30	30
Subulinidae	31	2	1	25	31
Ferussaciidae	30	1	1	30	30
Clausiliidae	24	16	12	24	30
Chondrinidae	30	7	3	28	30
Cionellidae	26	2	2	26	26
Achatinellidae	20	4	2	20	23
Enidae	24	6	6	24	24
Valloniidae	28	2	2	28	28
Partulidae	29	1	1	29	29