Problems in Pedigrees and Phylogenies

Peter Chi

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Vladimir Minin, Chair

Li Hsu

Kenneth Rice

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Problems in Pedigrees and Phylogenies

Peter Chi

Chair of the Supervisory Committee:
Assistant Professor Vladimir Minin
Department of Statistics

The footprint of genetic inheritance can be observed among both familial relationships between individuals and evolutionary relationships between species. The former can be depicted by a pedigree, whereas the latter can be depicted by a phylogeny. These two distinct objects can be viewed as analogous, in the sense that they both describe a correlation structure on the units of observation. We discuss statistical methods for various problems within both settings. We begin with heritability estimation using pedigree data, with an ordinal outcome trait. We discuss the use of the threshold model for heritability estimation, exploring the consequences of model misspecification and sample size requirements under this model. Next, we move to phylogenetic methods for the detection of recombination, or the exchange of genetic material between species. We propose a new recombination detection statistic, in which false positive detection due to convergent evolution can be avoided while detecting recombination. Finally, we conclude with an improvement to least squares phylogenetic inference, by considering a new loss function that does not use standard evolutionary distances. Instead, we use a distance measure that considers the sequence data simultaneously with substitution model parameters, the topology, and branch lengths of the phylogenetic tree.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to acknowledge my parents, for instilling a drive for learning in me at an early age, and continuing to be supportive throughout my entire academic career. I am also very thankful for the first professor who sparked the passion for statistics in me, Dr. Edward Rothman at the University of Michigan. I was very fortunate to take his class during my last semester as an undergraduate there, and made the decision to pursue graduate work in statistics or biostatistics after completing his course.

I am very grateful for my entire cohort in the Biostatistics Department at UW. The sense of community that we had immediately upon starting our program was something that I'll never forget, and I truly could not have made it through our first couple of years without having each of you to lean on. I also very much appreciate my interactions with the rest of the students in the department too. In particular, I am thankful for the opportunity to help create and captain the Improper Priors, the Ultimate Frisbee team of the Statistics and Biostatistics Departments, which competed in both the UW intramural league and the city of Seattle league (DiscNW).

My academic advisor, Dr. Li Hsu, was of tremendous help and support throughout my time at UW. From helping me study for the qualifying exams, to serving on my dissertation committee, she went well above the call of duty for an academic advisor and I will be forever grateful for her. I also thank Dr. Kenneth Rice, for challenging me and giving me insightful comments on my dissertation as he served on my reading committee. Dr. Lurdes Inoue was also a remarkable figure for me at UW, as we had a wonderful meeting when I first visited the department as a prospective student, and

then I later served as her TA, and she served on my dissertation committee. I am very appreciative both for her helpful comments on my dissertation, and the opportunity to learn from her as an instructor. I am also grateful for each of the other professors for whom I served as a TA, including Drs. Norman Breslow, Jim Hughes, and Daniela Witten. One of the best learning experiences for me during my time at UW was the opportunity to be a substitute lecturer, and I greatly appreciate this opportunity given to me by Drs. Lurdes Inoue, Daniela Witten, and Scott Emerson.

Dr. Adam Leaché taught me in the Applied Phylogenetics class, and I thank him for his inspiration with this subject, and for serving as the GSR on my dissertation committee. I also thank Dr. Joseph Felsenstein, for teaching numerous courses in the Genome Sciences Department that I was fortunate enough to take, and for organizing the weekly Population Genetics seminar. I am particularly appreciative of the conversation that I had with him while I was seeking a dissertation advisor. Specifically, he highly recommended Dr. Vladimir Minin, of the Statistics Department.

I simply could not have asked for a better advisor than Vladimir. While it was of course fantastic to work with such a sharp mind, I am even more grateful for his enthusiasm and encouraging nature. This applied both to my dissertation work, and my longer-term career aspirations. When I told him that I had a passion for teaching and wanted to become faculty at a Liberal Arts or Comprehensive University, he immediately helped me to strategize a plan for getting there. I am certain that I would not currently have a tenure-track position at California Polytechnic State University in San Luis Obispo, without Vladimir's guidance.

Finally, I am very blessed to have found a fantastic faith community here in Seattle: the Catholic Newman Center at the University of Washington. The friends that I have made there are too numerous to list, but I will acknowledge Marcellino Tanumihardja, who led the church's music group when I joined it upon my arrival

to Seattle. I will never forget his charisma, his friendship, and the music that we all made together. It is also in this community where I met Elisha, my wife. Words cannot adequately express how deeply grateful to her that I am, for her love and support throughout all of these years, and especially during the last push to complete my dissertation.

DEDICATION

To my father,

Dr. Minn-Shong Chi (1940-1999)

To my mother,

Cecilia Chi

And to my wife,

Elisha Chi

Chapter 1
INTRODUCTION

Within the field of Statistical Genetics lie the subfields of pedigree analysis and phylogenetics (Falconer and Mackay, 1996; Felsenstein, 2004). A pedigree illustrates familial relationships among individuals, e.g. mother-child, siblings, cousins, etc. Similarly, a phylogeny is a graph representation of the evolutionary relationships among species, across evolutionary time. Thus, pedigrees and phylogenies are two analogous objects, in the sense that both of them impose a correlation structure on the units of observation (either individuals or species), based on their relationships (familial or evolutionary). Hence, problems in each subfield can be seen to have similar solutions.

In Chapter 3, we explore the pedigree setting with a study of heritability estimation. Our aim is to estimate heritability of osteoarthritis in a population of pig-tailed macaques. Typically, heritability estimation with large pedigree data proceeds by assuming that the trait of interest follows a normal distribution, and then performing likelihood-based estimation in either a maximum likelihood or Bayesian framework. However, our problem is non-standard, as our outcome measure of osteoarthritis consists of ordinal measurements along the spinal cord of each macaque, with many missing values and a high proportion of measurements equal to 0. We thus explore heritability estimation under different model choices, including the threshold model (Wright, 1934; Sorensen et al., 1995). While this model is not new conceptually, the appearance of widely-available and actively maintained software is fairly recent (Hadfield, 2010), and the statistical properties of heritability estimation under this model has not yet been thoroughly explored. Thus, we examine the consequences

of model misspecification and sample size requirements for heritability estimation of an ordinal trait, with simulation studies. This work may have possible extensions to phylogenetic regression with the threhold model (Felsenstein, 2005, 2012).

In Chapter 4, we turn to the phylogenetic setting with a study of recombination detection. Here, we define recombination as an exchange of genetic material between two different taxa (or species). The presence of recombination has implications for phylogenetic inference: if recombination between two taxa has occurred, then there will be two different evolutionary histories for the observed data (typically a DNA sequence alignment of several species of interest). Thus, it is important to identify the presence of recombination when performing phylogenetic inference; otherwise, accurate tree estimation can be severly compromised (Posada and Crandall, 2002; Felsenstein, 2004). Additionally, a phenomenon known as convergent evolution can also produce a similar signal to that of a recombination event, as selective pressures act on two distantly related species and cause their DNA sequences to be more similar to each other than would be expected over evolutionary time (Christin et al., 2012). We therefore propose an improvement to an existing recombination detection method, which avoids false positive signals due to convergent evolution.

Finally, in Chapter 5, we propose an improvement to least squares phylogenetic inference. Currently, least squares phylogenetic inference relies on a matrix of pairwise genetic distances, which are estimated from molecular sequence data, thus making the inference a two-step procedure. We propose a novel loss function to replace the usual least squares criteria, which condenses the usual two steps of least squares phylogenetic inference into one coherent procedure, and also uses the data more directly. We show that this leads to more efficient least squares estimation, and also provides a natural framework to account for more complex models of DNA evolution.

Chapter 2

PHYLOGENETIC INFERENCE: OVERVIEW

2.1 Preliminaries

2.1.1 What is a Phylogeny?

Evolutionary biologists are often interested in describing relationships between different species. This can be traced at least as far back as 1837, when Charles Darwin drew what is believed to be the first depiction of a phylogeny, shown from his notebook in Figure 2.1. Similar to a "family tree" or pedigree, which represents familial relationships between individuals, a phylogeny represents evolutionary relationships between species.

At first, phylogenies such as the one below were created by biologists' "expert opinion," as evidenced in this case by the words "I think" written next to the image. Numerical methods for inferring phylogenies began to appear with the advent of computing in the late 1950s (Michener and Sokal, 1957; Sneath, 1957; Sokal and Michener, 1958). In the work by Michener and Sokal (1957), the first phylogeny created by numerical methods was demonstrated for the genus *Proteriades* (bees), using differences in morphological characteristics. Present-day work in phylogenetics typically uses molecular sequence data, e.g. DNA or amino acid sequence alignments (Yang, 2006, Chapter 1).

For taxonomists, the phylogeny itself is usually the end goal of an analysis, characterizing the evolutionary relationships between a group of species (or more generically, "taxa") of interest. However, phylogenies are also used to answer a variety of scientific questions. For example, phylogenetic analyses were used to answer the questions of when, where and why rapid genetic diversification of salamander species occurred

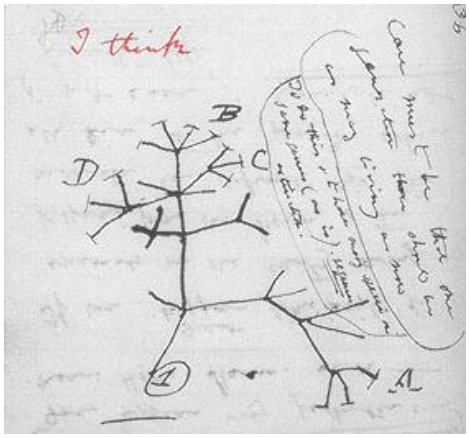

Figure 2.1: **From the notebook of Charles Darwin, 1837**. This is believed to be the first phylogeny ever drawn. *This work is in the public domain in the United States because it was published (or registered with the U.S. copyright office) before January 1, 1923.*

throughout their evolutionary history (Vieites et al., 2007). Viral strains have also been studied phylogenetically, e.g. to trace the origin of avian influenza (Lemey et al., 2009), or to help convict two suspected rapists of intentionally infecting their victims with HIV (Scaduto et al., 2010). Even individual cells within an organism can be studied phylogenetically, as daughter cells acquire some mutations during cell division, particularly cancerous cells which have a higher mutation rate than normal cells. For example, in leukemia, a phylogenetic analysis of affected cells has been used to answer questions about relapse mechanisms (Shlush et al., 2012).

When inferring a phylogenetic tree, one important distinction is whether the tree is "rooted" or "unrooted;" examples of each are shown in Figure 2.2. For the class of rooted phylogenies, the root is taken to be the origin of the lineage in question, with evolutionary time represented by the distance away from the root, typically expressed as the expected number of substitutions per site. However, most phylogenetic methods inherently infer "unrooted" phylogenies; to obtain a rooted tree, the two prevailing strategies are to either use a known outgroup, or to assume a molecular

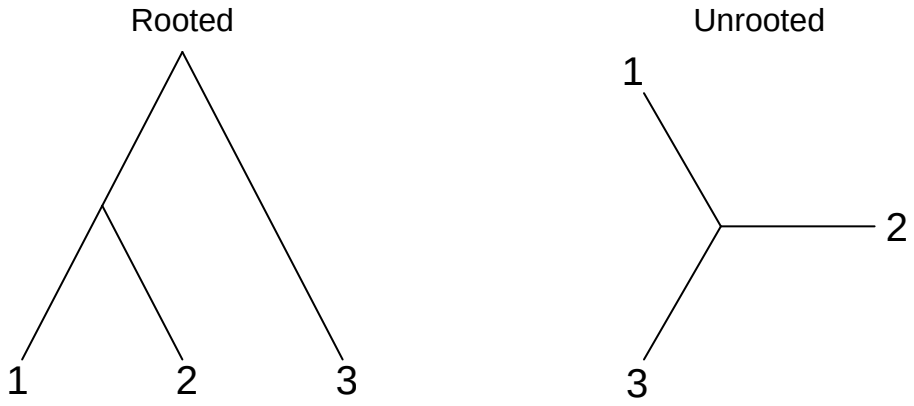

Figure 2.2: **A rooted and an unrooted phylogeny, with three taxa**.

clock (Yang and Rannala, 2012), which we will not discuss further here. We instead focus our attention on unrooted trees throughout the remainder of this work, with the assumption that we can root them if needed.

Another characterization of phylogenies is whether they are bifurcating or multifurcating. Bifurcating trees are those in which every split results in exactly two taxa, or in other words, that every internal node connects exactly three branches. We observe this to be true of both the rooted and unrooted trees in Figure 2.2; thus, these are bifurcating trees. Multifurcating trees are those in which at least one split results in three or more taxa. In this work, we focus our attention on bifurcating trees, mainly because a multifurcating tree can be represented by a bifurcating tree with an appropriate internal branch set to zero.

Finally, there is the distinction between labeled and unlabeled phylogenies. Unlabeled phylogenies do not have names associated with any of the tips, whereas labeled phylogenenies do. Clearly, in applied work when one wants to uncover the evolutionary history of a group of subjects (whether they are species, virus strains, cells, or otherwise), the phylogenies of interest are labeled. Thus, we focus our attention on labeled phylogenies.

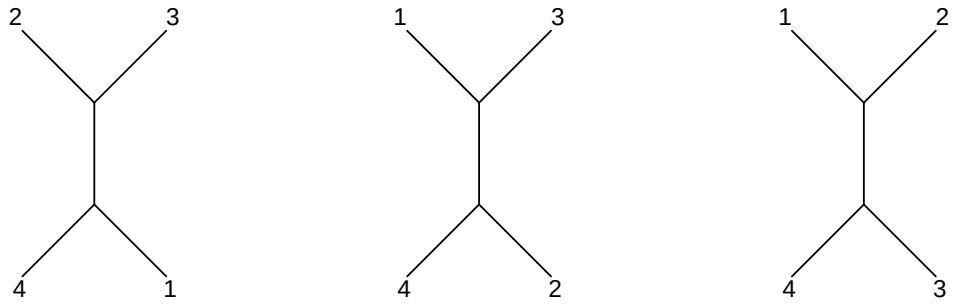

Figure 2.3: **All possible topologies for four taxa**. Labeled, unrooted and bifurcating phylogenies only.

2.1.2 Tree Space

When the goal is to infer the "best" phylogeny for a group of species, it may be useful to know how many possible phylogenies there are to choose from. To begin to answer this question, it is important to note that the space of phylogenies has both a discrete nature and a continuous nature: branch lengths are continuous and can take any non-negative value, whereas the shape, or topology, of the phylogeny is discrete. For four taxa, the set of all possible unrooted, bifurcating and labeled topologies is shown in Figure 2.3.

When the number of taxa is small, an exhaustive search through the set of all topologies is possible, in order to determine which one best describes the data. Estimation of branch lengths will be discussed later, but typically this would be done for each topology before choosing the overall optimum phylogeny. An exhaustive search with branch length estimation for each topology would guarantee that the best tree (according to some chosen criteria) is found. Unfortunately, as the number of taxa increases, this strategy becomes virtually impossible. For n taxa, the number of unrooted, bifurcating and labeled topologies is equal to $3 \times 5 \times \ldots \times (2n-5) = (2n-5)!!$.

Table 2.1: **Total number of possible topologies for a given number of taxa**. For labeled, unrooted, bifurcating phylogenies.

Number of taxa	Number of topologies
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
20	$\approx 2.216 \times 10^{20}$
30	$\approx 8.678 \times 10^{36}$
40	$\approx 1.311 \times 10^{55}$
50	$\approx 2.838 \times 10^{74}$

Table 2.1 shows this for increasing values of n.

Thus, for datasets of any size greater than roughly seven sequences, alternative strategies for searching through tree space are necessary. There exists a class of heuristic search techniques, in which, fundamentally, a candidate tree is evaluated according to the chosen criteria, and then branches are rearranged. The simplest of these techniques is known as nearest-neighbor interchanges, or NNI (Moore et al., 1973; Waterman, 1978). Essentially, this technique simply swaps two adjacent branches on a tree, to move from one topology to another. More formally, consider the three topologies in Figure 2.3, but instead of having the four taxa, imagine that each taxa represents an entire subtree, each with more branches and taxa. Then, an example of a nearest-neighbor interchange would be to start with the left-most tree, and swap the branches so that it becomes either of the other two trees in the figure. The entire tree can then be manipulated in this manner, to move from one topology to any other possible topology of the same number of taxa (after some number of steps).

Although we have not yet discussed any specific criteria for assessing whether a

tree is the "best" one (see Section 2.2), we have hinted at it when discussing exhaustive searches above. In contrast, one could start from any given tree, calculate the chosen criteria, and then perform nearest-neighbor interchanges in search of the tree which has the best score for that criteria. One common strategy is to only consider the tree (or trees, in case of ties) which have the best score among all neighbors, at each step of the search; that is, to use a "greedy" algorithm (Cormen et al., 2009, Chapter 16). Such a strategy would ensure that a local optimum is obtained, but as there is no guarantee (and in fact, it is highly unlikely) that any optimality criteria would be monotone across tree space, there is thus no guarantee that a greedy algorithm would obtain the global optimum.

Strategies to counter this are to use multiple starting trees, and to "be less greedy;" that is, explore some nearest-neighbor interchanges that are not the best among all neighbors. Also, there are other techniques for moving through tree space, such as subtree pruning and regrafting (SPR), and tree bisection and reconnection, or TBR (Felsenstein, 2004). We will not discuss the details of these here, but their benefit is that each of them defines a greater number of one-step neighbors than NNI does. For example, for 11 taxa trees, NNI defines 16 neighbors in each step, whereas SPR defines 288 neighbors, and TBR defines 296 neighbors. Thus, since each step is taken among a greater number of candidates, the chance that the global optimum will be reached is greater (even with a greedy algorithm).

Each of these strategies for increasing the chance of obtaining the global optimum, however, also require that more trees will be searched, and so provide less benefit relative to performing an exhaustive search. Also, with any strategy other than an exhaustive search, there is simply no way to be certain that the global optimum has been obtained. Optimization strategies continue to be an active area of research, and while there are others not discussed here, this is a common unresolved theme throughout the field of phylogenetics since tree space implies a complex landscape to search, under any optimality criteria (Felsenstein, 2004).

2.1.3 Models of DNA Evolution

Most methods for phylogenetic inference are model-based; thus, we include a discussion of various models for DNA evolution here. The most commonly-used models for this purpose are continuous-time Markov chains, modeling the change between DNA nucleotides, independently at each site of the sequence alignment. The main feature of these models is the Markov property: the probability of substitution from one nucleotide to any other nucleotide depends on the current nucleotide state, but does not depend on any past states. Generically, the substitution rate matrix is designated by $\mathbf{\Lambda} \equiv \{\lambda_{ij}\}$ for $i, j \in (A, C, G, T)$, the nucleotide state space. Further specifications lead to some of the commonly-used models, which we discuss below.

JC69

The first and simplest continuous-time Markov chain model of DNA evolution was proposed by Jukes and Cantor (1969), called JC69. This model assumes that the substitution rates between each pair of distinct nucleotides are identical; that is, $\lambda_{ij} \equiv \alpha$ for some α and for all $i \neq j$. As each row of the substitution rate matrix must sum to 0, this gives us

$$\mathbf{\Lambda} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \tag{2.1}$$

as the substitution rate matrix. Then, the transition probability matrix is given by $\mathbf{P}(t) = e^{\mathbf{\Lambda} t}$ (Guttorp, 1995), where $\mathbf{P}(t) \equiv \{p_{ij}(t)\}$, and each $p_{ij}(t)$ is the probability that a nucleotide site will be in state j after time t, given that it started in state i.

Of note is that the stationary distribution $\boldsymbol{\pi} \equiv (\pi_A, \pi_C, \pi_G, \pi_T)$ for the JC69 model is equal to $(1/4, 1/4, 1/4, 1/4)$; that is, as $t \to \infty$, the probability of being in any of the

four states is equal to 1/4. It is then easy to see that the continuous-time Markov chain for this model is reversible; that is: $\lambda_{ij}\pi_i = \lambda_{ji}\pi_j$ for all i, j. This allows us to solve for $\mathbf{P}(t)$, since reversibility of a continuous-time Markov chain implies diagonalizability of its substitution rate matrix $\mathbf{\Lambda}$. Thus, matrix exponentiation can be accomplished by diagonalizing $\mathbf{\Lambda}$ and arriving at its spectral decomposition (Schott, 1997), which for this model gives

$$\mathbf{P}(t) = e^{\mathbf{\Lambda} t} = \frac{1}{4}\begin{pmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{pmatrix}, \tag{2.2}$$

where we note that the diagonal entries are all the same, and likewise for the off-diagonal entries; we will refer to them as $p_0(t)$ and $p_1(t)$ respectively.

One class of methods for phylogenetic inference relies on pairwise distances between sequences (see Section 2.2.2). Evolutionary distances between two sequences are generally defined as the expected number of substitutions between them (Yang, 2006). For the JC69 model, we observe from (2.1) that the total rate of substitution from any given nucleotide is 3α. Then, if two sequences are seperated evolutionarily by time t, the distance between the sequences would be

$$d = 3\alpha t. \tag{2.3}$$

This relationship in (2.3) allows us to relate the distance d to a quantity that is observable from our data, $\hat{p}$: the proportion of sites that are different between two sequences, where p is the theoretical probability that a site is different among the two sequences. One might assume that $\hat{p}$ itself, also known as the Hamming distance, is a reasonable distance metric. The main drawback of using $\hat{p}$ is that it does not account for multiple substitutions at a given site; thus, Hamming distances are not additive,

and are not a linear function of evolutionary time (Felsenstein, 2004; Yang, 2006).

From (2.2), we note that the probability that any given site is different in two sequences after time t is $p = 3p_1(t)$. Then,

$$p = \frac{3}{4} - \frac{3}{4}e^{-4\alpha t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}, \tag{2.4}$$

using (2.2) and (2.3). By the method of moments, we can then obtain

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right), \tag{2.5}$$

to estimate the distance between a pair of DNA sequences under the JC69 model. We note that this same expression for distances as a function of $\hat{p}$ could also be obtained by maximum likelihood.

Also, it is important to observe that $\mathbf{\Lambda}$ and t appear as a product in the expression for $\mathbf{P}(t)$, and are not separable. Hence, we can estimate d, but we cannot estimate $\mathbf{\Lambda}$ or t individually. This is also true for the remainder of the models which follow.

K80

In real data, the class of substitutions known biologically as transitions ($T \leftrightarrow C$ or $A \leftrightarrow G$) often occur at a higher rate than transversions ($T, C \leftrightarrow A, G$). Thus, Kimura (1980) proposed a model that accounts for these potentially different rates, called K80. The substitution rate matrix for K80 is thus

$$\mathbf{\Lambda} = \begin{pmatrix} -(\alpha+2\beta) & \beta & \alpha & \beta \\ \beta & -(\alpha+2\beta) & \beta & \alpha \\ \alpha & \beta & -(\alpha+2\beta) & \beta \\ \beta & \alpha & \beta & -(\alpha+2\beta) \end{pmatrix}, \tag{2.6}$$

where the row and column labels are in alphabetical order: (A, C, G, T). Then, the distance between two sequences can be expressed as $d = (\alpha + 2\beta)t$.

Instead of using the parameters α and β directly, it is often more convenient to consider the ratio $\kappa = \alpha/\beta$ and then estimate this along with d. Letting $\hat{p}_S$ be the proportion of sites with transitional substitutions, and $\hat{p}_V$ be the proportion of sites with transversional substitutions observed in a pair of sequences, it was shown that

$$\hat{d} = -\frac{1}{2}\log(1 - 2\hat{p}_S - \hat{p}_V) - \frac{1}{4}\log(1 - 2\hat{p}_V), \tag{2.7}$$

$$\hat{\kappa} = \frac{2\log(1 - \hat{p}_S - \hat{p}_V)}{\log(1 - 2\hat{p}_V)} - 1, \tag{2.8}$$

by Kimura (1980) and Jukes (1987).

F84

Adding another layer of complexity, we note that the JC69 and K80 models both induce the stationary distribution of $\boldsymbol{\pi} = (1/4, 1/4, 1/4, 1/4)$. This assumption is unreasonable for most real datasets, since nucleotide frequencies are not typically observed to be all equal in real data. The F84 model, which is one of many models to account for this, was first implemented in the DNAML program of version 2.6 (1984) of the PHYLIP package (Felsenstein, 1989), but the first appearance of its substition rate matrix in print was by Hasegawa and Kishino (1989). It is:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \cdot & \beta\pi_C & (1 + \frac{\kappa}{\pi_R})\beta\pi_G & \beta\pi_T \\ \beta\pi_A & \cdot & \beta\pi_G & (1 + \frac{\kappa}{\pi_Y})\beta\pi_T \\ (1 + \frac{\kappa}{\pi_R})\beta\pi_A & \beta\pi_C & \cdot & \beta\pi_T \\ \beta\pi_A & (1 + \frac{\kappa}{\pi_Y})\beta\pi_C & \beta\pi_G & \cdot \end{pmatrix}, \tag{2.9}$$

where $\pi_R = \pi_A + \pi_G$, $\pi_Y = \pi_C + \pi_T$, and the diagonal entries are the appropriate values to make each row sum to 0.

Similar to the K80 model, the F84 model accounts for different rates of transitions and transversions, but with the added flexibility that the stationary distribution $\boldsymbol{\pi}$ is unrestricted. Expressions for estimators of d and κ can be found in the work by Tateno et al. (1994).

GTR

The general time-reversible (GTR) model was first suggested by Tavaré (1986), with subsequent developments by many others. The GTR model allows for the rates of change between any two nucleotide states to be distinct, in addition to having no restrictions on the stationary distribution $\boldsymbol{\pi}$. The substitution rate matrix is given by:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \cdot & \alpha\pi_C & \beta\pi_G & \gamma\pi_T \\ \alpha\pi_A & \cdot & \mu\pi_G & \tau\pi_T \\ \beta\pi_A & \mu\pi_C & \cdot & \rho\pi_T \\ \gamma\pi_A & \tau\pi_C & \rho\pi_G & \cdot \end{pmatrix}, \tag{2.10}$$

where again the diagonal entries are the appropriate values to make each row sum to 0. Thus, this model allows for the most flexibility among commonly used reversible continuous-time Markov chain models, at the expense of having the most parameters. Distance estimation under this model has been a complex issue throughout the literature, as various expressions for estimators of pairwise distances have been demonstrated in numerous publications (Rodriguez et al., 1990; Gu and Li, 1996; Yang and Kumar, 1996; Waddell and Steel, 1997).

Summary

There are many other models that we did not discuss here. However, these selected models represent some of the key features among all continuous-time Markov chain

models of DNA evolution, and are among the most commonly used. Each of these models is reversible, although non-reversible models do exist (Yang, 1994b). Generally speaking, while the more complex models are often more realistic, it comes at the cost of increasing the number of parameters. Also, at least for small distances (i.e. when sequences differ by roughly 10% or less), the choice of model does not have a significant impact on distance estimates (Yang, 2006).

2.2 Methods for Phylogenetic Inference

In this section, we introduce the most commonly used methods for estimating a phylogeny. As mentioned above, most of these methods are model-based. We begin, however, with the model-free class of methods known as parsimony. We include these in part because they are arguably the most intuitive class of methods, but also because they are still in use today, although perhaps not as widely as before.

2.2.1 Parsimony Methods

Consider a simple, four taxa dataset with five DNA nucleotide sites:

Table 2.2: **Four taxa sequence alignment**.

Taxon 1	A	T	G	T	T
Taxon 2	T	A	G	T	T
Taxon 3	T	A	G	A	T
Taxon 4	A	T	C	C	C

The parsimony method, in its most basic form, then attempts to identify the topology that allows for the fewest possible number of required state changes. We will consider each of the three topologies shown previously in Figure 2.3 (in Section 2.1.2), recalling that this is the set of all possible labeled, unrooted and bifurcating trees for four taxa.

Starting with the leftmost tree, we examine the sites one at a time, and keep track of the cumulative number of state changes required across the alignment. For the

sake of illustration, let us start at Taxon 1 at the lower-righthand corner of the tree, in state A (although the result would be the same regardless of where we start). We then trace the lineage, propagating the state A until a change is required. When a split occurs, the state continues in both directions: e.g. the first split leads to Taxon 4 in one direction, and to the remainder of the tree in the other direction. On the branch that leads to Taxon 4, we note that no state change is required, since Taxon 4 is also in state A. However, Taxa 2 and 3 are both in state T; thus, a change is required. This result can be observed with a change to state T at some point on the middle branch. Then, no further changes are required, to observe the site pattern of (A, T, T, A) for Taxa (1, 2, 3, 4).

Thus, for the first site, one change is required. We can then proceed through the remainder of the sites, and would observe that a minimum of six changes are required on this tree to obtain the given dataset. Therefore, the parsimony score for this tree is six. Examining the two remaining trees in Figure 2.3, we would find that the parsimony scores are eight for each of them. Therefore, the maximum parsimony tree for this dataset is the leftmost tree in Figure 2.3.

For this small dataset, it is easy to find the maximum parsimony score manually as we just did. However, for larger datasets, algorithms will prove to be useful. A few of the first ones were proposed by Camin and Sokal (1965), Kluge and Farris (1969), Farris (1970), and Fitch (1971). These essentially accomplish the calculation of the parsimony score as described above. Further developments included assigning a cost matrix to weight each type of substitution differently (Sankoff, 1975), and strategies for economizing computations (Gladstein, 1997; Ronquist, 1998).

One drawback to parsimony methods is that they inherently do not estimate branch lengths. Also, their simplicity is of course paid for by some undesirable qualities, most notably that parsimony methods have been shown to be statistically inconsistent under certain scenarios (Cavender, 1978; Felsenstein, 1978). This will be discussed in more detail in Section 2.3, but for now we proceed towards introducing

model-based methods.

2.2.2 Distance matrix methods

The earliest class of model-based methods are those using the pairwise distances that were introduced in Section 2.1.3 (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967). Here, we will briefly describe a few of the distance-based methods.

UPGMA

UPGMA stands for "unweighted pair group method using arithmetic averages," and it is a clustering procedure proposed by Sokal and Michener (1958). Contrary to all other methods that we discuss, UPGMA actually infers rooted trees. Also, it assumes that trees are ultrametric, meaning that the distance from the root to every tip is the same.

The algorithm for UPGMA starts by considering the matrix of pairwise distances, that can be obtained by any of the models in Section 2.1.3 or otherwise. For example, with four taxa, we might have a distance matrix $\mathbf{D} = \{d_{kl}\}$:

Table 2.3: **Distance matrix with four taxa**.

	Taxon 1	Taxon 2	Taxon 3	Taxon 4
Taxon 1	0	50	45	20
Taxon 2	50	0	**15**	55
Taxon 3	45	**15**	0	52
Taxon 4	20	55	52	0

Then, the first step of this algorithm identifies the pair (k, l) which have the smallest distance, designated in bold above. These two taxa are then connected to a new node, at a height (or distance from the tips) of $d_{kl}/2$. The columns and rows for k and l are then removed and replaced by one column and row whose elements are the average of the distances from k and l. In this example, we would then have:

Table 2.4: **New distance matrix after one step of UPGMA algorithm**.

	Taxon 1	Taxon (2,3)	Taxon 4
Taxon 1	0	47.5	20
Taxon (2,3)	47.5	0	53.5
Taxon 4	20	53.5	0

In subsequent steps, the averages are weighted if a particular node is the conglomeration of two or more, as in Taxon (2,3) currently. For this example, the following tree would be produced:

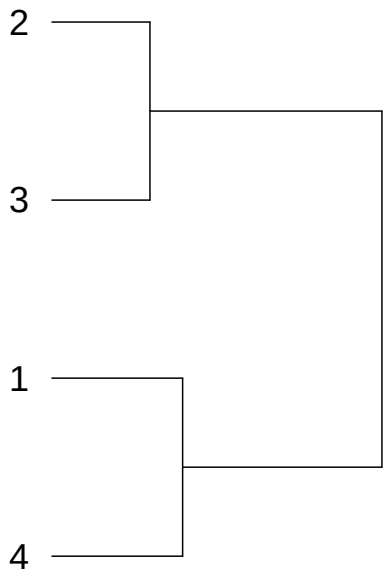

Figure 2.4: **UPGMA tree**.

Neighbor-joining

Because of its similarity to UPGMA, we will not describe the neighbor-joining algorithm (Saitou and Nei, 1987) in detail, but instead will point out the features that make it different. Firstly, the neighbor-joining algorithm does not assume that trees

are ultrametric, and also does not infer rooted trees. The algorithm begins by computing

$$u_k = \frac{1}{n-2} \sum_{l \neq k} d_{kl} \tag{2.11}$$

for each u_k, where k and l are tip labels, and n is the number of tips. Then, we choose the k and l for which the quantity $d_{kl} - (u_k + u_l)$ is smallest. This is in contrast to simply choosing the smallest d_{kl} as in the UPGMA algorithm, which solves one fundamental issue with UPGMA. Suppose the true tree is the one depicted in Figure 2.5. Then, we observe that Taxa 3 and Taxa 4 are the closest together in terms of evolutionary time, but yet they are on separate lineages of the tree. The UPGMA algorithm would incorrectly cluster Taxa 3 and Taxa 4 together based on their proximity to each other; conversely, the neighbor-joining algorithm is able to distinguish this nuance, and will cluster Taxa 4 with Taxa 1, and Taxa 2 with Taxa 3 as the true topology indicates.

Least Squares

While the UPGMA and neighbor-joining methods use the distance matrix, they do not explicitly optimize any criteria in their attempt to reconstruct the tree; that is, there is no score assigned to potential candidate trees, as in the parsimony methods discussed in Section 2.2.1. In constrast, least squares methods for phylogenetic inference draw on the well-established statistical framework after which it is named, in which the least squares criterion is minimized in order to find the optimal model.

First, let $\boldsymbol{\tau}$ represent the topology of a tree, and $\mathbf{b}$ represent its branch lengths. Then, we define the tree distance $t_{kl}(\boldsymbol{\tau}, \mathbf{b})$ between taxa k and l as the sum of the branch lengths between them on the particular phylogeny determined by $\boldsymbol{\tau}$ and $\mathbf{b}$.

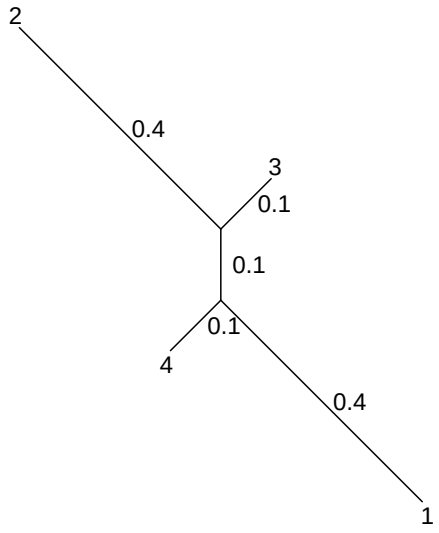

Figure 2.5: **Unbalanced tree**. Numbers adjacent to each branch indicates branch length.

Using the distances d_{kl} defined in Section 2.1.3, we then have

$$\sum_{k=1}^{n}\sum_{l=1}^{n}\left(\hat{d}_{kl} - t_{kl}(\boldsymbol{\tau}, \mathbf{b})\right)^2, \tag{2.12}$$

which is essentially the usual least squares linear regression criterion, adapted to our context of phylogenetics. The solution $(\hat{\boldsymbol{\tau}}, \hat{\mathbf{b}})$ to

$$\underset{\boldsymbol{\tau},\mathbf{b}}{\operatorname{argmin}} \sum_{k=1}^{n}\sum_{l=1}^{n}\left(\hat{d}_{kl} - t_{kl}(\boldsymbol{\tau}, \mathbf{b})\right)^2 \tag{2.13}$$

then gives the least squares phylogeny. This method was first proposed by Cavalli-Sforza and Edwards (1967).

2.2.3 Maximum likelihood methods

Since its introduction by Fisher (1912), maximum likelihood methods have become perhaps the most standard framework for most statistical problems. Maximum likelihood was first applied to phylogenetics by Edwards and Cavalli-Sforza (1964). However, its feasibility for even a moderate number of DNA sequences was lacking until an important contribution by Felsenstein (1981).

We begin by making two key assumptions: 1) Different sites across a sequence alignment evolve independently; 2) evolution in different lineages is independent, conditional on the most recent ancestor of these lineages. Then, using notation from Section 2.2.2, we define our likelihood in the usual way:

$$L = P(\text{data}|\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}) = \prod_i P(\text{data}_i|\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}), \tag{2.14}$$

where data_i is the data at the i^{th} site (whether it is a nucleotide or amino acid sequence), and $\boldsymbol{\theta}$ is the vector of parameters from the chosen model of evolution (see Section 2.1.3). Consider, then, the simple dataset from Section 2.2.1:

Table 2.5: **Four taxa sequence alignment, revisited**.

Taxon 1	A	T	G	T	T
Taxon 2	T	A	G	T	T
Taxon 3	T	A	G	A	T
Taxon 4	A	T	C	C	C

Our aim is to determine which tree gives this data the highest probability of occurring, given a model (e.g. Section 2.1.3). Consider the following tree:
Each b_j represents the length of the branch it labels, and y, z are the states at the respective interior nodes. Suppose temporarily that the states of y and z are known. Since we have assumed that lineages are independent, then for the first site of the

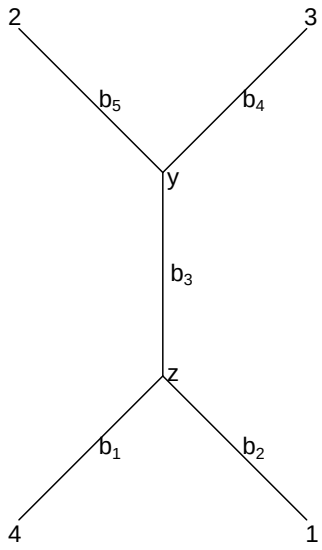

Figure 2.6: **Four taxa tree with branch lengths**.

dataset, with this topology we have:

$$P(\text{data}_1|\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}) = \pi_z p_{zA}(b_1) p_{zA}(b_2) p_{zy}(b_3) p_{Ty}(b_4) p_{Ty}(b_5) \tag{2.15}$$

where y and z represent their known states (e.g. A, C, G or T for each), π_z comes from the stationary distribution of the model, and $p_{..}(b_j)$ are the finite-time transition probabilities from the model (see Chapter 2.1.3). This seemingly implies that the root of the tree is at the node designated by z. However, it is easy to show that the choice of where to place the root does not change the value of this expression; thus, the maximum likelihood method infers unrooted trees only.

We note, however, that the interior nodes are typically unobserved. In theory, this does not pose much of a problem, since we can simply use the law of total probability to sum over all possible nucleotide states, for both interior nodes. This results in 4^2 terms to sum, where each term is calculated as the product of probabilities as in

(2.15). Sixteen terms to sum is not a problem, but the number of terms increases exponentially as the number of taxa increases; for n taxa, there are $n-2$ interior nodes, meaning that there will be 4^{n-2} terms to sum. At $n = 20$, this results in approximately 6.8×10^{10} terms. This must also then be done at every nucleotide site, and for every candidate tree (whether an exhaustive search or some other tree searching strategy is performed, as described in Section 2.1.2). The contribution by Felsenstein (1981) was to apply the method well-known to computer scientists as dynamic programming, to economize the calculations greatly. Here, the computational complexity was reduced from $\mathcal{O}(4^{n-2})$ with naive calculations to $\mathcal{O}(4 \times (n-2))$ with dynamic programming, and making maximum likelihood phylogenetic inference feasible for moderately large datasets (Felsenstein, 2004).

2.2.4 Bayesian methods

The goal of Bayesian phylogenetic inference is to obtain

$$P(\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}|\text{data}) = \frac{P(\text{data}|\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}) \times P(\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta})}{P(\text{data})}, \tag{2.16}$$

which is the posterior probability of each tree given the data, and then perhaps choose the tree which has the maximum posterior probability (Rannala and Yang, 1996). For the prior distribution $P(\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta})$, typically some sort of noninformative prior is typically chosen, putting equal prior probability on all possible trees or some class of trees (Mau and Newton, 1997; Li et al., 2000). The quantity $P(\text{data}|\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta})$ is simply the likelihood, discussed in the previous section. The real issue, computationally, is the marginal probability $P(\text{data})$, for which an explicit calculation would require summing over all possible topologies and integrating over branch lengths, and all substitution parameters. Thus, as is typical in practically all current Bayesian analyses, the posterior distribution is approximated by Markov chain Monte Carlo (MCMC).

Of course, it is impossible to literally answer the question of which tree has the

maximum posterior probability, since branch lengths are continuous and thus each tree is equally infinitely improbable. Even when constrained to the probability of topologies, if the number of taxa is large and the data are noisy, there may be millions of topologies represented somewhat equally in the posterior distribution, and thus it still might not be optimal to simply choose one of them. One common strategy is to instead us a majority rule consensus tree (Margush and McMorris, 1981), which summarizes a group of trees by consisting of the splits which occur the most frequently among the group of trees; this applies nicely to the posterior sample of trees, and the split frequencies have the convenient interpretability of being the posterior probability of each clade (Huelsenbeck et al., 2001).

2.3 Statistical Comparisons of Methods

While current research is more active for some than others (both methodological and applied), all of the above methods are still in use today. Each of them has its benefits and drawbacks, and we summarize some key points as follows.

2.3.1 Consistency

Statistical consistency is the property that, as the amount of data approaches infinity, the estimated parameter values approach their true values; in phylogenetics, this means that as the sequence length approaches infinity, the estimated tree will approach the true true, with respect to both topology and branch lengths. We have already mentioned that parsimony methods are inconsistent under some circumstances (Section 2.2.1). All of the model-based methods (distance-based, maximum likelihood and Bayesian) can suffer from this as well, if the assumed model is too simplistic. However, without model misspecification, each of them does have the property of statistical consistency (Yang and Rannala, 2012). This was proven analytically for maximum likelihood phylogenetic inference, under the assumption that the model is well-formulated so that all trees are identifiable (Yang, 1994c).

2.3.2 Efficiency

Efficiency is typically measured by comparing the mean squared error (MSE) of two estimators (or simply the variances, if both estimators are unbiased), and the estimator with the smaller MSE or variance is the more efficient estimator (Shorack, 2006). If an estimator attains the Cramér-Rao lower bound (Cramér, 1945; Rao, 1945) as the sample size tends towards infinity, then the estimator is called asymptotically efficient. Under what are typically considered to be mild regularity conditions, the maximum likelihood estimator is known to be asymptotically efficient; however, these regularity conditions do not hold for phylogenetic trees (Yang and Rannala, 2012). The efficiency of various phylogenetic estimators has been studied, however, and while not all of them have been compared rigorously, it has been shown through numerous simulation studies that likelihood-based methods have better efficiency than both parsimony and distance methods (Saitou and Imanishi, 1989; Jin and Nei, 1990; Hasegawa and Fujiwara, 1993; Kuhner and Felsenstein, 1994; Tateno et al., 1994; Huelsenbeck, 1995b).

2.3.3 Robustness

Model misspecification to some extent is inevitable in phylogenetic analyses. The extent to which a method performs well when model assumptions are incorrect is known as robustness. One method that relies heavily on its assumptions is the UPGMA algorithm, which, as mentioned above, assumes that trees are ultrametric. This is equivalent to assuming a molecular clock, which states that all sites on all lineages evolve at the same rate (Ho, 2008). When this assumption fails to hold (as is true in most real data cases), UPGMA has been shown to fail drastically (Felsenstein, 2004).

Not surprisingly, certain assumptions are more critical than others. For example, misspecification of the transition/transversion ratio and assumptions about each nucleotide base frequency do not seem to affect phylogenetic inference greatly (Huelsen-

beck, 1995a). An assumption that seems to be more important is that the sites evolve at the same rate, which is important both for calculating pairwise distances (Section 2.1.3) and for formulating and computing the data likelihood for a given tree (Section 2.2.3). If sites do evolve at different rates, then it has been shown that both distance-based and maximum likelihood phylogenetic inference is inconsistent (Chang, 1996). Distances can be adjusted for evolutionary rate heterogeneity, if given a model and parameter value(s), whereas maximum likelihood phylogenetic inference can incorporate evolutionary rate heterogeneity explictly into the model, and uses the information more efficiently (Felsenstein, 2004). Overall, likelihood methods appear to be more robust than distance-based methods, but the extent is unclear (Yang, 2006).

2.3.4 Speed

While computational speed is not, strictly speaking, a statistical criteria, it is nonetheless important to consider when evaluating different methods. Not surprisingly, the more complex (and thus, generally speaking, more realistic) methods are slower and more computationally intensive. UPGMA and neighbor-joining are typically the fastest, as they are clustering algorithms and thus do not need to search tree space; least squares phylogenetic inference is also very fast, as the problem can be formulated as a system of equations and thus can be solved by matrix algebra as in the usual least squares regression. In some cases, least squares phylogenetic inference is seen as the "sweet spot" in terms of its balance between computational speed and statistical justifiability (Felsenstein, 2004). Regarding maximum likelihood, even with the improvements by Felsenstein (1981), maximum likelihood phylogenetic inference is still fairly computationally expensive for moderate to large datasets. Bayesian methods rely on MCMC estimation of the posterior distribution, and thus their speed depends entirely on how long the MCMC chain is run; typically, it takes longer to achieve convergence of the MCMC chain than it does for maximum likelihood to obtain a point estimate from a heuristic tree search (Yang, 2006). However, a complete maxi-

mum likelihood phylogenetic analysis is typically slower than a Bayesian analysis (see below).

2.3.5 Miscellaneous

If computational speed were of no concern, then in general, the maximum likelihood and Bayesian methods would be superior. Compared to each other, their performance appears to be similar, as they are both able to use complex mutational models and have good statistical properties. One benefit of Bayesian methods is that, as mentioned in Section 2.2.4, the procedure produces the posterior probabilities of each clade on the tree. To obtain a measure of uncertainty for the maximum likelihood tree, one typically uses a parametric bootstrap (Felsenstein, 1985), which can then place bootstrap support on each clade. However, the interpretation of these is less clear than the posterior probabilities obtained by the Bayesian method, as competing interpretations have been offered in the literature (Felsenstein, 1985; Felsenstein and Kishino, 1993; Efron et al., 1996). The desired interpretation generally appears to be that of accuracy, or that the bootstrap proportion is the probability that the clade is true; clearly, this is the Bayesian interpretation. Furthermore, the parametric bootstrap is computationally expensive, as the maximum likelihood estimation procedure must be performed again repeatedly; typically, maximum likelihood estimation with parametric bootstrap support appears to be slower than Bayesian estimation, which inherently provides posterior probabilities of each clade (Yang, 2006).

While parsimony and distance-based methods are relatively easy to program, they are also available in software packages such as PHYLIP (Felsenstein, 1989) and TNT (Goloboff et al., 2008). Maximum likelihood methods are available in PhyML (Guindon et al., 2010), RAxML (Stamatakis, 2006), and PHYLIP. Based on citation count (Yang and Rannala, 2012), two of the most popular phylogenetic reconstruction packages are Bayesian: MrBayes (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012) and BEAST (Drummond et al., 2012).

Chapter 3

HERITABILITY ESTIMATION OF AN ORDINAL TRAIT: OSTEOARTHRITIS IN PIG-TAILED MACAQUES

3.1 Introduction

Osteoarthritis is a condition that is characterized by the breakdown of cartilage in joints between bones, and can occur in any joint in the body. Those who suffer from osteoarthritis may experience pain and soreness in the affected area, and even a lack of mobility, particularly in spinal osteoarthritis. Thus, spinal osteoarthritis is a serious worldwide public health concern, and a better understanding of this disease can lead to better treatment and care of patients who suffer from it (Hadjipavlou et al., 1999). This disease is characterized by several radiological features, including narrowing of the intervertebral disk space, bone spurs along the spinal cord (osteophytosis), and vertebral end-plate sclerosis (Lawrence, 1969). The conglomeration of these features is generally referred to as degenerative disk disease, or DDD (Vernon-Roberts and Pirie, 1977), although this term is also used to indicate the presence of a single one of these features (Cohn et al., 1997; Lawrence, 1969).

Specific aspects of DDD in humans have been well-characterized throughout the literature. For example, evidence for associations between DDD and various factors have been demonstrated, including age (Frymoyer et al., 1984; Riihimaki et al., 1990), body mass (Riihimaki et al., 1990), trauma (Kerttula et al., 2000), type and level of activity (Caplan et al., 1966; Riihimaki et al., 1990; Videman et al., 1990; Videman and Battie, 1999), and gender (Jones et al., 1988; Miller et al., 1988). Research in other mammals has corroborated the contribution of biomechanical stress to the development of DDD (DeRousseau, 1985; Schultz, 1969). Indeed, the bipedality and erect posture of humans has been assumed to be one of the primary causes of DDD in

our species (Bridges, 1994; Jurmain and Kilgore, 1995; Knusel et al., 1997; Schultz, 1969; Shore, 1935).

Nevertheless, much is still unknown about the etiology of DDD. In particular, the extent to which genetics plays a role in DDD development has not yet been uncovered. Since there are safety concerns posed by radiography, the macaque monkey is often used as an animal model for humans in the study of bone diseases, due to its close genetic relatedness to humans (Duncan et al., 2011, 2012). One may question its appropriateness for DDD as macaque monkeys are not bipedal, but this concern was addressed by Kramer et al. (2002), which explored DDD specifically in the macaque species known as pig-tailed macaques (*Macaca nemestrina*), and concluded that they are indeed an appropriate animal model for DDD in humans.

Here, we use a population of captive pig-tailed macaques to explore the question of whether there is a genetic component to DDD. To this end, we examine whether DDD is heritable. Heritability is a statistically defined quantity that describes the degree to which a trait is determined by genetics. Heuristically, if a trait has high heritability, then individuals who are more related to each other would appear more similar to each other than average, with regard to this trait. While genotyping in humans is now cheap and ubiquitous, heritability estimation in primates is still often performed to determine whether a trait warrants genotyping in the animal model, with the goal of mapping genes that control the trait.

Methodologically, we are interested in a model for the trait that would allow for a transparent heritability estimation. A common assumption is that the trait follows a normal distribution. This is generally justified by the polygenic model, which postulates that complex traits are under control by several additive, independent loci, with similar variances (Fisher, 1918). However, this assumption may be drastically violated in some real data problems. In particular, if the trait is ordinal with only a few categories, it is clear that the trait would not follow any distribution resembling normality. Likewise, normality may be violated even if there are many categories, but

one category is severely over-represented. Such is the case with our trait distribution in the pig-tailed macaques.

Heritability estimation with discrete data was first demonstrated for the binary case by Dempster and Lerner (1950), and extended to multiple ordered categories by Gianola (1979), using transformations of a continuous trait. However, some intrinsic difficulties to these tasks quickly presented themselves. First, unlike continuous traits, the variance of a binary trait is closely tied to the mean or prevalence of the trait, and thus provides no useful information about the inherent biological variability of interest (Burton et al., 2007). Furthermore, the observed scale of an ordinal trait may not be additive (e.g. an observation of "4" may not be equal to twice the value of an observation of "2"), thus leading to biases in parameter estimates (Gianola, 1982; Höschele, 1986). Thus, some authors have completely abandonded transforming/thresholding continuous models and has attempted to estimate heritability of discrete traits directly under Poisson and negative binomial mixed models (Foulley et al., 1987). However, these models have their own drawbacks as well, including the issue of whether heritability is even well-defined in these contexts (Matos et al., 1997).

Here, we consider the threshold model for ordinal data (Wright, 1934). This model makes the assumption that the value of the ordinal trait is dictated by an unobserved latent variable with a normal distribution, referred to by Wright (1934) as the *liability*. For example, with a binary variable that has observed states of 0 and 1, the value for any given individual would be determined by whether that individual's value of the liability is above or below some threshold. We choose this model primarily because of its biological justifiability, through applying the polygenic model to the liability. That is, regardless of the distribution of the observed trait, if the phenotype is determined by many genetic loci, then it is plausible that an underlying normally distributed liability would exist.

A comprehensive Bayesian framework for heritability estimation under the threshold model was formulated by Sorensen et al. (1995). Further work was done through

the next 10 years, e.g. improving the MCMC convergence (Cowles, 1996), and extending the framework to a censored normal outcome variable (Sorensen et al., 1998). There was, however, no widely available and actively maintained open-source software to perform such analyses until the recent appearance of the `MCMCglmm` package in R (Hadfield, 2010). Thus, the time has come when biologists with ordinal data wishing to estimate heritability using the threshold model can begin to do so more easily than before.

Of course, ordinal data will not necessarily follow the threshold model, and although we obviously cannot "know" what the true distribution of any real data trait is, it is still useful to examine the statistical properties of heritability estimates under various scenarios and models, to reveal the consequences of misspecifying the distribution of the trait. That is, if the trait follows the threshold model with a liability trait that follows a normal distribution, but we incorrectly assume that the trait itself follows a normal distribution, how is the estimation of heritability affected by this incorrect assumption? We examine this under a variety of scenarios and number of categories. We then examine heritability estimation on our actual dataset under these different models, which illuminates the concern of how much data are needed to obtain estimable heritability under this model. Thus, we conclude with an exploration of sample size requirements, which will be useful in guiding future studies.

3.2 Methods

3.2.1 Heritability

When estimation of heritability is performed with pedigree data, the structure of these data allows for the identifiability of the quantities that define heritability. The kinship coefficient $\mathbf{\Phi}$ and coefficient of identity $\boldsymbol{\kappa}_2$, also commonly referred to as Δ_7 (Wright, 1922; Jacquard, 1966), are two quantities well established by classical population genetics. Briefly, $\mathbf{\Phi}$ is a matrix whose components Φ_{ij} are defined as the probability

at a given locus that two gene copies chosen at random from two individuals i and j are Identical-By-Descent (IBD), and $\boldsymbol{\kappa}_2$ is also a matrix whose components $\kappa_{2_{ij}}$ are defined as the probability at a given locus that two individuals i and j share two gene copies IBD. Then, for any trait vector $\mathbf{Y}$ of measurements taken on individuals within the pedigree, the polygenic model (Fisher, 1918) posits that $\mathbf{Y}$ will have a multivariate normal distribution with covariance matrix

$$\boldsymbol{\Sigma} = 2\sigma_A^2\Phi + \sigma_D^2\kappa_2 + \sigma_E^2\mathbf{I}, \tag{3.1}$$

where σ_A^2 is the variance of the additive genetic effect, σ_D^2 is the variance of the dominant genetic effect, σ_E^2 is the variance of the environmental effect, and $\sigma_A^2+\sigma_D^2+\sigma_E^2 = \sigma_Y^2$ if there are no other effects to consider (such as household or maternal), and there is no interaction or correlation between effects (Lange, 2002). Heritability of the trait $\mathbf{Y}$ is then defined as the ratio of the additive genetic variance to the total variance of the trait: $h^2 = \sigma_A^2/\sigma_Y^2$.

3.2.2 Estimation under Normality

The framework of the polygenic model then leads us to consider a multivariate normal model for the vector of trait values from the whole sample. Under this model, the partitioning of the covariance matrix in (3.1) allows for estimation of these variance components through maximum likelihood. Furthermore, in this framework it is easy to adjust for covariates, as we can state that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\mathbf{X}$ is an $(n \times p)$ matrix for n individuals and p covariates of interest (e.g. age, weight, gender), and then $\boldsymbol{\beta}$ is a $(p \times 1)$ column vector of mean components. The βs are nuisance parameters since our object of interest is still just the variance components, but incorporating them into the model allows for control over confounders. Thus, we can write the usual multivariate normal likelihood: $L(\boldsymbol{\beta}, \sigma_A^2, \sigma_D^2, \sigma_E^2) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-0.5(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)$ where $\boldsymbol{\Sigma}$ is explicitly partitioned

into our variance components of interest; thus we have a tractable likelihood that we can attempt to maximize with respect to our parameters. Computational issues in solving for the roots of the likelihood equations for variance components estimation were addressed by Lange et al. (1976) and implemented in the `MENDEL` software package (Lange et al., 2001). One could also proceed using restricted maximum likelihood for fitting linear mixed models, available in software packages such as `ASReml` (Gilmour et al., 1995); however, we do not consider this here.

3.2.3 Threshold Model: Ordered Probit Regression

For ordinal data, a more realistic assumption than normality of the trait may be to assume that this trait is dictated by an underlying normally distributed latent variable. Then, an individual's value of the liability trait would determine which category that individual falls into for the observed trait. Formally, we consider the following model:

$$\mathbf{U} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \boldsymbol{\epsilon}; \quad P(Y_i = j) = P(t_{j-1} < U_i \leq t_j), \tag{3.2}$$

where $\mathbf{U} = (\mathbf{U_1}, \ldots, \mathbf{U_n})'$ is the vector of unobserved liabilities for each individual, and $\mathbf{a}$ is a random vector representing the breeding values for each individual, with $\mathbf{a}|\sigma_A^2 \sim N(0, 2\boldsymbol{\Phi}\sigma_A^2)$. Then, with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_E^2\mathbf{I})$, the latent variable vector $\mathbf{U}$ has the same covariance structure given by $\boldsymbol{\Sigma}$ in (3.1) above, assuming here that $\sigma_D^2 = 0$. Finally, $\mathbf{t} \equiv (\mathbf{t_0}, \ldots, \mathbf{t_C})$ are the true but unknown cutpoints on the distribution of the latent variable, which, along with the values of each U_i, determine the values of each Y_i, where $\mathbf{Y}$ is the observed categorical outcome vector. This forms the basis for Ordered Probit Regression (OPR).

Heritability estimation under this model could be performed through either Maximum Likelihood or Bayesian approaches. Since open-source implementation for Bayesian approaches to heritability estimation under this framework are readily avail-

able, we proceed in that manner. Namely, we use the R package `MCMCglmm` (Hadfield, 2010). Ideally, we would like to approximate the posterior distribution of σ_A^2 and σ_E^2 so that we can estimate heritability. Here, this is done along with concurrent estimation of $\mathbf{U}$, $\boldsymbol{\beta}$, and $\mathbf{t}$, given the data $\mathbf{Y}$ that we observed and the pedigree. We impose inverse gamma prior distributions on σ_A^2 and σ_E^2, with shape and scale parameters (α_A,γ_A and α_E,γ_E, respectively) of 0.01. We note here that these distributions on the individual variance components impose a Beta(0.01,0.01) prior distribution on h^2, which will be discussed later.

To facilitate the Gibbs sampling, data augmentation of the unobserved liability $\mathbf{U}$ is included as a latent variable, which we have already assumed to have a normal distribution, given $\boldsymbol{\beta}$ (Tanner and Wong, 1987; Albert and Chib, 1993). Then, the joint posterior distribution of the parameters and latent variables is given by:

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{U}, \mathbf{t}, \sigma_A^2, \sigma_E^2 | \mathbf{Y}) \quad \propto \quad & p(\boldsymbol{\beta}) p(\mathbf{t}) p(\mathbf{U} | \boldsymbol{\beta}, \sigma_A^2, \sigma_E^2) \times \\ & p(\sigma_A^2 | \alpha_A, \gamma_A) p(\sigma_E^2 | \alpha_E, \gamma_E) p(\mathbf{Y} | \mathbf{U}, \mathbf{t}), \end{aligned} \tag{3.3}$$

where most of these distributions have already been mentioned above, but the prior $p(\boldsymbol{\beta})$ follows a normal distribution with a variance of to 10^{10} and appropriate dimensions for the number of fixed effects (e.g. age, weight), the prior $p(\mathbf{t})$ for the thresholds is flat and improper, and $p(\mathbf{Y}|\mathbf{U}, \mathbf{t})$ is simply a vector of indicator functions of whether each Y_i falls into the category corresponding to the true value of U_i and $\mathbf{t}$. To improve convergence, a Metropolis-Hastings-within-Gibbs strategy is implemented in `MCMCglmm`, where $\mathbf{U}$ and $\mathbf{t}$ are updated jointly using a Metropolis-Hastings step at each iteration, $\boldsymbol{\beta}$ is sampled jointly from the entire vector's full conditional distribution, and σ_E^2 and σ_A^2 are each sampled independently from their individual full conditional distributions (Cowles, 1996; Hadfield, 2011).

3.2.4 Identifiability of variance components and heritability

In latent models with an ordinal response variable, individual variance components many not be identifiable (Harville and Mee, 1984; Mizstal et al., 1989; Luo et al., 2001; Stock et al., 2007; Ødegård et al., 2010). A common solution to this problem is to fix one of the variance components to a known constant c (e.g. $\sigma_E^2 = 1$). This solution is viable, because even when individual variance components are not identifiable, heritability – the main object of interest – may still be (Stock et al., 2007; Ødegård et al., 2010). In our case, fixing $\sigma_E^2 = c$ allows us to re-parameterize our model in terms of heritability instead of variance components, yielding the following posterior distribution:

$$p(\boldsymbol{\beta}, \mathbf{U}, \mathbf{t}, h^2|\mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{U}, \mathbf{t}) p\left(\mathbf{U}|\boldsymbol{\beta}, \sigma_A^2 = \frac{h^2 c}{1-h^2}, \sigma_E^2 = c\right) p(h^2), \tag{3.4}$$

where $p(h^2)$ is density of the Beta distribution, as discussed in the previous section.

Although the outlined approach to solving the identifiability problem is theoretically valid, in practice, fixing one of the variance components results in severe mixing problems of MCMC algorithms designed to approximate the posterior (3.4) (Ødegård et al., 2010). An alternative solution is to use MCMC to sample from the posterior of the unidentifiable model (3.4), but draw inferences based on only the posterior of heritability parameter, h^2. This latter approach can be viewed as MCMC with auxiliary variable augmentation of the state space, where σ_E^2 plays the role of an auxiliary variable. Using simulated data, we demonstrate that the auxiliary MCMC approach is superior in practice to the MCMC targeting the posterior (3.4), at least when using `MCMCglmm` package. Figure 3.1 shows traceplots of variance component(s) and heritability under both MCMC sampling schemes, using two different pedigree structures. For the first pedigree, fixing $\sigma_E^2 = 1$ results in such slow mixing that the Markov chain does not reach stationarity, while the auxiliary MCMC mixes very well, settling on the true value of heritability, which we set to 0.6 for both pedigrees. Using

the second pedigree and fixing $\sigma_E^2 = 1$, we observe possible stationary behavior of the heritability traceplot, but still very slow mixing with 1000 MCMC iterations corresponding to an effective sample size of 15. The auxiliary MCMC mixes much faster with 1000 MCMC iterations corresponding to an effective sample size of 615. These two examples and results of our extensive simulation study, outlined below, demonstrate that the auxiliary MCMC, even though unconventional, appears to work well in practice.

3.3 Data

3.3.1 Simulations

We simulate several datasets under a variety of conditions. The simplest scenario is that of a three generation pedigree shown in Figure 3.2, where trait data are simulated over 40 such distinct extended families, each of eight individuals: two unrelated founders with two children, each with an unrelated spouse and one child of their own. The trait data are simulated according to a multivariate normal distribution with mean vector determined by an additional covariate (e.g. age), and covariance structure dictated by the relationship matrix determined by this pedigree: that is, using the model in (3.2), $\mathbf{X}$ is a vector of ages which are in agreement with the real data when available, or simulated at random when unavailable, and β was set to a value of 1.5 to indicate a positive relationship between age and OST. Also, in concordance with (3.1), the unrelated parents have 0 covariance, each parent-offspring pair has a covariance of $0.5\sigma_A^2$; and the extended relationship pairs have covariances determined similarly.

Using this same pedigree, we also simulate data according to the threshold model. First, a latent variable is simulated according to a multivariate normal distribution with the same mean and covariance structure as described above. This is followed by discretization of the latent variable into categories. While we explore inference

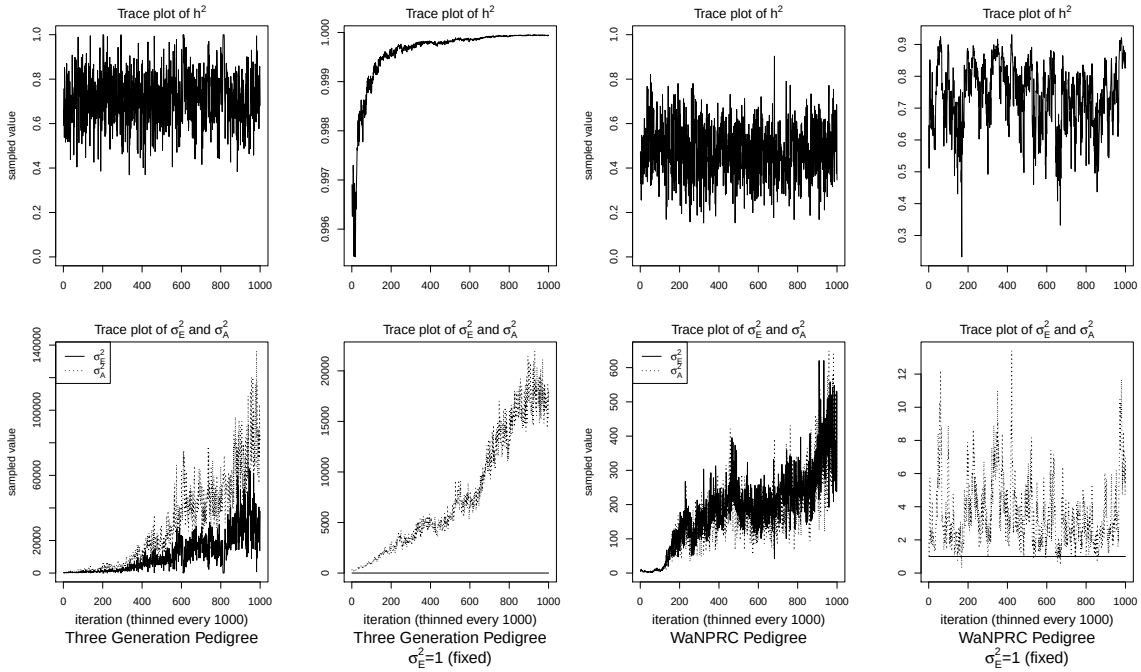

Figure 3.1: **Variance components and heritability traceplots**. Four scenarios are shown here, with traceplots of $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$ on top, and traceplots of the individual variance components on the bottom. The first scenario (shown in the first of the four columns of panels) is the three generation pedigree. While the MCMC samples of each individual variance component clearly do not show convergence (bottom), we observe that when we examine the corresponding values of h^2, this does appear to be stable (top). Conversely, when we fix $\sigma_E^2 = 1$, this does not appear to stabilize the MCMC samples of σ_A^2 here, and $h^2 \to 1$ as shown in the top and bottom panels of the second column. With the WaNPRC pedigree (third and fourth columns), we again observe that without fixing $\sigma_E^2 = 1$, the MCMC samples of h^2 does indicate convergence despite the fact that those for σ_A^2 and σ_E^2 individually do not. On the other hand, when fixing $\sigma_E^2 = 1$, we observe that σ_A^2 does not "blow up" like it did in the three generation pedigree case, but mixing appears to be poorer with regard to the traceplot of h^2. Indeed, in these 1000 MCMC samples, our effective sample size is 15, compared to 615 when σ_E^2 is not fixed to 1.

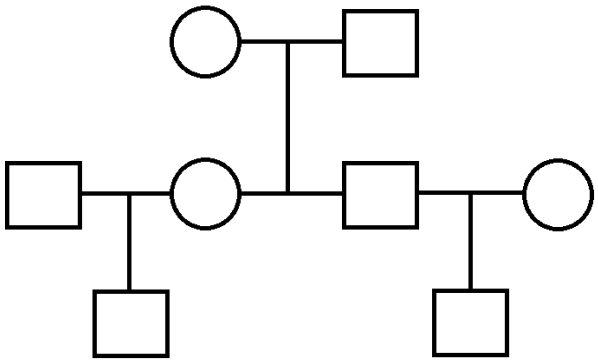

Figure 3.2: **Three generation pedigree**. The simpler scenario used for some simulations. Our simulated data consist of 40 repeated independent iterations of this pedigree structure, for a total sample size of 320.

with various numbers of categories, our primary interest is in a discretization into 10 categories, to mimic the actual data that we observed in the pig-tailed macaques. Specifically, the discretization is done in such a way to reflect the zero-inflated nature of our data. A graphical representation of this is shown in Figure 3.3.

We also consider the pedigree of our actual data of 542 pig-tailed macaques, with multivariate normal trait data simulated with covariance structure dictated by this pedigree structure. Again, we consider simulation of both a normally distributed trait, and a zero-inflated ordinal trait dictated by a normally distributed latent variable as per the threshold model. (again represented by Figure 3.3). Under each scenario, four "true" heritabilities are considered: $h^2 = 0.4, 0.6, 0.75, 0.90$. The number of simulated datasets for each value of heritability is 200.

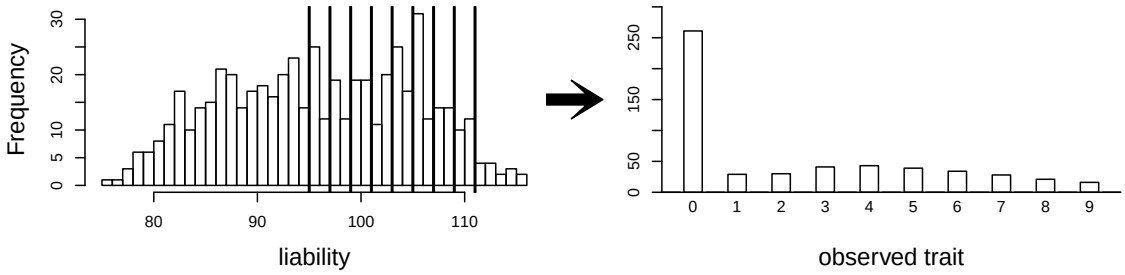

Figure 3.3: **Simulating a zero-inflated trait**. On the left-hand side is one simulated realization of a normally distributed liability trait, with cut-points shown for the transformation to the observed zero-inflated ordinal trait.

3.3.2 WaNPRC Pig-tailed Macaques

The study population consists of six generations of pedigree data for 542 pig-tailed macaques at the University of Washington National Primate Research Center (WaNPRC). Phenotypic data are available for 189 female monkeys present at the center in 2002, between the ages of 4.7 and 29.2 years old at that time with a mean of 10.11 years old. Younger monkeys are over-represented (with n=45 for monkeys between the ages of 4.7 and 6 years old), and older monkeys are under-represented (with n=12 for monkeys between the ages of 17 and 29.2 years old).

As a proxy for DDD, we measured osteophytosis (OST), also known as bone spurs. OST trait values for each monkey were determined through radiography at each of a total possible 16 intervertebral spaces through each monkey's spinal cord, and each space was recorded as 0, 1, 2, or 3 for unaffected, slight, moderate or severe bone changes, respectively. Details of the data collection and primate facility can be found in the study by Kramer et al. (2002).

From these raw data, there are a number of possible ways to summarize them into one number per monkey to use as the putative outcome trait. Perhaps the

most obvious choice, the simple sum of the values from all intervertebral spaces, was removed from consideration because each monkey had data from a different number of the 16 total intervertebral spaces recorded; thus, there would be an upward bias in this value corresponding to the monkeys which had more spaces recorded. Therefore, we choose to focus on a subsample of the intervertebral spaces for which a large majority of the monkeys had complete data. Specifically, with the seven intervertebral spaces from location L5 to T10, there are a total of 173 of the 189 monkeys with complete data on these spaces. We then combined adjacent categories that had less than three monkeys, to give a phenotype which has a total of 10 ordered categories.

3.4 Results

3.4.1 Simulations: Comparison of methods

The simulations were performed to assess both the consequences of assuming a normal distribution on an ordinal trait with normal liability, and also the performance of threshold model estimation under extreme discretization (e.g. our zero-inflated data). Under both pedigree structures, we first simulate a trait under multivariate normality with covariance structure dictated by the respective pedigree, and then perform heritability estimation of that trait under both maximum likelihood and Bayesian methods with a normality assumption. Results for the simulations under normality are shown on the left half of each panel in Figure 3.4. Next, we simulate a latent trait under multivariate normality again with covariance structure dictated by the respective pedigree, and then discretize the latent trait as described earlier. We then perform heritability estimation under both maximum likelihood and Bayesian methods, but now the Bayesian method assumes the threshold model via OPR, while maximum likelihood still assumes normality. The aim of this experiment is to illustrate the potential consequences of incorrectly assuming a normal distribution, when the trait actually follows the threshold model. Results are shown on the right half

of each panel in Figure 3.4. Also, trace plots for chains initialized using different starting points are shown in Figure 3.5 for one representative simulation scenario (WaNPRC pedigree with $h^2 = 0.60$), showing no sign of nonstationarity in each case. The starting values for σ^2_E varied from $(0.1, 1, 1000, 100000)$, and the starting values for σ^2_A varied from $(0.1, 1, 10)$ as indicated on the plots. Starting values for $\boldsymbol{\beta}$, $\mathbf{t}$ and $\mathbf{U}$ are obtained heuristically as described in (Hadfield, 2010).

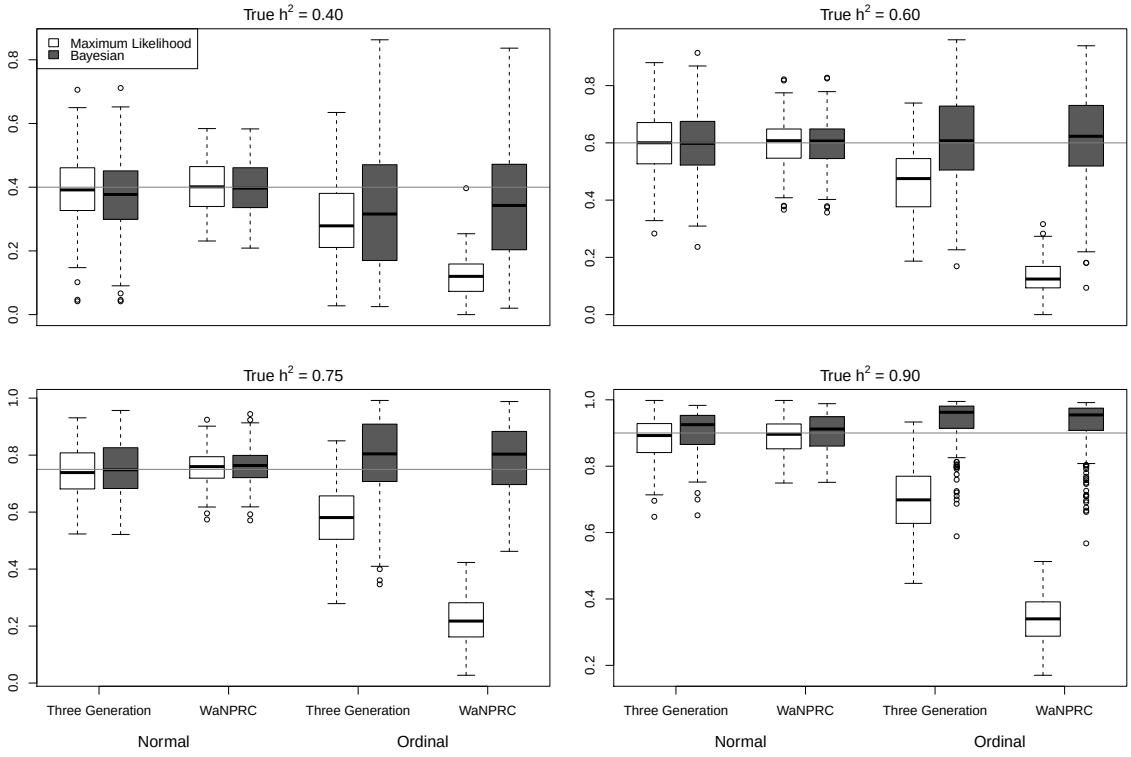

Figure 3.4: **Comparison between maximum likelihood and Bayesian methods**. Data were simulated both under normality (left half of each panel) and the threshold model (right half of each panel). Under normality, both maximum likelihood and Bayesian methods correctly assume normality. Under the threshold model, maximum likelihood still (incorrectly) assumes normality, whereas the Bayesian method correctly assumes the threshold model.

Under the scenarios with a normally distributed trait, maximum likelihood and

Bayesian estimations both show estimates that are centered around the true values of heritability. In the scenarios with an ordinal trait, maximum likelihood gives estimates that are quite far from the true values of heritability, tending to underestimate it severely. The Bayesian OPR performs much better under these scenarios, showing estimates that are closer to the true values. This is as expected, as the OPR in fact assumes the "correct" model under these simulations. In most of the scenarios, the medians of the heritability estimates from OPR are within roughly 5% of the true value used for the simulations. We do note that under the scenario with true $h^2 = 0.90$, the estimates are centered above the true value, close to 1. Examination of some trace plots showed that the chain for σ^2_E tended to be equal to exactly 0 for a substantial portion of the iterations, thus leading to sampled values of $h^2 = 1$ (not shown). It is thus possible that under such a high value of h^2, MCMC has a hard time approximating the posterior distribution of h^2.

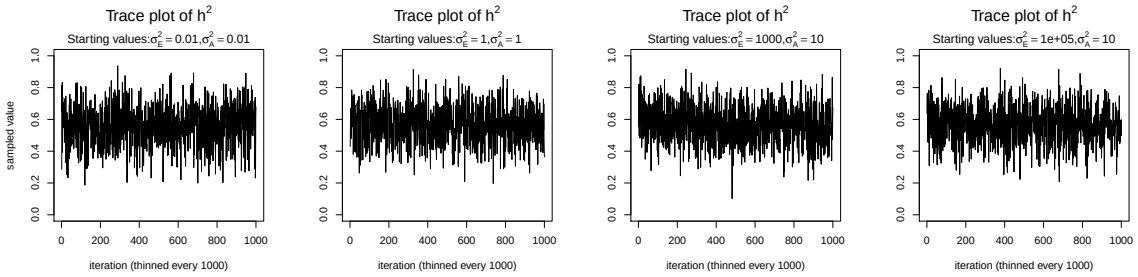

Figure 3.5: **Trace plots of heritability**. Chains for various starting values, for the scenario with $h^2 = 0.60$ using the WaNPRC pedigree.

3.4.2 Data Analysis: WaNPRC Pig-tailed Macaques

Descriptive statistics for the study population of pig-tailed macaques are shown in Table 1. Skewness in OST, age and parity are evident, as the mean is less than the center of the range in each case. For age, those between 5-6 years old are over-

Table 3.1: **Descriptive Statistics.** The mean, median, minimum, maximum, and standard deviation of each variable in the dataset are shown here.

	mean	median	min	max	sd
OST	1.64	0	0	9	2.79
Age (years)	9.83	7.40	4.70	29.20	5.08
Body mass (kg)	7.06	6.92	4.53	12.35	1.40
parity (#)	2.13	1	0	15	3.03

represented (n=40), and those between 18-29 are under-represented (n=11). For parity, 86 of the 173 monkeys had a value of 0.

The OST trait distribution is shown as the darkest bars in Figure 3.7 (the left-most of each value). All analyses were adjusted for age, mass and parity, according to results from a previous study indicating that these may be potential confounders of the association between genetic factors and OST (Kramer et al., 2002). The first two rows of Table 3.2 show maximum likelihood and Bayesian results from naively using the average OST value and assuming normality. The third row shows the result from using Bayesian ordered probit regresion on the ordinal phenotype described above.

Table 3.2: **Heritability estimates.** Adjusted for age, mass and parity. For maximum likelihood, CI = Confidence Interval; for Bayesian, CI = Credible Interval.

Trait	Model	$\hat{\sigma}^2_A$	$\hat{\sigma}^2_E$	h^2	95% CI
Average OST	ML normal	0.0394	0.0781	0.335	(-0.089, 0.760)
Average OST	Bayes normal	0.0400	0.0815	0.326	(0.0364,0.717)
Binary OST	Bayes OPR	$6.45 \cdot 10^8$	$8.31 \cdot 10^8$	0.430	$(1.70 \cdot 10^{-12}$,1)
Ordinal OST	Bayes OPR	$1.06 \cdot 10^{10}$	$4.53 \cdot 10^9$	0.700	$(5.56 \cdot 10^{-11}$, 1)

Maximum likelihood and Bayesian heritability estimates under the normality assumption are comparable (0.335 and 0.326 respectively). The Bayesian OPR on the ordinal trait shows a heritability estimate that is greater (0.700), but what is remarkable is that the estimated variance components are very large ($\hat{\sigma}^2_A = 1.06 \cdot 10^{10}$ and

$\hat{\sigma}^2_E = 4.53 \cdot 10^9$). An examination of the trace plots over MCMC generations suggested that the total variance may be unidentifiable (Figure 3.1). However, this also seemed to be the case in the ordinal simulations with both the three generation pedigree and WaNPRC pedigree, where the quantity of heritability was recovered successfully (as estimates tended to be centered near the true values, as shown in Figure 3.4). While this is of some technical concern, it thus seems more important for our current purposes to examine the posterior distribution of heritability as estimated from the MCMC. In our real data, we find that the posterior distribution simply reflects the information provided by the prior; that is, our estimation procedure was not able to extract substantial information from the data. This is shown in the left two panels of the top row of Figure 3.6. A similar posterior distribution of heritability was observed in the binary case (not shown). Also, results using different prior distributions are shown in subsequent rows of Figure 3.6. We observe that with $n = 173$ in either the real data or simulated case, the estimated posterior distributions tend to reflect the prior distributions. In some cases, mixing appears to be good, in the sense that the MCMC chain travels between 0 and 1 with no discernible pattern, such as with the Beta(0.01,0.01) or Beta(0.1,0.1) priors using the WaNPRC data. In other cases, mixing appears to be poor, such as with a Beta(0.2,0.2) prior using the WaNPRC data, or the Beta(0.01,0.01) prior using the simulated dataset with n=173; in these cases, the posterior distribution reflects one of the two modes of the prior distribution. These cases do raise uncertainty as to whether the posterior distribution is simply hard to estimate here, or if the posterior distribution truly contains no information about h^2. However, with increased sample size such in the three panels with $n = 542$, we obtain much stronger indications of stationarity of the MCMC chain in all cases, and unimodal posterior distributions of h^2, thus leading us to hypothesize that the true posterior distribution of h^2 contains more information about h^2 with larger sample sizes.

3.4.3 Sample size exploration

Since we were not able to extract any conclusive information from our data, we explored simulations to determine how much data would be necessary for heritability estimation under the threshold model. First, we simulate two extreme cases: 173 monkeys (identical to that of our real data), and the full 542 monkeys in the entire WaNPRC pedigree. In each case, a zero-inflated trait is simulated under the threshold model. Again, we focus on the posterior distributions of heritability, which are shown in the right two panels of Figure 3.6.

As shown, with 173 monkeys, threshold model heritability estimation typically produces estimated posterior distributions that mimic the prior distribution, even when the data are simulated according to the same threshold model that we are using for estimation. In contrast, with the full pedigree of 542 monkeys, threshold model heritability estimation succeeds in producing a spread of MCMC samples around our truth of $h^2 = 0.60$. We next aimed to determine the minimal number of monkeys required to estimate heritability under simulation.

Our criteria for labeling a particular sample size as having estimable heritability follows from our observations with the sets of 173 and 542 monkeys. That is, we examined the resulting estimated posterior distributions of h^2 at each sample size. Specifically, we checked the proportions of the estimated posterior distribution that fell into each of the 10 bins of size 0.1, from 0 to 1. Then, if the bins of $0-0.1$ and $0.9-1.0$ had the smallest proportion of mass from the estimated posterior distributions, we determined that the sample size had estimable heritability. In each such case, we also observed a unimodal posterior distribution with its mode near the simulated true h^2, so while our criteria only depends strictly on the decreasing tails of the posterior distribution, the result is that a sensible posterior distribution of h^2 indicates that h^2 is estimable.

To this end, we created subsets of the full WaNPRC pedigree, proceeding by

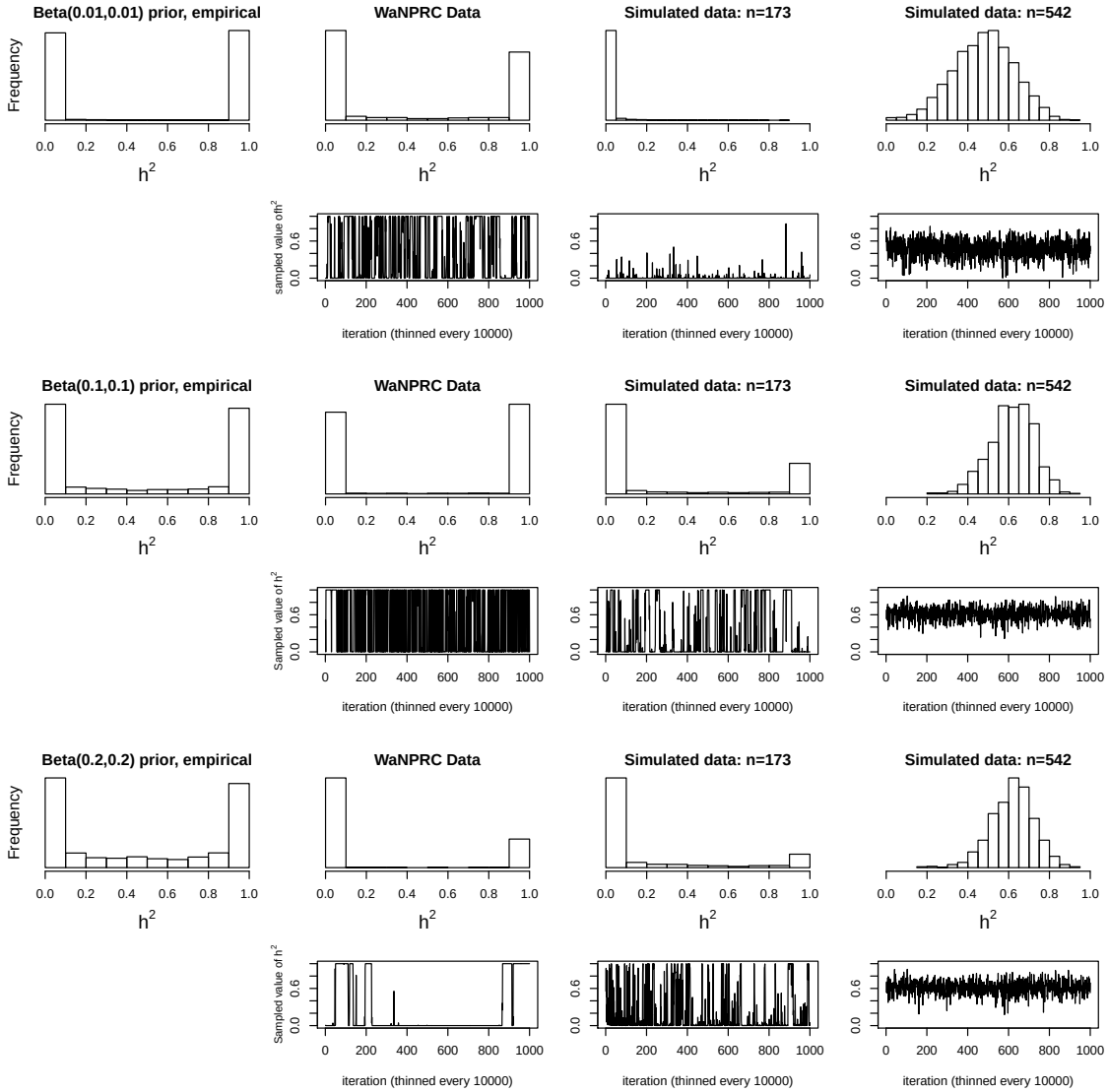

Figure 3.6: **Distributions of heritability**. Three scenarios with different prior distributions are shown consecutively, with two rows of panels for each scenario. Within each scenario, the first panel shows an empirical realization of the prior distribution of heritability, according to inverse-gamma prior distributions on each of the individual variance components. The second panel shows the posterior distribution of heritability from the real data analysis. The third and fourth panels show the posterior distributions of heritability from 173 simulated monkeys and 542 simulated monkeys, respectively. The bottom panels show trace plots of heritability, thinned to 1000. Simulated heritability was 0.60 in each case.

starting with the original 173 monkeys and adding the most related monkeys to that set, based on cumulative pairwise kinship coefficient. That is, the candidate monkey who is the "most related" to the current set would be the one who has the greatest sum of kinship coefficients with each monkey in the set, and is not currently in the set itself. Also, 28 of the 173 monkeys actually are not related to any of the others in this set, so these were first removed. We then added monkeys based on the maximum kinship criteria to create larger subsets of monkeys (e.g. n=200, n=210, etc.), and proceed with our simulations as if these were the monkeys for which we had data. We note that with sample sizes for which h^2 appeared to be estimable, stationarity of the MCMC was typically observed within roughly 1 million iterations, at which point the above criteria for estimable heritability was always satisfied. For sample sizes in which the posterior distribution did not satisfy our criteria for estimable heritability, the trace plot for h^2 would typically appear similar to the prior distribution of h^2, with trace plots showing no sign of nonstationary behavior by the MCMC chain, as it bounces back and forth between 0 and 1 (such as in select panels of Figure 3.6). Additionally, when we ran certain scenarios with insufficient sample sizes for up to 200 million iterations, the trace plots for h^2 appeared the same as at 1 million iterations, adding further evidence that the chain's repeated jumping from 0 to 1 exhibits its stationary behavior. This suggests that our MCMC appears to be providing a good approximation of the true posterior in both cases when the sample sizes lead to estimable h^2, and when sample sizes are low, with the true posterior not containing much information about h^2.

Additionally, we wanted to explore the effect of attenuation on the degree of zero-inflatedness in our trait distribution, and whether a less extreme distribution may lend itself to better heritability estimation. This has direct relevance to our real OST phenotype, as it is a trait which manifests itself gradually over the lifespan of monkeys: in an older sample of monkeys we expect to see a less zero-inflated trait distribution. By simply increaing the value of our age covariate in our simulations by five years for

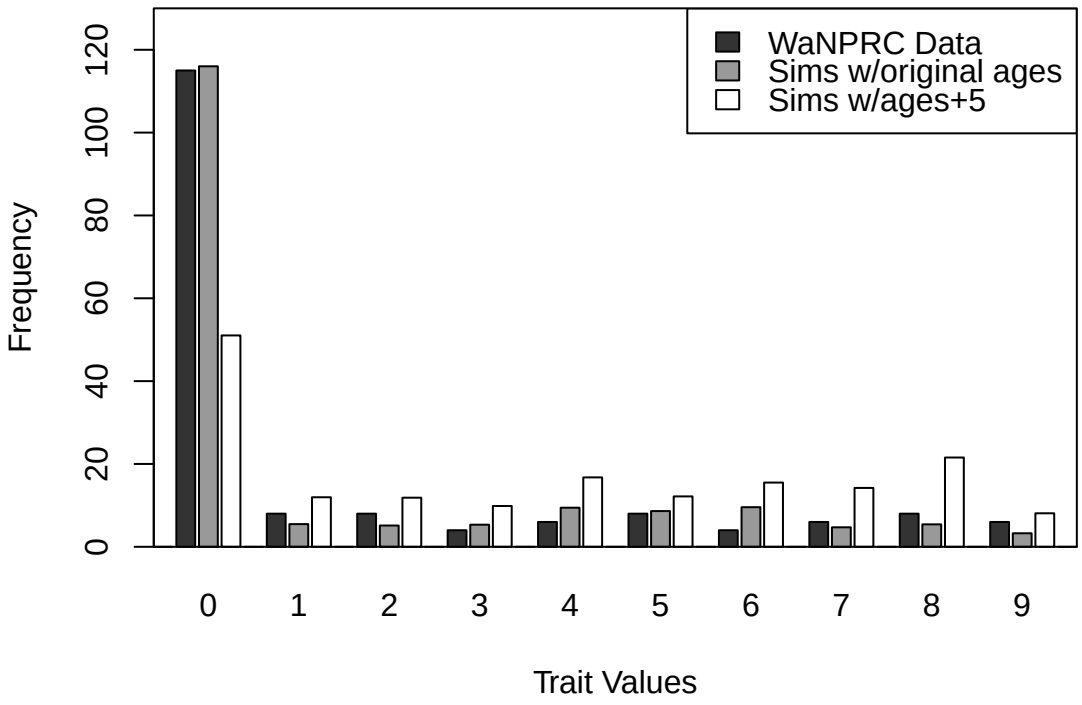

Figure 3.7: **Distributions of the real and simulated OST phenotype, with age shifts**.

each monkey, we obtain this effect. An illustration of the trait distribution resulting from the five-year age increase is shown in Figure 3.7, with empirical averages from 100 datasets for the simulated cases. Based on the posterior distribution histograms of heritability (not shown, but similar to the right two panels of Figure 3.6), we determine whether there was enough information in the simulated data for each case.

Alternatively, we also examine phenotype data from another population of monkeys, in the Wisconsin National Primate Research Center. These monkeys are older than our WaNPRC center monkeys, with a mean age of 21.55 years old. Therefore, almost all of the monkeys have exhibited some degree of the OST trait and there is no zero-inflatedness. We perform simulations with a trait distribution that mimics this, to again determine what sample size is required for estimable heritability.

Although it has less relevance to our primary WaNPRC data, we also explore whether having phenotype data on a different subset of monkeys than the original 173 may be more optimal, with regards to heritability estimation. That is, thus far we have merely added additional monkeys to the original set in which our real dataset has phenotype data. However, these original monkeys are not all highly related to each other, which provides less information for heritability estimation than if they were all highly related. Thus, it is also of interest to know whether a smaller sample size would be necessary to obtain estimable heritability under a more related set of monkeys. We therefore sample monkeys based on maximum cumulative kinship coefficient starting from the single monkey which is the most related to all other monkeys under this criteria, and add monkeys as before. We simulate data under both the original trait distribution, and that of the Wisconsin dataset.

Finally, we explore sample size requirements under the more simple three-generation pedigree, i.e. the same one as in our previous simulations shown in Figure 3.4. These previous simulations were performed with a sample size of 40 families, or 320 individuals. We find that we can reduce the sample size to 20 families, or 160 individuals, and still obtain reasonable heritability estimation through the threshold model. Also,

with a trait distribution that is less zero-inflated (again as through an increased age by 5 years), not much improvement is obtained; we can further reduce the sample size by just one family, to 19 families or 152 individuals. A summary is shown in Table 3.3.

Table 3.3: **Minimum sample size required for estimability of heritability under the threshold model**.

Pedigree	Phenotyped	Trait Distribution	Min. Sample Size
WaNPRC	Original	Original	250
WaNPRC	Original	Age + 5 years	230
WaNPRC	Original	Wisconsin	250
WaNPRC	Optimal	Original	190
WaNPRC	Optimal	Wisconsin	180
Three Generation	All	Original	160
Three Generation	All	Age + 5 years	152

3.5 Discussion

Here, we examine heritability estimation of an ordinal trait. Our ultimate aim is to determine whether osteoarthritis is heritable, and we explored a number of modeling considerations that take account of the ordinal nature of the data that were collected. We discovered that heritability estimates can vary greatly based on the choice of model, from both our simulation study and our real data analysis. In our WaNPRC macaques, under the naive assumption of normality of the average OST value, we observed an estimate that indicates a slight-to-moderate amount of heritability (0.335 under maximum likelihood estimation). This is also observed in the Bayesian estimate, under the same model (and with non-informative priors).

However, our simulations illustrate the degree to which inference can be biased, if normality is assumed when the data actually follow the threshold model. Ordered Probit Regression was able to obtain heritability estimates that were centered closer to

the true value in each case than maximum likelihood estimates under the normality assumption. While it is no surprise that Ordered Probit Regression was able to obtain good estimates from these datasets since they were simulated under the exact model that the Ordered Probit Regression assumes, it is more to the point that using a standard maximum likelihood approach with an assumption of normality yielded estimates that were quite far from the true values, even when the number of categories was large (e.g. 10 in two of the scenarios).

These scenarios were also designed to mimic a plausible imitation of our real data, in the fact that most of the monkeys (115 out of 173) were "normal" with respect to the second OST trait. We simulated the liability trait and then put the bulk of the data into the first category in order to attain a similar distribution of the observed trait. Whether or not this model exactly reflects the biological mechanism of the OST trait, these simulations nevertheless illustrate that incorrectly assuming normality of an ordinal trait invites the risk of producing misleading heritability estimates, while Ordered Probit Regression has a better chance of producing estimates that are closer to the truth. Furthermore, while it is true that we do not know whether our actual data follow the threshold model, this assertion could be justified by applying the polygenic model to the liability; that is, even if what we observe is ordinal with a very non-normal distribution, it is defensible to assume that, if the trait is determined by many loci, there may be an underlying latent variable which does have an approximately normal distribution.

Thus, it is interesting that our heritability estimate rises to 0.700 under estimation with the threshold model. However, there are several alarming aspects to this: 1) the estimates of the individual variance components are very large; 2) the 95% Credible Interval spans essentially the entire range of (0,1); 3) the posterior samples of heritability almost exactly mimic its prior distribution. These observations suggest that the information content of our data is not high, which may be surprising given that we do have 173 monkeys with trait data. However, as our regression setting here

is a non-standard one, we find it useful to perform simulations to explore how much is required to obtain estimable heritability.

For the sake of its direct relevance to our real data, we first examine the effect of increasing the sample size on our actual WaNPRC pedigree. Our original sample of phenotyped monkeys was a convenience sample that was not specifically intended for heritability estimation, and many of the monkeys which were not originally phenotyped are still alive and could still be obtained. Obtaining these data from another 80+ monkeys, however, is non-trivial, and we are still investigating this possibility.

It is somewhat surprising that we do not gain much improvement in sample size reduction with a more balanced trait distribution that was induced by shifting the age distribution. It is possible that there are nuances in our simulated trait distributions which are causing difficulties that we do not understand, particularly because all of our threshold locations were placed in an *ad hoc* manner, simply to create trait distributions that appeared reasonable. On the other hand, it is also possible that the threshold model does not inherently struggle with zero-inflated data (at least when such data truly arose from the threshold model itself), and so an improvement is not to be expected with less zero-inflated data. This possibility is corroborated by the fact that, using the Wisconsin trait distribution, we also see limited and/or no improvement to sample size requirements, depending on the set which was phenotyped. It is thus interesting to note that the actual trait distribution seems to be far less of a factor than the set of monkeys for which phenotype data are available, in terms of obtaining estimable heritability.

Of the previously mentioned alarming aspects to our heritability estimate on our real data, the one that our simulations does not address is that of the extremely high estimates of the individual variance components. In fact, this seems to be a recurring observation even when the posterior distribution of heritability appears to be well-behaved. While we are reasonably satisfied to simply obtain sensible posterior distributions of heritability from this implementation of the threshold model, this

suggests that the individual variance components in fact are not identifiable in our scenarios. A resolution to this concern is a possibility for further study.

Chapter 4

UNTANGLING CONVERGENT EVOLUTION AND RECOMBINATION

4.1 Introduction

Biologists are often interested in investigating evolutionary relationships between different species. The field of phylogenetics aims to describe these evolutionary relationships, by inferring a phylogeny, or evolutionary tree, using data from the species of interest. While the phylogeny itself may often be of interest, many important scientific questions can be answered using phylogenetics. For example, phylogenetics has been instrumental in recent studies involving viral evolution: to trace the origin of avian influenza (Lemey et al., 2009), and to help convict two suspected rapists of intentionally infecting their victims with HIV (Scaduto et al., 2010).

While any attribute that accumulates changes over time can technically be used to infer a phylogeny, modern phylogenetics typically relies on molecular sequence data, e.g. DNA or amino acid (Yang and Rannala, 2012). In the estimation of a phylogenetic tree from molecular sequence alignment data, the absence of recombination is frequently assumed: that every site along the sequence alignment has the same evolutionary history. However, if recombination is in fact present, it has direct implications on not only tree estimation (Posada and Crandall, 2002; Felsenstein, 2004), but also on all downstream analyses, such as the detection of selection (Anisimova et al., 2003). To illustrate why, suppose there is one recombination breakpoint within a sequence alignment. This suggests that the DNA bases before the breakpoint have been inherited according to one particular tree, and the DNA bases after the breakpoint have been inherited according to another. Phylogenetic estimation procedures that ignore

the recombination in this example may produce misleading results, such as inferring only one of the true underlying trees, inaccurately estimating branch lengths, and overestimating substitution rate heterogeneity (Schierup and Hein, 2000).

The important implications of ignoring recombination in phylogenetic inference have motivated the development of tests for the presence of recombination (Posada and Crandall, 2002; Awadalla, 2003). The most extensively used methods for recombination detection include phylogenetic methods that examine estimated phylogenies from different portions of the sequence alignment to see if these phylogenies differ; if so, this would be evidence that recombination had occurred (Posada and Crandall, 2001). Among these phylogenetic recombination detection methods, one popular approach is to use a "sliding window," so-called because adjacent, overlapping windows of the DNA sequence alignment are examined sequentially across the alignment, to look for phylogenetic disagreement between windows (Grassly and Holmes, 1997; McGuire et al., 1997; Husmeier and Wright, 2001).

One potential drawback of any recombination detection technique is that selective pressures may produce a similar signal to that of a true recombination. For example, suppose that selective pressure acts upon two taxa to make them appear more closely related to each other than they are under the true evolutionary history, such as the development of wings in both birds and bats, which are not closely related, evolutionarily. This phenomenon is known as convergent evolution (Wake et al., 2011). Now, if convergent evolution occurs between these two sequences only at a localized region of the alignment, then it will appear as if this region has a different evolutionary history than the remainder of the alignment. This could occur, for example, if one gene is critical for the survival of the organism in a particular environment, so this particular gene might undergo convergent selective pressure whereas its surrounding genes may not. Thus, under this scenario, the sequences will appear similar to how they would if one of the sequences had simply given a portion of DNA to the other sequence, i.e. if recombination had occurred. Any recombination detection algorithm might,

therefore, falsely identify a recombination event when in fact the conflicting evolutionary histories are due to convergent evolution. Our aim in this work is to introduce a fundamental modification to one of the recombination detection methods that will avoid falsely identifying recombination in the presence of convergent evolution.

As a starting point, we consider the Dss method proposed by McGuire et al. (1997) and implemented in the TOPALi software (Milne et al., 2004). Dss, an abbreviation for "difference in the sum of squares," is a sliding window approach that scans across the sequence alignment in question, with the assumption that if a recombination breakpoint is present within any given window, then the portions of the window on opposite sides of the breakpoint would have distinct evolutionary trees. The test statistic produced by the Dss method is based upon the pairwise distance matrices for each half of every window across the alignment, and an extreme value of the statistic is expected to occur in a window that contains a recombination event at its center.

Our proposed modification is to base the test statistic on a measure of evolutionary distance that considers only synonymous substitutions: the codon changes that do not result in amino acid changes. Each of the 20 amino acids are coded by a triplet of nucleotides; since there are 64 possible triplets (4^3), this directly implies the redundancy that is observed in the genetic code as almost every amino acid has more than one codon which codes for it. Thus, sometimes a change in codon does not result in a change in amino acid. The codon changes that do result in an amino acid change are known as nonsynonymous substitutions.

Assuming that selection acts on the amino acid level, it follows that selection has an effect only on nonsynonymous substitutions. That is, if selective pressures favor a particular amino acid over another in a given protein sequence, then by definition this means that at the nucleotide level, it favors a particular nonsynonymous substitution. On the other hand, since synonymous substitutions do not correspond to amino acid changes, then selective pressures on the amino acid level could not possibly favor a

particular synonymous substitution. However, synonymous substitutions do provide information about evolutionary relationships of sequences under study (Lemey et al., 2005; Yang, 2006; O'Brien et al., 2009). Thus, we postulate that using a distance metric which counts only synonymous substitutions within the Dss framework would still allow for recombination detection, but will avoid the false positives resulting from convergent evolution. We develop a new statistic and a novel parametric bootstrap method to access the distribution of this test statistic under the null hypothesis of no recombination.

To test our new recombination detection method, we first proceed via simulations to compare its performance to the original Dss statistic, both in terms of their ability to identify true recombination events, and to avoid false positives due to convergent evolution. We also examine two real data examples. The first is an HIV dataset, which comes from nine Belgian patients that belong to a known HIV transmission chain (Lemey et al., 2005). This dataset has been of particular interest because phylogenetic reconstructions can be compared to the known transmission chain, providing a real data example in which estimation procedures can be validated. In their work, Lemey et al. (2005) studied two distinct HIV genes: *pol* and *env*, and concluded that the *pol* gene was under convergent selective pressures, whereas the *env* gene was not. Here, we revisit this question in a coherent fashion with our method, by examining a concatenated alignment of the *pol* and *env* genes. Our second real data example comes from the evolutionary study of *Salmonella* – a genus of bacteria that is responsible for various illnesses, including typhoid fever and food poisoning (Groisman and Ochman, 1997). Its FimH adhesin, an adhesive protein responsible for both cell adhesion and invasion, is believed to thus be useful in the study of *Salmonella* adaptive evolution (Bäumler et al., 1997; Naughton et al., 2001; Boddicker et al., 2002; Althouse et al., 2003). Previous work on FimH adhesin of the *Salmonella* species *Salmonella enterica* found evidence of convergent selective pressures on this gene (Kisiela et al., 2012). We revisit this dataset with our Dss statistics to compare our results with theirs.

4.2 Methods

4.2.1 Evolutionary Distances

Let $\mathbf{Y}$ be a matrix that represents a DNA sequence alignment, composed of row vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where n is the number of taxa/species/sequences. Then, $\mathbf{y}_k = (y_{k1}, \ldots, y_{kL})$, where L is the length, or number of sites in the sequence alignment. For a given DNA sequence alignment, one common statistic of interest is the distance matrix, whose elements are the distances between sequences of each pair, e.g. $(\mathbf{y}_k, \mathbf{y}_l)$ for $k, l \in (1, \ldots, n)$. Intuitively, each pairwise distance simply indicates how different two sequences are from each other. For example, two sequences that are identical at every site along the alignment would have a distance of 0, under most sensible measures of distance.

In principle, evolutionary distances can be either model-free or model-based. For example, the Hamming distance, perhaps the simplest of all distances, is just the fraction of discordant sites between two sequences and hence is model-free. Unfortunately, its simplicity is overshadowed by the fact that Hamming distances are not additive, since substitutions can be hidden when multiple substitutions occur at a given site (Felsenstein, 2004; Yang, 2006).

In contrast, model-based distances assume a substitution model defined by the rates of change between the possible states. For a pair of taxa, the evolution of each site (y_{ks}, y_{ls}) for $s \in (1, \ldots, L)$ can thus be described by a continuous-time Markov chain (CTMC) with infinitesimal generator $\mathbf{\Lambda} = \{\lambda_{ij}\}$ for $i, j \in (1, \ldots, M)$, where M is the number of states (e.g. for DNA nucleotide data, $M = 4$ as the state space is $\{A, C, G, T\}$; for DNA codons, $M = 64$ as the state space is $\{A, C, G, T\}^3$), and the rate of leaving state i is $\lambda_i \equiv \sum_{j \neq i}^{M} \lambda_{ij}$. The simplest DNA nucletide mutational model is the Jukes-Cantor model, typically abbreviated as JC69 (Jukes and Cantor, 1969), which assumes that the substitution rates between any pair of distinct nucleotides are equal. With stationary distribution $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_M)$, we then can calculate distances

at each site as

$$\hat{d}_{kl} = \sum_{i=1}^{M} \hat{\pi}_i \hat{\lambda}_i, \tag{4.1}$$

with the necessary parameter estimates being calculated from the data. As this quantity is equal to the average number of jumps in a stationary continuous-time Markov chain, evolutionary distances are thus defined as the expected number of substitutions per site, according to the given continuous-time Markov chain model.

Specifically, to calculate $\hat{d}_{kl}$ as in (4.1), $\boldsymbol{\pi}$ is often estimated as the empirical state frequencies (unless it is known from the model, such as in JC69 where $\boldsymbol{\pi} = (1/4, 1/4, 1/4, 1/4)$). Next, we use the finite-time transition probabilities $\mathbf{P} = \{p_{ij}(t)\} = \exp(\boldsymbol{\Lambda} t)$ to write the likelihood of the data as

$$\mathbf{Pr}(\mathbf{y}_k, \mathbf{y}_l | \boldsymbol{\Lambda}, t) = \prod_{s=1}^{L} \hat{\pi}_{y_{ks}} p_{y_{ks}, y_{ls}}(t). \tag{4.2}$$

Noting that the infinitesimal generator $\boldsymbol{\Lambda}$ and time t appear in the transitions probabilities only as the product $\boldsymbol{\Lambda} t$, they are not both individually identifiable. Thus, one can either impose constraints on either $\boldsymbol{\Lambda}$ or $t = 1$ to obtain identifiability. Here, we will set $t = 1$ and estimate $\boldsymbol{\Lambda}$, which can then be done by maximizing the likelihood in (4.2). Alternatively, with some models, closed-form expressions for distances can also be obtained with the method of moments (see Section 2.1.3).

4.2.2 Least Squares Phylogenetic Inference

With an estimated distance matrix, one can then infer a phylogeny. A phylogeny is an object that describes the relationships between taxa over evolutionary time, and consists of a topology $\boldsymbol{\tau}$ and branch lengths $\mathbf{b}$. The topology refers simply to the shape of the tree (branching order), and in this work, we will consider only the class of topologies known as unrooted (lacking any indication of historical ancestry of the

taxa) and bifurcating (each split results in exactly two new lineages). The branch lengths refer to the evolutionary distance between two nodes, where a node can be either an observed taxon at the tips of the topology, or an internal and possibly unobserved taxon.

The tree distance $t_{kl}(\boldsymbol{\tau}, \mathbf{b})$ between taxa k and l is then the sum of the branch lengths between them on any particular topology. With $\hat{d}_{kl}$ estimated from DNA sequence data as above, least squares phylogenetic inference then proceeds by finding

$$\underset{\boldsymbol{\tau},\mathbf{b}}{\operatorname{argmin}} \sum_{k,l} \left(\hat{d}_{kl} - t_{kl}(\boldsymbol{\tau}, \mathbf{b})\right)^2, \tag{4.3}$$

which is the usual least squares criterion. The solution $(\hat{\boldsymbol{\tau}}, \hat{\mathbf{b}})$ to (4.3) then gives the least squares phylogeny.

4.2.3 Distance-Based Recombination Detection: Dss Statistic

Evolutionary distances and least squares phylogenetic inference form the basis of the Dss method for the detection of recombination. Here, we define recombination as an exchange of genetic material between two taxa that results in different evolutionary histories for the different respective parts of the sequence alignment. The method based on the Dss statistic then uses a sliding window approach (McGuire et al., 1997; McGuire and Wright, 2000; Milne et al., 2004), as illustrated in Figure 4.1, where the two panels show a window (in red) moving across a sequence alignment.

First, the average of all estimated pairwise distances from the entire sequence alignment is recorded as $\overline{d}$. Next, the distance matrix is estimated with a DNA substitution model for the first half of a given window, along with its mean, $\overline{w}^{\{1\}}$. This distance matrix is then standardized by multiplying each entry by $\overline{d}/\overline{w}^{\{1\}}$, and the resulting standardized distance matrix for the first half of the window is recorded

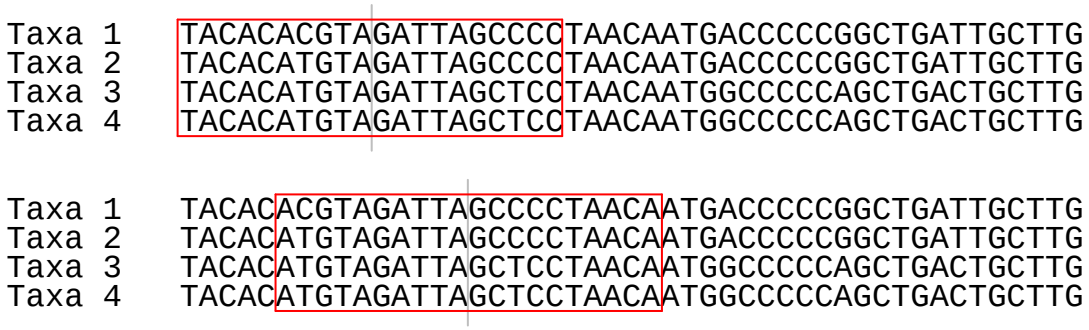

Figure 4.1: **Illustration of a sliding window across a sequence alignment of four taxa**.

as $\hat{\mathbf{d}}^{\{1\}}$. Using phylogenetic least squares as described above, we then calculate

$$\underset{\boldsymbol{\tau},\mathbf{b}}{\operatorname{argmin}} \sum_{k,l} \left[(\hat{d}_{kl}^{\{1\}} - t_{kl}(\boldsymbol{\tau}, \mathbf{b}) \right]^2. \tag{4.4}$$

The estimated topology is recorded as $\hat{\boldsymbol{\tau}}^{\{1\}}$, and the minimized value of the sum of squares in (4.4) is recorded as SSa_w^F.

For the second half of the window, again the distance matrix is estimated, with its mean stored as $\overline{w}^{\{2\}}$ and again the standardized distance matrix is calculated by multiplying each entry by $\overline{d}/\overline{w}^{\{2\}}$ to obtain $\hat{\mathbf{d}}^{\{2\}}$. Now, we calculate

$$SSb_w^F = \min_{\mathbf{b}} \sum_{k,l} \left[\hat{d}_{kl}^{\{2\}} - t_{kl}(\hat{\boldsymbol{\tau}}^{\{1\}}, \mathbf{b}) \right]^2. \tag{4.5}$$

That is, the topology from the first half is imposed as fixed in (4.5), and only the branch lengths are optimized according to the sequence alignment of the second half of the window.

For each window, we then have $Dss_w^F = (SSa_w^F - SSb_w^F)$. The entire procedure is repeated in the reverse direction, by starting with a window at the end of the alignment, swapping the roles of each half of the window, and then sliding it backwards

across the sequence alignment; this gives Dss_w^B for each window. Then, $Dss_w = \max(Dss_w^F, Dss_w^B)$. Finally, here we consider only the maximum Dss statistic from all windows, giving us $Dss_{max} = \max_w(Dss_w)$.

To assess statistical significance, McGuire and Wright (2000) propose a parametric bootstrap to generate the null distribution of the test statistic. Under this parametric bootstrap, the distribution of Dss_{max} is simulated under the null hypothesis as follows: first, the OLS tree for the entire sequence alignment is obtained, under the chosen model of substitution. In this manner, the data are treated as if the sequences were inherited through one true tree and substitution model, in accordance with the null hypothesis. Next, sequence data are simulated under this tree, B times (here, $B = 100$ for the simulation studies, and $B = 500$ for the real data analyses). Finally, the Dss values are calculated under the same procedure as outlined above for each simulated sequence alignment, and saving only the maximum from each simulated realization. Thus, one obtains the distribution of the maximum Dss statistic under the null hypothesis. This gives the basis for determining how extreme an observed Dss statistic is, and we calculate the Monte Carlo estimate of the p-value as the proportion of simulated null Dss values that are more extreme than our observed value in question. Important modifications to the parametric bootstrap for our implementation are discussed in the next section.

In Figure 4.2, we show output from a Dss analysis on a simulated sequence alignment with a recombination event (indicated by the vertical red line). This plot shows the value of the Dss statistic at every window across the sequence alignment. To complete the analysis, we would simply look at the maximum value across the alignment, and check whether this value is above the desired significance level threshold (here, 95%).

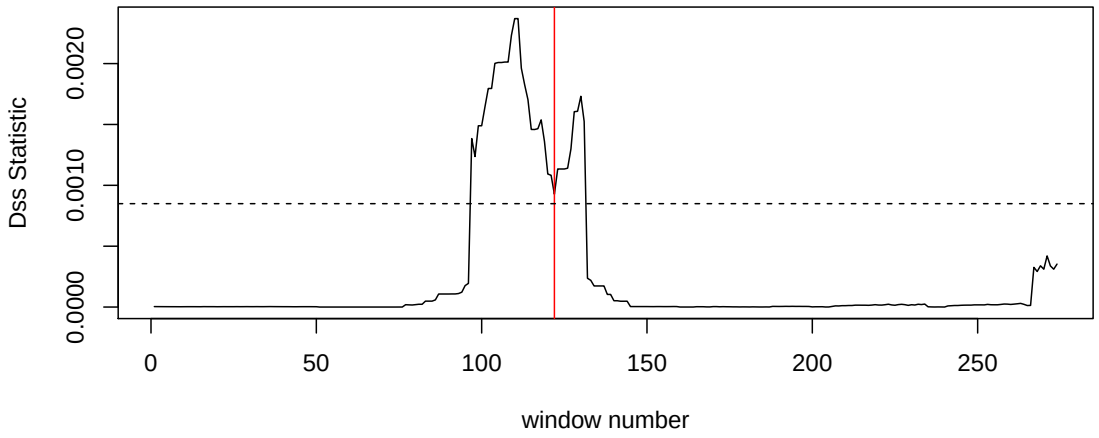

Figure 4.2: **Output from one representative analysis using the Dss statistic on a sequence alignment with a recombination event**. The vertical red line indicates the location of the simulated recombination event. The horizontal dotted line represents the 95% significance level based on the parametric bootstrap as described above.

4.2.4 Labeled Distances

The notion of an evolutionary distance can be generalized to consider certain subsets of substitutions. For example, it is sometimes of biological interest to count only transitions ($A \rightleftharpoons G$ and $C \rightleftharpoons T$), or transversions ($A \rightleftharpoons T$, $A \rightleftharpoons C$, $G \rightleftharpoons T$, and $G \rightleftharpoons C$). A variety of *ad-hoc* strategies could be used to account for this (Felsenstein, 2004), but it can also be formally incorporated into the framework of CTMC models of DNA evolution as was done by O'Brien et al. (2009). First, we define the set $\mathcal{L}$ to be the subset of the lattice $\{1, \ldots, M\}^2$ that indicates the substitutions which we wish to count; that is, $(i, j) \in \mathcal{L}$ if $i \to j$ is a substitution of interest. Then, we can

express distances for any labeled subset of substitutions as

$$\hat{d}_{\mathcal{L}} = \sum_{i=1}^{M} \hat{\pi}_i \sum_{j \neq i}^{M} \hat{\lambda}_{ij} 1_{(i,j) \in \mathcal{L}}, \tag{4.6}$$

for each pair of sequences (k, l). The indicator function $1_{(i,j)\in\mathcal{L}}$ in (4.6) is equal to 1 if $i \to j$ is a substitution of interest, and 0 if it is not. In this manner, the labeled distance metric does not count the substitutions that are not of interest. In this work, we consider the labeled subset known as synonymous substitutions: the changes in codon state which do not result in a change in amino acid.

Specifically, we propose a modification of the Dss statistic which uses estimates of labeled distances for synonymous substitutions, in the place of $\overline{d}$ and $\hat{d}_{kl}^{\{\cdot\}}$ in the calculation of Dss_{max} above. Recall from Section 4.2.3 that our definition of recombination is an exchange of genetic material, resulting in different evolutionary histories for different parts of the sequence alignment. The current Dss method tests for discrepant phylogenies throughout windows across the sequence alignment. However, it cannot distinguish between the case when the discrepancies are due to an exchange of genetic material, as opposed to convergent selective pressure. Here, our null hypothesis is that the sequence alignment has one true evolutionary history that has been affected neither by recombination nor convergent evolution. The current Dss method detects departures from the null hypothesis due to either recombination or convergenet evolution, or due to the presence of both of these events. With our new synonymous Dss statistic, our null hypothesis remains the same, but we hope that our new test will detect only departures from the null that are due to an exchange of genetic material. Under the assumption that selective pressure acts on the amino acid level, synonymous substitutions are presumed to be neutral, and therefore distances based upon them would ignore selective pressures. In this manner, potential false positive signals for recombination due to selection can be avoided.

Further, the manner in which we calculate these distances is robust to model

misspecification. Appealing to the fact that distances are the expected number of substitutions per site, we thus have $d_{\mathcal{L}} = \mathbb{E}(N_t^{\mathcal{L}})$, where $N_t^{\mathcal{L}}$ is the number of labeled substitutions in time t. The main idea, then, is that predicted probabilities of site patterns are replaced with the empirical frequencies of each state; in other words, conventional evolutionary distances implicitly assume that $P(y_{ks} = i, y_{ls} = j) = \pi_i p_{ij}(t)$. Instead, O'Brien et al. (2009) note that this can be replaced with empirical frequencies to obtain $P(y_{ks} = i, y_{ls} = j) = (1/L)\sum_{s=1}^{L} 1_{\{y_{ks}=i, y_{ls}=J\}}$, which provides partial freedom from the model-based assumptions of the continuous-time Markov chain. Thus we arrive at the robust estimate: $\hat{d}_{\mathcal{L}}^{robust} = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}(N_{\mathcal{L}} | X_0 = y_{1l}, X_1 = y_{2l}; \hat{\mathbf{\Lambda}})$.

Since these expectations are not directly observable from the data, it can be shown that we can compute them given $(\hat{\boldsymbol{\pi}}, \hat{\mathbf{\Lambda}})$, which are obtained empirically (Ball and Milne, 2005; Minin and Suchard, 2008). To compute them for our synonymous substitutions, we first note that we are working in codon state space. Fitting standard codon models (Goldman and Yang, 1994; Muse and Gaut, 1994) require computationally costly and numerically unstable optimization. We instead use a composition of separate nucleotide models for each of the three codon positions in order to efficiently estimate $\boldsymbol{\pi}$ and $\mathbf{\Lambda}$. To calculate $\hat{\boldsymbol{\pi}}$, we find empirical nucleotide frequencies at each of the three positions, and multiply them together to obtain estimates for each of the 64 codons. $\hat{\mathbf{\Lambda}}$ is calculated as the Kronecker sum of the infinitesimal generators from each position; that is, $\hat{\mathbf{\Lambda}} = \hat{\mathbf{\Lambda}}^{(1)} \oplus \hat{\mathbf{\Lambda}}^{(2)} \oplus \hat{\mathbf{\Lambda}}^{(3)}$, where the superscripts refer to each of the three positions of any given codon. Specifically, we use an F84 model (Felsenstein and Churchill, 1996) for each codon position, as it allows for unequal base frequencies and different rates of transitions and transversions. Each $\mathbf{\Lambda}^{(\cdot)}$ is then estimated via Maximum Likelihood as in (4.2) above, using the `markovjumps` package (Minin and Suchard, 2008; O'Brien et al., 2009). Finally, our labeled set for synonymous substitutions $\mathcal{L}_S$ is defined as follows: $(i, j) \in \mathcal{L}_S$ if there is a single nucleotide change in any of the three codon positions between codon i and codon j which does not

alter the translated amino acid. Thus, we replace the conventional distances in the Dss statistic with these labeled, robust distances, and evaluate performance of the modified statistic.

We also must make important modifications to the parametric bootstrap for assessing statistical significance of the Dss statistic as first proposed by McGuire and Wright (2000). First, we estimate the distances between sequences on the codon scale (e.g. the expected number of substitutions per codon site). Using this distance matrix, we then estimate the least squares tree, which represents the null hypothesis of the evolutionary history of the sequences: with no recombination or convergent evolution. Next, we estimate codon substitution parameters from the codon model proposed by Goldman and Yang (1994): κ (the transition/transversion ratio) and $\mathbf{\Omega} = (\Omega_1, \Omega_2, \Omega_3)$ where each Ω_i is a nonsynonymous/synonymous rate ratio, with corresponding proportions $\mathbf{p} = (p_1, p_2, p_3)$ where each p_i represents the probability that Ω_i will be selected as the nonsynonymous/synonymous rate ratio for any given site. This mixture of three codon models allows for estimation of variable nonsynonymous/synonymous rate ratios at each site, to simulate the bootstrap data as similarly as possible to the evolutionary process that created the original data. With the null evolutionary history, κ, $\mathbf{\Omega}$ and $\mathbf{p}$ estimated from the sequence data, we then simulate our parametric bootstrap sequence alignment datasets. Using these, we calculate the original Dss statistic and the synonymous Dss statistic in the manner described above, to then obtain the distribution of the maximum, for each.

4.3 Data

4.3.1 Simulations

To assess performance of each statistic, we simulate sequences under a codon model using the software package PAML (Yang, 2007). Three basic scenarios are considered: 1) null; 2) true recombination event; 3) localized convergent evolution. For

each scenario, we consider a sequence alignment with five taxa, and we set $\kappa = 2$ (transition/transversion ratio), and sample Ω (nonsynonymous/synonymous rate ratio) from a discrete mixture model with values $\mathbf{\Omega} = (0.1, 0.8, 3.2)$. We used three sets of sampling probabilities: $\mathbf{p}_1 = (0.74, 0.24, 0.02)$, $\mathbf{p}_2 = (0.85, 0.14, 0.01)$ and $\mathbf{p}_3 = (0.99, 0.009, 0.001)$ to produce average synonymous substitution proportions of 50%, 60% and 75% respectively.

Under the null scenario, we assume that every site along the sequence alignment is inherited according to one true evolutionary history. To simulate this, we provide PAML with one true phylogeny with five tips (shown in panel A of Figure 4.3), and simulate codon sequences along this phylogeny. Each codon sequence consists of 1032 codon sites (or 3096 nucleotides).

For the scenario with a true recombination event, we use two phylogenies corresponding to each side of the recombination breakpoint (shown in panels A and B of Figure 4.3, which are identical except that taxa 2 and taxa 5 are swapped). Partial sequence alignments of length 400 and 632 codons are simulated according to each phylogeny respectively, and then concatenated to form one mosaic sequence alignment that is 1032 codons in length.

For the scenario of localized convergent evolution, we simulate codon sequences along one true phylogeny, and then choose a region upon which to have selection act. Specifically, we use the latter 632 codons as this region (as in the recombination scenario). In this region, we again target taxa 2 and taxa 5, but here we make substitutions to change differing amino acids each into another (concordant) amino acid. We choose only codon sites in which this convergent evolution could occur with one nucleotide change in each of the sequences, and select a proportion of these sites, uniformly at random, in which to make this change. To make scenarios comparable, we convert the same proportion, on average, of all sites that have codon variations initially: we set this proportion to 25% in all cases. Noting that our convergent evolution scheme can only act on sites which had nonsynonymous substitutions initially

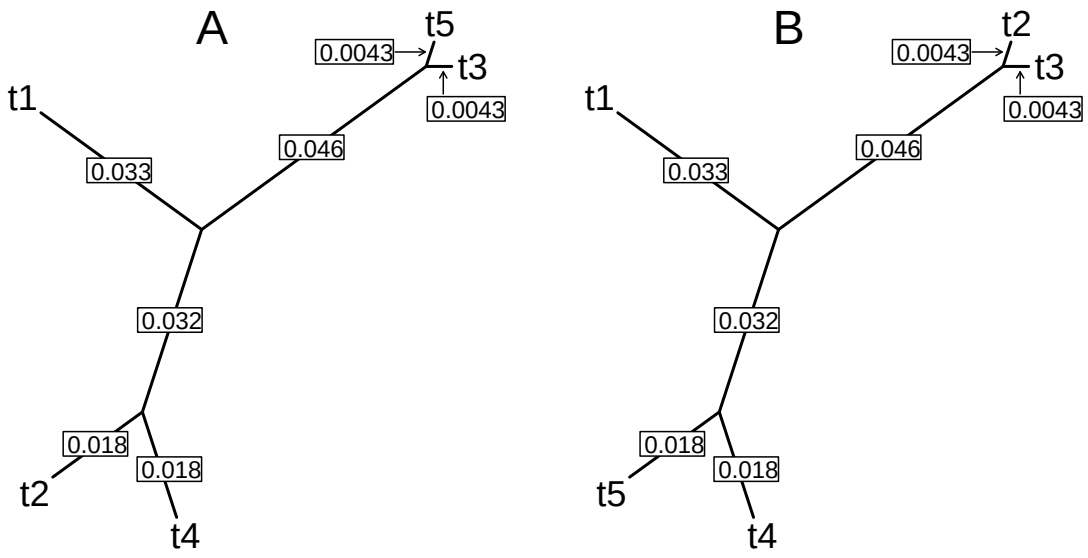

Figure 4.3: **Phylogenies used for simulations**. Numbers indicate branch lengths, in expected number of substitutions per site between two nodes.

(since they must be in different amino acid states), the proportion of eligible sites which get converted at random must be adjusted according to the percentage of non-synonymous substitutions in each scenario (described above), in order to maintain the overall proportion of 25% of all variable sites that will be converted.

Finally, for every scenario, we also vary the branch length, to effectively vary the number of substitutions, or diversity, in each simulation. The branch lengths shown in Figure 4.3 are the original set of branch lengths. We consider the original branch lengths, and also branch lengths that are scaled by 0.80 and 0.67 relative to the original branch lengths, to result in scenarios of "high diversity," "medium diversity," and "low diversity," respectively.

4.3.2 Viral Evolution

Phylogenetic analyses of HIV sequences are useful in characterizing its transmission and spread, and these analyses are particularly relevant to HIV's drug resistance mechanisms (Lemey et al., 2005). However, while HIV's high mutation rate and short generation time are conducive towards a phylogenetic reconstruction, phylogenetic inference can be confounded by both its high recombination rate, and selective pressures from antiretroviral therapies and selective pressures from each HIV strain's host body (Rambaut et al., 2004). Our method is the first to address the important issue of distinguishing between recombination and convergent evolution, and thus we apply it here.

Of particular interest are the *pol* and *env* genes of HIV-1, which are responsible for replication (Hill et al., 2005) and cell infiltration (Coffin et al., 1997), respectively. These two genes were studied through a transmission chain of nine Belgian HIV-positive patients (Lemey et al., 2005), in which it was found that a phylogenetic reconstruction using the sequenced *env* gene was compatible with the known transmission history among these nine patients; on the other hand, the phylogenetic reconstruction using the *pol* gene sequences was not compatible with the transmission history. This raised the question of whether selective pressures might be the cause of this incongruity.

Specifically, Lemey et al. (2005) explored whether the selective pressure may have been due to antiretroviral drug therapies applied to HIV-positive patients in the transmission chain. They hypothesized that patients on similar antiretroviral drug treatments may invoke convergent evolution on their HIV strains, due to the fact that their respective HIV viruses may develop the same drug resistance-associated mutations. By examinating known drug resistance-associated mutations within the *pol* gene, they found this was in fact the case with two of their individuals: "Patient A" and "Patient I." That is, these two individuals shared specific amino acid substitutions

that have been identified by the International AIDS Society as being associated with clinical resistance to HIV antiretroviral drugs (Johnson et al., 2003).

Following this observation, Lemey et al. (2005) then constructed phylogenetic trees for the *pol* gene based on synonymous distances and nonsynonymous distances separately, using the Syn-SCAN software (Gonzales et al., 2002). The synonymous tree was compatible with the transmission history, while the nonsynonymous tree was not, and showed Patient A's strains clustering with those of Patient I. For an illustration, see Figure 4.4 below, taken from Lemey et al. (2005). Thus, they concluded that the *pol* gene was under convergent selective pressure. Here, we revisit this question by examining the behavior of each Dss statistic on a concatenated *pol-env* sequence alignment. That is, if we join the two sequence alignments together as one, will either recombination detection method indicate the presence of intergenic or intragenic recombination?

4.3.3 Bacterial Evolution

Salmonella enterica is a pathogenic bacteria with roughly 2,500 serovars, defined by their specific surface antigens. While some serovars are responsible for severe diseases such as typhoid fever, most can be linked to milder infections such as gastroenteritis, in both humans and other animals. *Salmonella* are typically acquired by the consumption of contaminated food or water, and must have the ability to survive the harsh pH of the stomach if it is to infect its host. Further, those that cause systemic diseases must also be able to survive in the blood and replicate in the liver and spleen (Groisman and Ochman, 1997).

A growing body of research on its evolution has helped to characterize the pathogenicity of *Salmonella*. For example, it has recently been observed that such molecular mechanisms as gene loss via deletion, insertional inactivation and truncation are instrumental in the evolution of highly pathogenic *Salmonella* (Holt et al., 2009). Also, at a more general level, *Salmonella* differs from many other pathogens in that a large

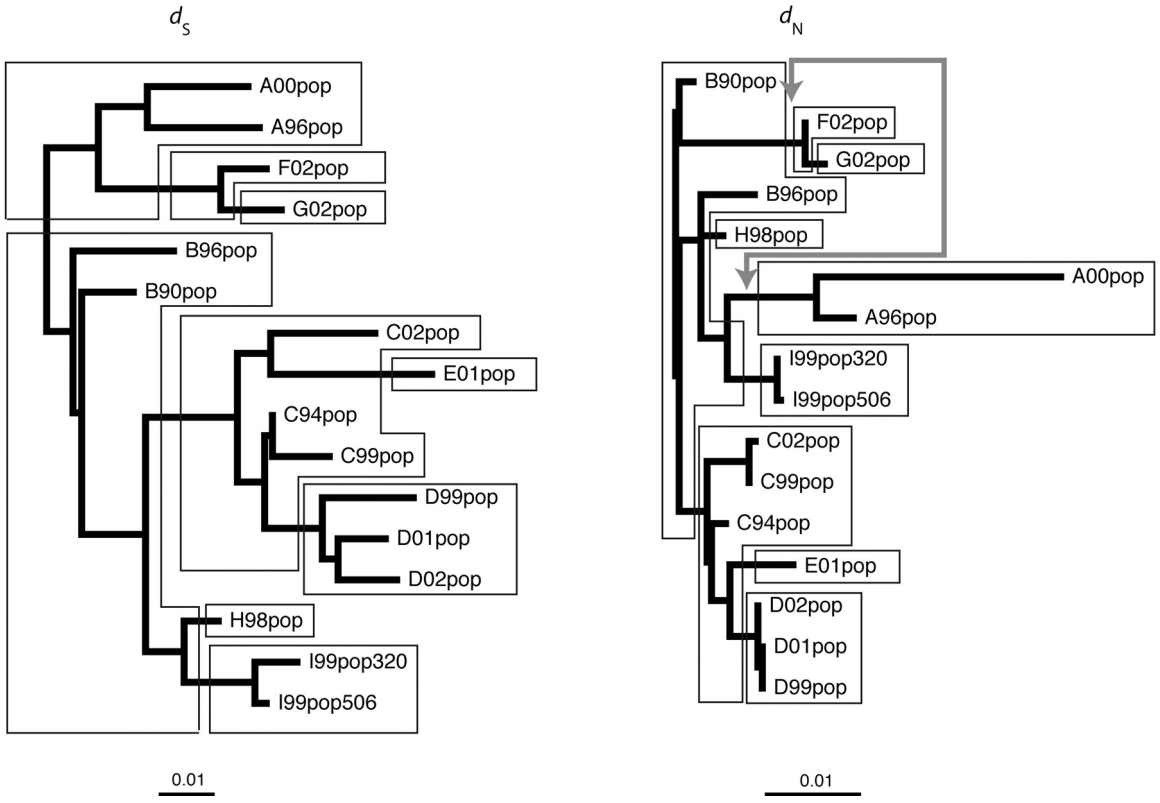

Figure 4.4: **Inferred phylogenies for the *pol* gene**. Based on synonymous and non-synonymous substitutions, separately. *The American Society for Microbiology has granted authorization to republish this figure in this dissertation.*

number of genes distributed around the chromosome are required for its virulence, as opposed to its virulence being controlled by a single region of its genome (Groisman and Ochman, 1997). This fact leads to the observation that many of these genes reside on so-called pathogenicity islands – large clusters of virulence genes – not found in related, nonpathogeneic species. This then raises the question: evolutionarily speaking, how would seemingly related species of *Salmonella* differ in so many genes that are found on these pathogenicity islands? A partial answer comes from a study of the pathogenicity island called SPI-1, which is indicated to have been acquired by *Salmonella* via recombination (Galán, 1997).

Here, we focus on the *Salmonella* FimH adhesin, an adhesive protein that may be useful in the study of evolutionary changes in *Salmonella* adaptation, as it is crucial in both cell adhesion and invasion (Bäumler et al., 1997; Naughton et al., 2001; Boddicker et al., 2002; Althouse et al., 2003). FimH expression is controlled by a number of regulatory proteins in response to environmental signals (Tinker et al., 2001; Chuang et al., 2008), which are also coupled with the expression of other virulence factors including the flagella (Clegg and Hughes, 2002) and LPS (Kwan and Isaacson, 1998). This suggests that FimH may be under differential selective pressures during different types of *Salmonella* infection, and Kisiela et al. (2012) showed evidence that these selective pressures act in a convergent manner throughout *Salmonella* evolution, using FimH adhesin variants from 33 serovars of *Salmonella enterica*. In this example, our aim is to determine whether we find evidence that corroborates the finding of Kisiela et al. (2012), that FimH has undergone convergent evolution, or rather whether there is evidence that recombination has occured, as in the case of SPI-1. Since our current implementation of the Dss statistics are limited to the number of taxa that they can handle, we isolated six representative strains from the original dataset, based on their evolutionary relatedness to each other.

4.4 Results

4.4.1 Simulations: Power and Type I Error

Using the trees shown in Figure 4.3, we simulate data according to the scenarios described above in Section 4.3.1. With $\alpha = 0.05$, Type I error rates under the null scenario appear to be well-behaved, with estimated Type I error rates of 6.6% and 6.3% for the original Dss and synonymous Dss tests, respectively. Distributions of the p-values resemble a uniform distribution as shown in Figure 4.5.

Then, we examine power to detect recombination, under the scenario with a true recombination event. We vary the expected number of substitutions (diversity) and

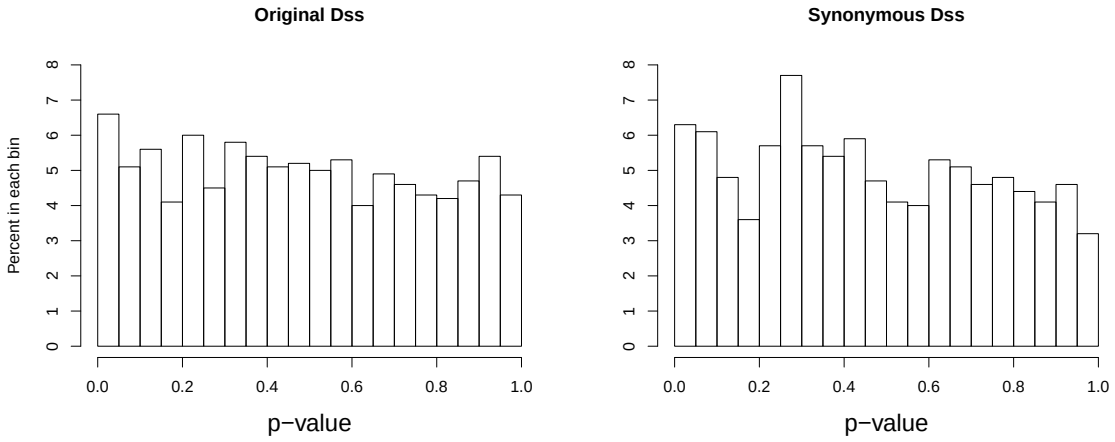

Figure 4.5: **Distribution of p-values under a simulated null scenario.** With 1000 replicates using original branch lengths ("high diversity"), and 50% synonymous substitutions.

proportion of synonymous substitutions, and examine the corresponding effect on the power of each version of the test statistic. Histograms of p-values from the original Dss statistic and synonymous Dss statistic from one scenario are shown in Figure 4.6, where we observe 90% power with the original Dss test statistic, and 76% power with our synonymous Dss test statistic. The results from all scenarios are shown in Table 4.1. Our synonymous Dss statistic has reduced power in every case, which is to be expected since we have reduced the amount of information used. The reduction in power is less dramatic in the scenarios where a greater proportion of the substitutions are synonymous (bottom row of Table 4.1), since less information is being discarded in these cases.

Under the scenario of convergent evolution, we compare the false positive rates under the original Dss statistic and the synonymous Dss statistic. That is, while the Dss statistic detects phylogenetic incongruence from any cause, we want to determine if the synonymous Dss statistic can avoid giving a significant p-value when the phylo-

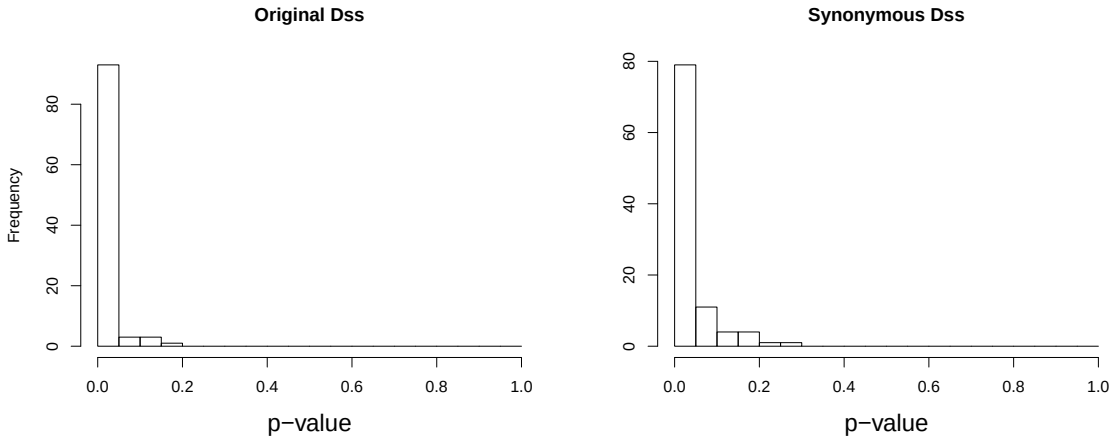

Figure 4.6: **Distribution of p-values under one representative simulated recombination scenario**.

genetic incongruence is due to convergent evolution. Under every scenario, we observe that the false positive rate of the synonymous Dss method is substantially lower than that of the original Dss statistic, as shown in Figure 4.7. For example, under high diversity and 50% synonymous substitutions, the estimated false positive rate for the original Dss statistic is 39%, vs. 7% for the synonymous Dss statistic. Results from all scenarios are shown in Figure 4.7, labeled as "Orig" and "Syn" respectively.

To further validate use of the synonymous Dss statistic, we examine whether this reduction in false positive rate might simply be due to the fact that the synonymous Dss statistic uses less information; that is, it considers only synonymous substitutions. To answer this question, we first examine the effect of removing a proportion of sites, corresponding to the proportion of synonymous substitutions. For example, under the scenario with 75% synonymous substitutions, we retain 75% of the alignment sites at random, and then obtain the original Dss statistic. We observe that the false positive rate under these simulations are similar to that of the original Dss statistic, as shown in Figure 4.7, labeled as "Del 1."

Table 4.1: **Power of each test under the recombination scenario**. Each column represents one set of branch lengths (equivalently, the diversity), which correspond to each average power of the original Dss test. Each cell represents the power of the synonymous Dss statistic, with 95% confidence intervals in parentheses. In each scenario, 100 simulated replications were analyzed.

	Power of original Dss		
	99%	90%	85%
50% syn	66 (56.6, 75.4)	38 (28.4, 47.6)	20 (12.1, 27.9)
60% syn	79 (70.9, 87.1)	48 (38.1, 57.9)	34 (24.6, 43.4)
75% syn	87 (80.3, 93.7)	76 (67.5, 84.5)	62 (52.4, 71.6)

However, this effort suffers from the fact that, while the sequence alignments are shorter, our window size has remained the same, thus resulting in fewer windows across the alignments. In our exploration of the Dss statistic behavior, we have noticed trends between window count and Power / Type I error (not shown) indicating that the "Del 1" regime is probably anti-conservative. Thus, we perform another validation experiment in which we also shrink the window size by the corresponding proportion; that is, if we removed 50% of the sites, we also shrink the window size by 50%. This is shown in Figure 4.7 as "Del 2." Based on our experimentation with the relationship between window size and power (not shown), we believe this to be a conservative effort, and yet in eight of the nine scenarios, we still obtain higher Type I error rates under this regime than that of the synonymous Dss statistic.

Finally, for the scenarios with 50% synonymous substitutions, we perform one additional set of experiments. Noting that in these scenarios, a nonsynonymous Dss statistic would have, on average, the same loss of information as our synonymous Dss statistic, we thus create a nonsynonymous Dss statistic in an analogous manner to which we created our synonymous Dss statistic, using labeled distances for nonsynonymous substitutions. We then run this nonsynonymous Dss statistic on the same set of data for each of the 50% synonymous substitution scenarios. In the high diversity case, we obtain a false positive rate of 19%, which is substantially higher than the

synonymous Dss statistic's false positive rate of 7%. For medium and low diversity, we obtain false positive rates of 13% and 7% respectively, which are identical to their respective estimated false positive rates from the synonymous Dss statistic.

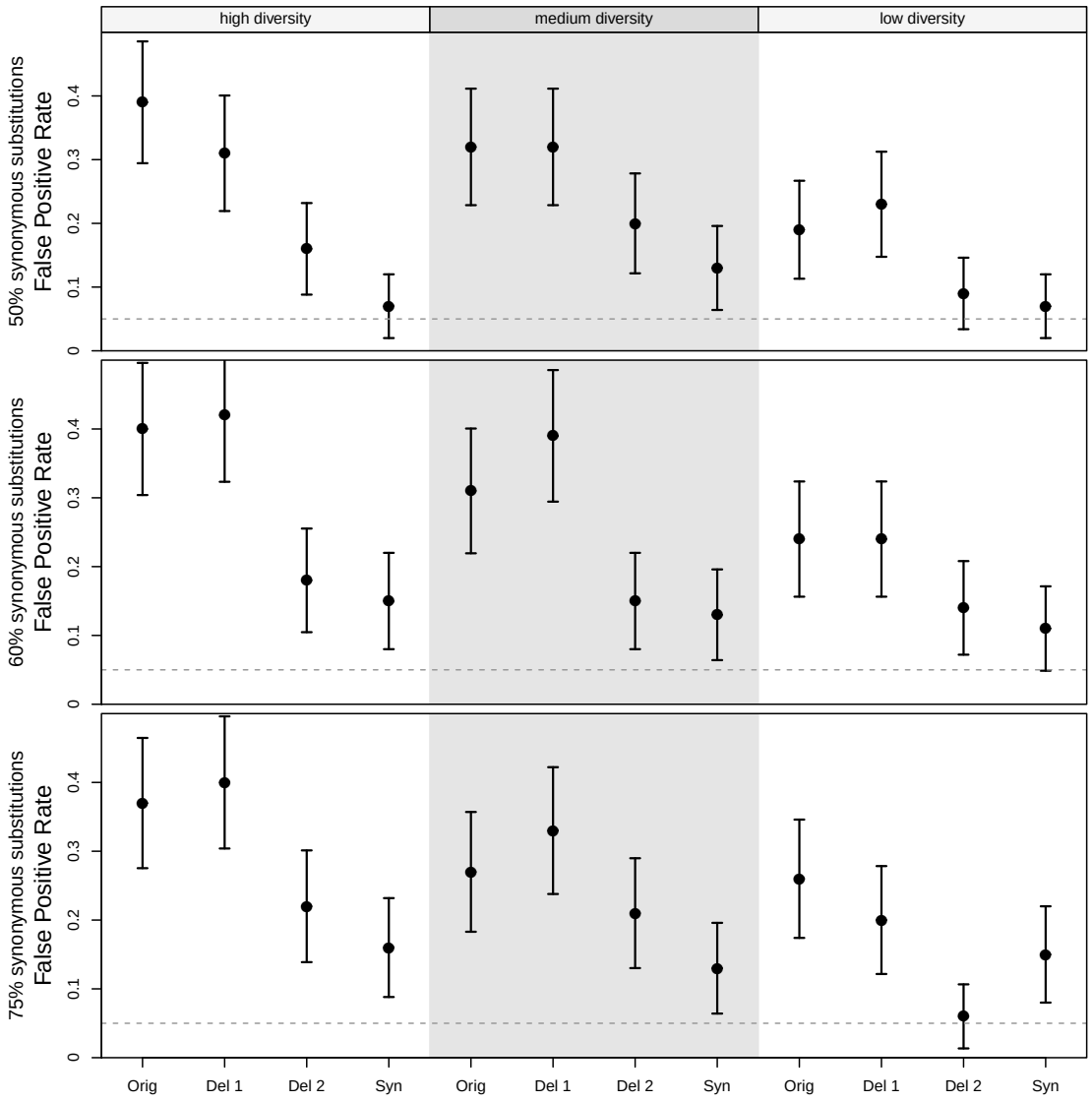

Figure 4.7: **False positive rate of each test under the convergent evolution scenario**. Using the same branch length sets (diversity) and synonymous substitution proportions as in the recombination scenarios, we induce convergent evolution on the alignment instead of a true recombination. "Orig" refers to the original Dss statistic; "Del 1" refers to the case in which we remove a proportion of substitutions corresponding to the non-synonymous substitution proportion; "Del 2" is similar to "Del 1" except that we also shrink the window size by the corresponding proportion; "Syn" refers to the synonymous Dss statistic. Error bars are based on the asymptotic binomial variance using the observed false positive rate as $\hat{p}$ to obtain standard errors. In each scenario, 100 simulated replications were analyzed.

4.4.2 Data Analysis I: Belgian HIV Transmission Chain

Lemey et al. (2005) previously studied the *pol* and *env* genes from nine Belgian HIV-positive patients, where the HIV transmission chain was known. Their study concluded that the *pol* gene was under converging selective pressure, whereas the *env* gene was not, based on whether phylogenetic reconstruction based on each of these genes individually was consistent with the known transmission chain or not (Lemey et al., 2005). Here, we revisit this question using our recombination detection statistics, by concatenating the two genes and treating them as one combined sequence alignment. Then, if the *pol* gene was in fact under convergent selective pressure, then we would expect that the original Dss test would find evidence for recombination, whereas our synonymous Dss test would not.

The original datasets consists of nine individuals, with multiple samples taken longitudinally from some of them. Because the current implementation of our method can only handle up to six taxa, we reduce the dataset by removing individuals with more recent transmission events. Our final dataset consisted of one sample from each of Patient A, B, F, H and I.

Results from our analyses are shown in Figure 4.8. We observe that the original Dss statistic crosses its 95% bootstrap significance threshold. Specifically, we observe that all of the original Dss crossings are in the *pol* region. While the synonymous Dss statistic has one peak that comes somewhat close to its 95% bootstrap significance threshold, it in fact does not reach it at any point. This suggests that the *pol* gene has undergone convergent evolution. We further explore this by examining the nonsynonymous Dss statistic, which crosses its 95% bootstrap significance threshold. This provides further evidence that the signal is due to convergent evolution, and that the synonymous Dss statistic's failure to cross its 95% significance threshold is not simply due to the reduced information content in the synonymous substitutions.

However, we note that in our data, the ratio of nonsynonymous to synonymous

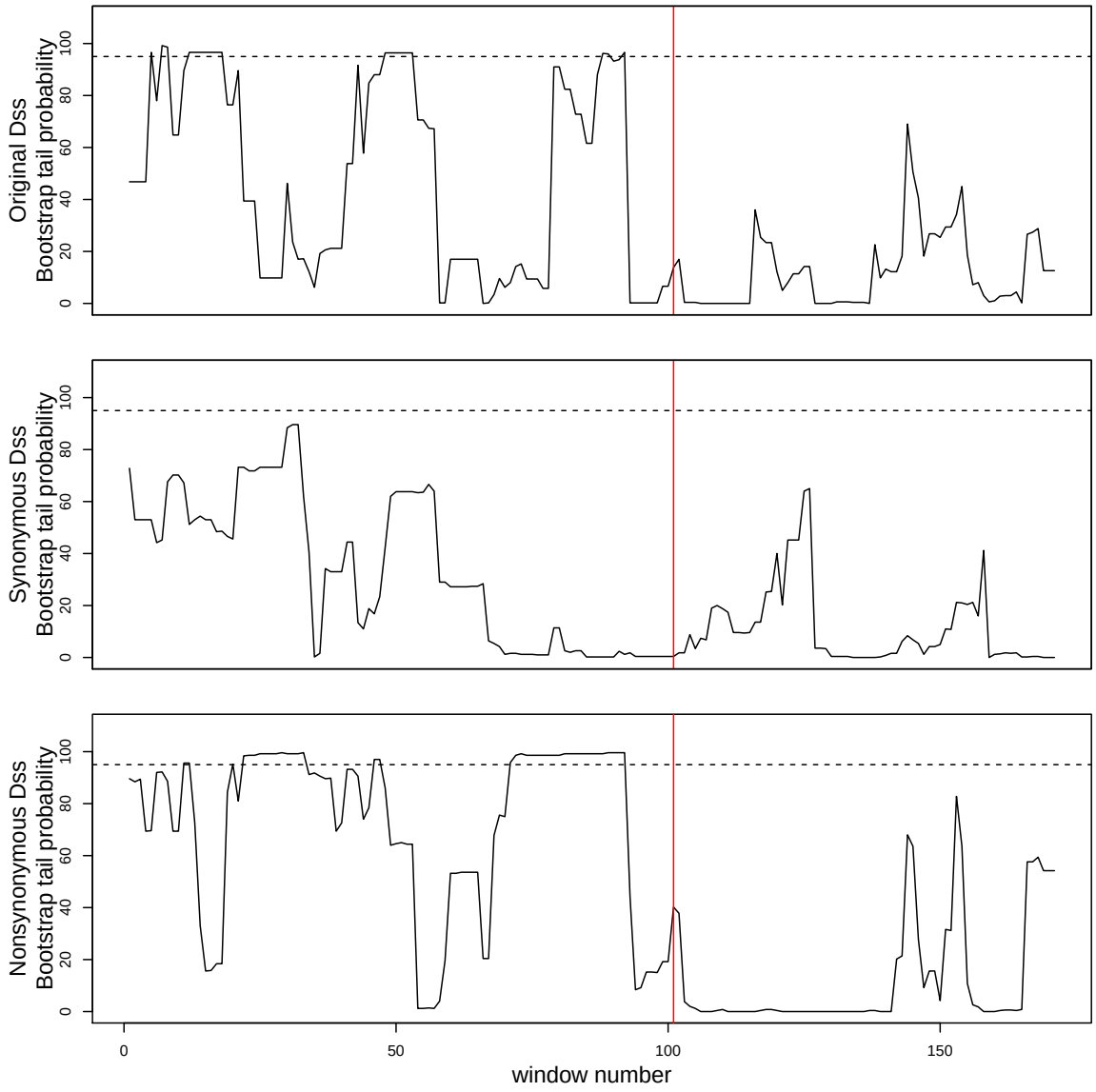

Figure 4.8: **Dss statistic landscapes for *pol-env* concatenation**. Dotted horizonal lines represent the 95% significance level for each test, from parametric bootstrap with $B = 500$. The red vertical lines represent the boundary between the two genes, with *pol* on the left, and *env* on the right.

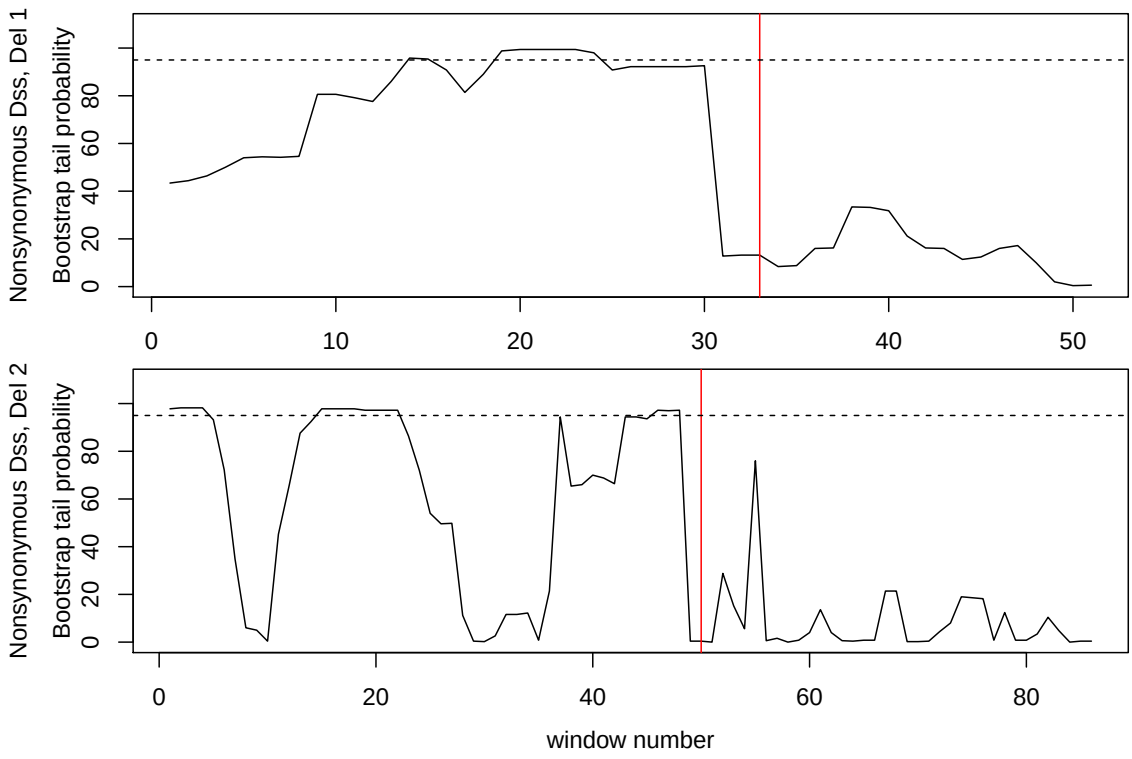

Figure 4.9: **Dss statistic landscapes for *pol-env* concatenation**. Dotted horizonal lines represent the 95% significance level for each test, from parametric bootstrap with $B = 500$. The red vertical lines represent the boundary between the two genes, with *pol* on the left, and *env* on the right.

substitutions is roughly $2:1$, in both the *pol* and *env* genes individually. Therefore, this comparison between the nonsynonymous Dss statistic and the synonymous Dss statistic is not based on each of them using the same amount of information. We thus turn to our "Del 1" and "Del 2" experiments as described in Section 4.4.1, in order to remove half of the information from the sequence alignment before running the nonsynonymous Dss method on it. The results from these analyses are shown in Figure 4.9. We notice that the nonsynonymous Dss statistic crosses its 95% bootstrap significance level in both the "Del 1" and "Del 2" experiments, adding further evidence that the *pol* gene is in fact under convergent evolution.

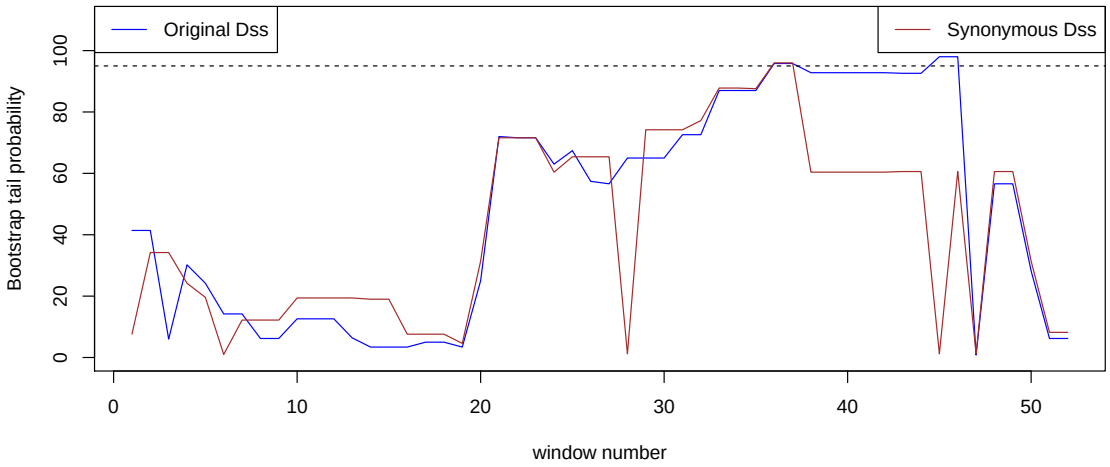

Figure 4.10: **Dss statistic landscape for *Salmonella enterica* FimH adhesin**. Dotted horizonal line represents the 95% significance level for each test, from parametric bootstrap with $B = 500$.

4.4.3 Data Analysis II: Salmonella enterica FimH adhesin

Previous work with these variants of FimH adhesin found evidence that convergent evolution has occurred throughout the evolutionary history of *Salmonella enterica* (Kisiela et al., 2012). Here, we revisit this question using our recombination detection statistics. If convergent evolution has indeed occurred, then we would expect that the original Dss test would find a statistically significant signal for recombination, whereas our synonymous Dss test would not, as in the HIV example.

Results from our analyses are shown in Figure 4.10. We observe that both the original Dss statistic and our synonymous Dss statistic cross their respective 95% bootstrap significance thresholds. This suggests that convergent evolution has not occurred, but rather that recombination is the evolutionary mechanism responsible for the genetic variation that has been observed here.

4.5 Discussion

In this chapter, we have introduced the synonymous Dss statistic, developed to give a statistical method which will allow us to distinguish between recombination and convergent evolution. Our simulations show that while our synonymous Dss statistic loses some power compared to the original Dss statistic, it does have a lower false positive rate when the signal is due to convergent evolution. Furthermore, we provide some verification that this lower false positive rate is not simply due to the loss of power, as suggested by the false positive rates of the various scenarios in which we remove a comparable portion of the information in the sequences, as shown in Figure 4.7. While the simulations with the nonsynonymous Dss statistic confirmed this improvement in only one of the scenarios (high diversity with 50% synonymous substitutions), we note that because the nonsynonymous Dss statistic's false positive rate was close to the $\alpha = 0.05$ in the other two cases, this could be making it difficult to discern any difference.

We are also able to validate the conclusion by Lemey et al. (2005) that convergent evolution has occurred in the *pol* gene lineage of HIV. The benefit of using our method is that we have created a statistical testing framework for addressing precisely this question. In contrast, Lemey et al. (2005) examined phylogenies constructed with synonymous and nonsynonymous substitutions separately, basing their conclusion on whether each phylogeny matched the known transmission chain. While their results certainly provided evidence for their conclusion, theirs was a somewhat circuitous strategy for answering the question. We also note that Lemey et al. (2005) did not provide any measure of statistical significance for their findings, which could have been shown via bootstrap support values for the bipartitions within their inferred phylogenies; in contrast, our method naturally assigns statistical significance to our individual findings. Furthermore, Lemey et al. (2005) implicitly assumed that the entire *pol* gene had one evolutionary history, and likewise for the *env* gene. If re-

combination had occurred within either gene, or if convergent evolution had occurred within only a specific region of either gene, then this assumption would be violated. In contrast, our method made no such assumption, which we believe to be an important advantage.

Lemey et al. (2005) also examined the nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$), which can be used as a test for selective pressure: $\omega = 1, \omega < 1, \omega > 1$ indicates neutral evolution, negative selection and positive selection, respectively (Miyata and Yasunaga, 1980). It has been noted, however, that the d_N/d_S test in its basic form has been shown to lack power for detecting selection, particularly in a similar case of HIV evolution where convergent evolution was known to have occured (Crandall et al., 1999). Thus, Lemey et al. (2005) turn to several more complex models (Nielsen and Yang, 1998; Yang et al., 2000a,b; Yang and Nielsen, 2002). Despite all of this, we note that fundamentally, the d_N/d_S test can only identify selection generically, regardless of the power of the test or the specific mutational model chosen; that is, it cannot determine whether the identified selection is of a converging nature or not. Our method presented here is the first to do so, and thus we believe our analyses to be more appropriate that the original, in regards to this particular question.

Conversely, our analyses do not agree with the finding by Kisiela et al. (2012) that FimH adhesin of *Salmonella enterica* is under convergent selective pressure, and instead we find statistically significant recombination. In this case, however, it is important to acknowledge that we reduced the number of analyzed sequences substantially, from roughly 50 down to six. Therefore, it is possible that the sequences excluded from our analysis have undergone convergent evolution. In fact, it is plausible that both recombination and convergent evolution play significant roles in the evolution of *Salmonella enterica* FimH.

The reason that our method can currently handle a limited number of sequences it can handle is because we perform an exhaustive search over the entire set of all

possible topologies, when optimizing over topologies for each window half as in (4.4). For smaller datasets (e.g. four or five sequences), this is not too costly in terms of computational time. However, for larger datasets, it quickly becomes prohibitive as the total number of topologies is equal to $(2n-5)!!$ where n is the number of sequences (Felsenstein, 2004). For example, with eight sequences, the total number of unrooted, bifurcating topologies is 10,395. In practice, there are various strategies for searching through tree space with higher numbers of sequences, such as nearest-neighbor interchanges, subtree pruning and regrafting, and tree bisection and reconnection (Felsenstein, 2004). However, in any non-monotone optimization problem, there is always the concern of falsely identifying a local optimum, if an exhaustive search is not performed. Since we are presenting here a proof of principle study, we deemed it to be critical to be certain that the true optimum is obtained at every step. For a real implementation, one potential strategy would be to first use a fast algorithm such as the neighbor-joining algorithm (Saitou and Nei, 1987) to obtain a starting topology, and then perform the least squares search from that point. Although this still would not guarantee the avoidance of a local optimum, at the very least it should result in reasonable trees, and would also likely be faster than starting the search from a random starting point.

A fundamental question that one might ask is how our method is advantageous over simply removing sites that contain nonsynonymous mutations. An illustration of the answer can be observed by considering an alignment containing a large number of sequences, in which multiple mutations per site would not be uncommon. Thus, if two mutations had occurred at a particular site, then one mutation could be synonymous and the other could be nonsynonymous. To use the brute-force approach of removing sites that contain nonsynonymous substitutions would necessarily remove the information contained in the synonymous mutation that had occurred at that site; that is, to remove the site means to remove the entire column from the sequence alignment, so all of the information contained in that site is lost. In contrast, our approach

of counting synonymous mutations under the framework laid out by O'Brien et al. (2009) removes the nonsynonymous mutation information in a more elegant manner, avoiding the total loss of information that would result from removing entire sites.

One concern with our simulation studies is that for the parametric bootstrap, we set $B = 100$. We were able to increase it to $B = 500$ for the real data, but it would have been too time consuming to do this for the simulations. We briefly examined how stable the 95% significance level was, when using $B = 100$. We found that the value of the significance level could differ by as much as 16% of Dss_{max}, on different runs with the same data. Therefore, it is very possible that, in our simulation studies, we are not accurately representing the null distribution of Dss_{max}. For future studies, the optimal value of B should be determined, and used.

A potential future development would be to create a coherent method to disintangle recombination and convergent evolution without a convoluted three-way comparison, between the original Dss statistic, the synonymous Dss statistic, and the nonsynoymous Dss statistic. That is, in this study, we would conclude that there is evidence for recombination if both the original Dss statistic and the synonymous Dss statistic show a positive signal. If the original Dss statistic shows a positive signal but the synonymous Dss statistic does not, then we would conclude that this is evidence of convergent evolution, further validated if the nonsynonymous Dss statistic also showed a positive signal. It would be preferable if a methodology could produce one coherent statistic to evaluate in order to answer this question, instead of two or three.

Finally, there is the potential that our concept here could be implemented in other recombination detection regimes, specifically those that are likelihood-based. Inherent to any distance-based, sliding window methodology such as that of the Dss statistic are some drawbacks. For example, the behavior of the statistic is somewhat influenced by the window size chosen, and there are few guidelines on how to select this tuning parameter (McGuire and Wright, 2000). Also, a multiple comparisons

issue exists, since each window produces a value of the test statistic. Although this issue is handled by considering only the maximum statistic value from the alignment and performing an appropriate parametric bootstrap test for statistical significance, it is unclear whether this completely alleviates all concerns; for example, how will the test statistic behave when there is a true recombination event producing a signal, and also a competing, spurious recombination signal elsewhere in the alignment? Thus, it may be advantageous to import our concept of synonymous recombination detection into a likelihood-based framework, such as those proposed by Husmeier and Wright (2003) or Minin et al. (2005).

Chapter 5

PHYLOGENETIC LEAST SQUARES INFERENCE WITHOUT DISTANCES

5.1 Introduction

A phylogeny is an object which describes evolutionary relationships between various groups, generically referred to as "taxa," where each taxon is typically a species. Phylogenies, or evolutionary trees, can be traced back to Charles Darwin, who drew what is believed to be the first one to exist, in his notebook in 1837 (Figure 2.1). However, statistical methods for inferring phylogenies are much younger, having existed for only roughly 40 years.

Fundamentally, methods for inferring phylogenies aim to group taxa with similarities to each other. These similarities can be with respect to whatever type of data are collected; for example, early phylogenies were based on morphological characteristics. In contrast, modern data are usually of the molecular form: DNA, amino acid or protein sequences.

Regardless of the data type, phylogenetic inference can be either model-free or model-based. For example, the method of Maximum Parsimony is perhaps the simplest numerical method, and is model-free: it simply counts the minimum number of necessary changes on each candidate phylogeny that would be required to observe the data, and finds the phylogeny which has the smallest of these necessary changes (Fitch, 1971). Unfortunately, its simplicity is overshadowed by the fact that it lacks some desirable statistical properties: most notably, it is inconsistent under some circumstances (Cavender, 1978; Felsenstein, 1978).

Model-based methods for inferring phylogenies include both distance-based meth-

ods and likelihood-based methods. While likelihood-based methods have attractive statistical behavior in general, they can suffer in terms of speed and computational burden. On the other hand, distance-based methods are generally the easiest to program, are very fast, and are well-justified statistically, leading to their continued popularity (Felsenstein, 2004, Chapter 11). Further, as we have seen in Chapter 4, distance-based methods have the flexibility of using labeled distances; an analogous use of labeled substitutions has not been demonstrated in any other class of methods to date. Therefore, the attractiveness of distance-based methods is quite high. As noted by Felsenstein (2004, Chapter 11), least squares phylogenetic inference generally performs almost as well as likelihood-based methods in terms of bias and variance, and better than all other methods. Our aim in this work is to devise an improvement to least squares phylogenetic inference that will decrease its bias and variance, bringing it closer to likelihood-based inference while maintaining flexibilities of distance-based methods.

Similar to estimation of a least squares regression line, least squares phylogenetic inference aims to minimize the squared differences between observed and predicted quantities, where the former are represented by the matrix of pairwise distances between each pair of molecular sequences, and the latter is a matrix of pairwise tree distances on the candidate tree, predicted by the tree topology and branch lengths (Cavalli-Sforza and Edwards, 1967). This makes least squares phylogenetic inference a two-stage procedure, since the distance matrix must first be estimated from the data, before tree estimation can then proceed based on this estimated distance matrix. In other words, inference is actually based on a summary statistic of the data, resulting in loss of information by the estimation procedure, and thus poorer performance (Penny et al., 1992).

Here, we aim to improve upon current least squares phylogenetic inference based on a modification of the least squares criterion, or loss function. Rather than treating distances as fixed quantities, our new loss function relies on conditional expectations

that depend both on the data and the candidate tree. Thus, the actual sequence data are considered simultaneously with the candidate trees, as opposed to first summarizing the sequence data into the distance matrix. Although our new proposed distances still use only pairwise comparisons (as opposed to accounting for the correlation which may exist in higher-order comparisons), we show that our new distances will result in better estimation of a phylogeny than using the usual least squares criterion.

Furthermore, it has been noted that distance-based methods are not able to use information about rate heterogeneity very well (Felsenstein, 2004, Chapter 11); that is, when substitution rates vary from site to site, the distance matrix can be corrected for this, but the correction is suboptimal as compared to likelihood-based modeling of substitution parameters. Conversely, using our modified least squares criteria leads to a natural approach for incorporating rate heterogeneity, which we hypothesize will lead to improved phylogenetic estimation in the presence of rate heterogeneity.

Thus, we examine the performance of least squares phylogenetic inference with our proposed loss function, and compare it to the performance when using the standard least squares criteria, under simulations. We begin by examining branch length estimation over the correct tree, thus temporarily setting aside the usual issues of searching through tree space, and evaluating whether the use of our new loss function results in improved branch length estimation, in terms of bias, variance and mean squared error (MSE). We then examine optimization of the loss functions while searching over topologies and evaluate whether the use of our new loss function results in correctly identifying the topology more frequently than use of the original loss function. Also, we explore branch length estimation across different models of DNA evolution, and compare performance of the two methods.

Finally, we conclude with an example using deep sequencing data from mouse tumor cells, in which Bayesian phylogenetic methods (Rannala and Yang, 1996; Yang and Rannala, 1997) were used to infer the lineage of the individual cells (Carlson et al., 2012). In one part of their analysis, Carlson et al. (2012) obtained some results

that differed slightly from the true, known cell lineage. We revisit this data with our two least squares methods, comparing their performance to the original Bayesian analysis, and to each other, in terms of correctly inferring the known cell lineage.

5.2 Methods

5.2.1 Least Squares Phylogenetic Inference

Least squares phylogenetic inference is based on evolutionary distances, which are generally defined as the expected number of substitutions per site, between two molecular sequence alignments (Yang, 2006, Chapter 1). Intuitively, they are a measure of how different two molecular sequences are from each other; two sequences that are identical to each other at every site would have a distance of 0, under most sensible distance metrics. Although distances can be either model-based or model-free, typically model-based distances have more desirable properties. For example, one model-free distance would simply be the proportion of discordant sites between sequences (also known as the Hamming distance). Its drawback is that it cannot account for multiple substitutions at a given site; thus, Hamming distances are not additive, and are not a linear function of evolutionary time (Felsenstein, 2004; Yang, 2006).

In order to define evolutionary distances, we must formally define notation to represent a molecular sequence alignment. Let $\mathbf{Y}$ be a matrix, composed of row vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where n is the number of taxa/species, or sequences. Then, $\mathbf{y}_k = (y_{k1}, \ldots, y_{kL})$, where L is the length, or number of sites in the sequence alignment. Each y_{ks} for $s \in \{1, \ldots, L\}$ takes on a value determined by the type of data; for DNA nucleotide sequences, the state space would be $\{A, C, G, T\}$.

Most commonly-used models of DNA evolution are continuous-time Markov chain models (Yang, 2006, Chapter 1). The simplest of these is the Jukes-Cantor model (also known as JC69), which assumes that the substitution rates between any pair of distinct nucleotides are equal (Jukes and Cantor, 1969). It is thus defined by the

substitution rate matrix

$$\mathbf{\Lambda} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}, \tag{5.1}$$

and stationary distribution $\boldsymbol{\pi} \equiv (\pi_A, \pi_C, \pi_G, \pi_T)$ equal to $(1/4, 1/4, 1/4, 1/4)$. In order to calculate distances, we must first obtain the finite-time transition probability matrix $\mathbf{P}(t) \equiv \{p_{y_{ks},y_{ls}}(t)\}$, where $p_{y_{ks},y_{ls}}(t)$ is the probability that site s is in state y_{ks} and y_{ls} when the two sequences are separated by time t. Thus, we note that $\mathbf{P}(t) = e^{\mathbf{\Lambda}t}$ (Guttorp, 1995), which can be solved to obtain

$$\mathbf{P}(t) = e^{\mathbf{\Lambda}t} = \frac{1}{4} \begin{pmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{pmatrix}, \tag{5.2}$$

where we note that the diagonal entries are all the same, and likewise for the off-diagonal entries; we will refer to them as $p_0(t)$ and $p_1(t)$ respectively.

Then, we further note that the total rate of leaving any particular state is equal to $\lambda_i \equiv \sum_{j \neq i} \lambda_{ij} = 3\alpha$. Thus, with distances defined as the expected number of substitutions per site between two sequences, then we can express distances as

$$d = \sum_{i \in \{A,C,G,T\}} \pi_i \lambda_i t \tag{5.3}$$

$$= 3\alpha t. \tag{5.4}$$

Using this relationship in (5.4) and the observation that the probability of any given site being different in two sequences after time t is $p = 3p_1(t)$ from (5.2), we thus

have

$$p = \frac{3}{4} - \frac{3}{4}e^{-4\alpha t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}. \tag{5.5}$$

Finally, using the method of moments, we obtain

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right), \tag{5.6}$$

to estimate distances from DNA sequences under the JC69 model. The same techniques can be used to obtain a method of moments estimator of distance for most other continuous-time Markov chain models of DNA evolution (namely, those that are stationary and reversible). Alternatively, one could also arrive at an estimator for distances by writing the data likelihood:

$$\mathbf{Pr}(\mathbf{y}_k, \mathbf{y}_l | \mathbf{\Lambda}, t) = \prod_{s=1}^{L} \hat{\pi}_{y_{ks}} \mathbf{P}_{y_{ks}, y_{ls}}(t), \tag{5.7}$$

and then obtain a maximum likelihood estimate for $\mathbf{\Lambda}$. We note that $\mathbf{\Lambda}$ and t appear only as a product in this expression, as observed in (5.2); thus, a constraint must be placed on t in order to obtain identifiability of $\mathbf{\Lambda}$. One possibility is to assume that $t = 1$ (O'Brien et al., 2009). The stationary distribution $\boldsymbol{\pi}$ might be implied by the continuous-time Markov chain model (as it is for JC69), or it can be estimated as the empirical nucleotide frequencies in the dataset. Then, using $\hat{\mathbf{\Lambda}}$, $\hat{\boldsymbol{\pi}}$ and a constraint on t, one can use the expression in (5.3) to obtain maximum likelihood estimates of evolutionary distances.

With an estimated distance matrix, one can then infer a phylogeny. Formally, a phylogeny consists of a topology $\boldsymbol{\tau}$ and branch lengths $\mathbf{b}$. The topology refers simply to the shape of the tree, or branching order, and in this work, we will consider only the class of topologies known as unrooted (lacking any indication of historical ancestry

of the taxa) and bifurcating (each split results in exactly two new lineages). The branch lengths refer to the evolutionary distance between two nodes, where a node can be either an observed taxon at the tips of the topology, or an internal and possibly unobserved taxon. A simple example of an unrooted bifurcating phylogeny is shown in Figure 5.1.

Least squares phylogenetic inference proceeds by defining the tree distance $t_{kl}(\boldsymbol{\tau}, \mathbf{b})$ between taxa k and l as the sum of the branch lengths between them on the particular phylogeny determined by $\boldsymbol{\tau}$ and $\mathbf{b}$. For example, tree distance between taxa t3 and t4 in Figure 5.1 is $0.0718 + 0.0716 + 0.0628 = 0.2062$. Next, recall from above that we have estimated $\hat{d}_{kl}$ from our DNA sequence data. Then, the quantity

$$L_1(\mathbf{d}, \mathbf{b}, \boldsymbol{\tau}) = \sum_{k=1}^{n} \sum_{l=1}^{n} \left[\hat{d}_{kl} - t_{kl}(\boldsymbol{\tau}, \mathbf{b}) \right]^2, \tag{5.8}$$

is the usual least squares criterion, adapted to our context of phylogenetics, where $\mathbf{b}$ and $\boldsymbol{\tau}$ are defined above, and $\mathbf{d}$ is the matrix of pairwise distances d_{kl}. The solution $(\hat{\boldsymbol{\tau}}, \hat{\mathbf{b}})$ to

$$\underset{\boldsymbol{\tau}, \mathbf{b}}{\operatorname{argmin}} \; L_1(\mathbf{d}, \mathbf{b}, \boldsymbol{\tau}) \tag{5.9}$$

then gives the least squares phylogeny.

5.2.2 New loss function

In this work, we propose an improvement to least squares phylogenetic inference, by considering a new loss function. Our new loss function bypasses the need for the intermediate step of estimating the distance matrix, and instead considers the sequence alignment directly. First, we define $N_{t_{kl}(\boldsymbol{\tau}, \mathbf{b})}$ as the number of substitutions between taxa k and l, along the tree defined by $\boldsymbol{\tau}$ and $\mathbf{b}$. Then, $\mathbb{E}(N_{t_{kl}(\boldsymbol{\tau}, \mathbf{b})} | y_{ks}, y_{ls})$

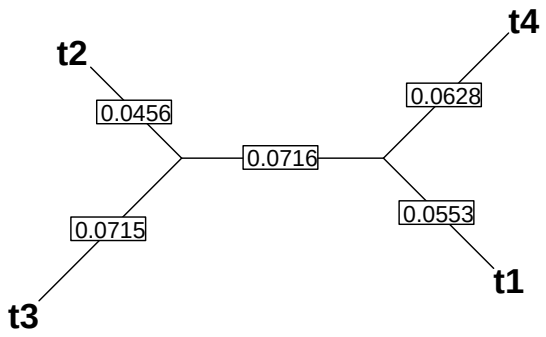

Figure 5.1: **An unrooted, bifurcating phylogeny**. This particular topology indicates that taxa "t1" and "t4" are more closely related to each other than to taxa "t2" and "t3," and vice-versa. The numbers placed on each branch indicates its length.

is the expected number of substitutions between taxa k and l, given their observed states at site s, and a chosen mutational model. The empirical average number of substitutions across the alignment is thus

$$\hat{e}_{kl}(\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}) = \frac{1}{L} \sum_{s=1}^{L} \mathbb{E}(N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})} | y_{ks}, y_{ls}), \tag{5.10}$$

where $\boldsymbol{\theta}$ refers to parameter(s) of the model of DNA evolution (for JC69, the one parameter α is actually subsumed into $t_{kl}(\boldsymbol{\tau}, \mathbf{b})$, but for other models, there may be more parameters – see Section 2.1.3). Finally, we define our new loss function

$$L_2(\mathbf{Y}, \mathbf{b}, \boldsymbol{\tau}, \boldsymbol{\theta}) = \sum_{k=1}^{n} \sum_{l=1}^{n} \left(\hat{e}_{kl}(\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}) - t_{kl}(\boldsymbol{\tau}, \mathbf{b}) \right)^2, \tag{5.11}$$

which we note is similar to L_1 in (5.8), but with the replacement of the original model-based estimates of distance $\hat{d}_{kl}$ with the so-called robust distance $\hat{e}_{kl}$ (Minin and Suchard, 2008; O'Brien et al., 2009). Explicitly, $\hat{e}_{kl}$ depends on $(\mathbf{y}_k, \mathbf{y}_l, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{b})$ simultanously, whereas $\hat{d}_{kl}$ is a summary statistic of $(\mathbf{y}_k, \mathbf{y}_l)$ based on the previously estimated model parameters $\boldsymbol{\theta}$. Details for the calculation of (5.10) was demonstrated

by Ball and Milne (2005), among others.

We note that marginally, $\mathbb{E}(N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}) = t_{kl}(\boldsymbol{\tau},\mathbf{b})$, as branch length is defined as the expected number of substitutions between two nodes. Thus, if we have the mutational model exactly correct (i.e. if our model exactly describes the evolution of our data from the true tree), then our least squares criterion L_2 should be minimized at exactly 0 when we have identified the correct tree, as the sequence length $L \to \infty$. We postulate, however, that even with the inevitable model misspecification, we will still benefit from gains due to the fact that L_2 considers the sequence data directly, and that $\hat{e}_{kl}$ incorporates the branch length directly and avoids the necessity of estimating pairwise distances as an intermediate step.

5.2.3 Optimization

Under loss function L_1, the least squares solution can be obtained through standard matrix algebra approaches, as a phylogeny can be fully represented by the collection of all $t_{kl}(\boldsymbol{\tau},\mathbf{b})$, and the $\hat{d}_{kl}$ are constants, once they are obtained from the data (Felsenstein, 2004). However, finding the least squares solution becomes non-trivial under L_2, since $\mathbb{E}(N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}|y_{k\nu}, y_{l\nu})$ depends on $t_{kl}(\boldsymbol{\tau},\mathbf{b})$. We must therefore resort to iterative approaches in which we can update $\mathbb{E}(N_{t_{kl}}|y_{k\nu}, y_{l\nu})$ with each change in $t_{kl}(\boldsymbol{\tau},\mathbf{b})$.

Furthermore, there exists here an additional concern that, in reality, $L_2 \to 0$ as each $t_{kl}(\cdot,\cdot) \to \infty$. We demonstrate this here in the case of the JC69 model. For one site and a given pair of sequences, we have:

$$
\begin{aligned}
L_2(\mathbf{Y},\mathbf{b},\boldsymbol{\tau},\boldsymbol{\theta}) &= \left(\hat{e}_{kl}(\boldsymbol{\tau},\mathbf{b},\boldsymbol{\theta}) - t_{kl}(\boldsymbol{\tau},\mathbf{b})\right)^2 && (5.12)\\
&= \left(\mathbb{E}\left[N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}|y_{ks}=i, y_{ls}=j\right] - t_{kl}(\boldsymbol{\tau},\mathbf{b})\right)^2 && (5.13)\\
&\overset{\text{JC69}}{=} \left(\mathbb{E}\left[N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}|y_{ks}=i, y_{ls}=j\right] - 3\alpha t\right)^2 && (5.14)
\end{aligned}
$$

where (5.14) follows from our definition of distance (5.4). Then, it has been shown

that

$$\mathbb{E}\left[N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})} \cdot 1_{(y_{ks}=i)} | y_{ls} = j\right] = t \cdot [\boldsymbol{\Lambda} - \text{diag}(\boldsymbol{\Lambda})] e^{\boldsymbol{\Lambda} t} \tag{5.15}$$

(Ball and Milne, 2005; Minin and Suchard, 2008). Dividing the matrix $t \cdot [\boldsymbol{\Lambda} - \text{diag}(\boldsymbol{\Lambda})] e^{\boldsymbol{\Lambda} t}$ element-wise by $p_{ij}(t)$, we thus obtain

$$\mathbb{E}\left[N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})} | y_{ks} = i, y_{ls} = j\right] = \begin{cases} t\alpha\left(\frac{3(1-e^{-\alpha t})}{1+3e^{-\alpha t}}\right) & i = j, \\ t\alpha\left(2 + \frac{1+3e^{-\alpha t}}{1-e^{-\alpha t}}\right) & i \neq j. \end{cases} \tag{5.16}$$

Now, for $i = j$, we have

$$\begin{aligned} \mathbb{E}\left[N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})} | y_{ks} = i, y_{ls} = j\right] - 3\alpha t &= t\alpha\left(\frac{3(1-e^{-\alpha t})}{1+3e^{-\alpha t}}\right) - 3\alpha t & (5.17) \\ &= 3\alpha t\left(\frac{1-e^{-\alpha t}}{1+3e^{-\alpha t}} - 1\right) & (5.18) \\ &= 3\alpha t\left(\frac{-4e^{-\alpha t}}{1+3e^{-\alpha t}}\right) & (5.19) \\ &= \frac{-12\alpha t e^{-\alpha t}}{1+3e^{-\alpha t}}, & (5.20) \end{aligned}$$

which converges to 0 as $(\alpha t) \to \infty$, by taking limits of the numerator and denominator separately. A similar argument can be made for the case $i \neq j$. Intuitively, this observation follows from the fact that, marginally, $\mathbb{E}\left[N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}\right] = t_{kl}(\boldsymbol{\tau}, \mathbf{b})$ by definition; then, as $t_{kl}(\boldsymbol{\tau}, \mathbf{b}) \to \infty$, $N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}$ becomes independent of the observed states. This is because $N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}$ and the observed states are both properties of the Markov chain path that rely on the amount of time that has passed. Specifically, $N_{t_{kl}(\boldsymbol{\tau},\mathbf{b})}$ counts the number of jumps along the entire path, and y_{ks}, y_{ls} can be considered as the start and end points, respectively. As the length of the path grows large, the count starts to rely less on where it began and ended.

Of course, although the above demonstrates the "true" solution to our minimiza-

tion problem at hand, it is clearly not the solution which we desire to obtain. To circumvent this issue, we utilize box-constrained optimization (Box, 1965), to bound the estimated branch lengths at a point before which they will tend towards ∞. In our experiments, we place an upper bound of 1 on our branch lengths, and a lower bound of e^{-10}, or effectively 0.

In order to implement our desired optimization, we turn to currently available packages in `R`. We compare the performance of seven routines which are capable of iterated box-constrained optimization over several variables (Table 5.1), both in their ability to actually minimize the loss function, and in their speed. We simulated 30 DNA sequence alignments, and ran each optimization routine with our loss function L_2.

Although each routine has convergence criteria that it aims to satisfy in order to report whether the minimum has been obtained or not, any given two routines did not necessarily agree on the returned loss function value, even when both had independently determined that it has achieved convergence. Thus, in comparing these seven routines, we deem that convergence has been achieved by the one which returned the smallest value of the loss function, and record this under the "# best" column in Table 5.1. The average time per iteration is shown under "Time (s)."

Out of the 30 iterations, `nlminb` found the smallest value 26 times, and `Rvmmin` found the smallest value 4 times. We note that although we did not observe any ties in the minimized loss function value, in most cases 3-4 of the routines would agree within 10 decimal places. This in fact was the case with `nlminb` and `Rvmmin` in the 4 times when `Rvmmin` found the smallest value. On the other hand, `nmk` (the fastest routine, on average) often returned values far above the smallest value, sometimes as much as 50 times greater. Since `nlminb` is the next fastest at an average of 25.5 seconds per iteration, and has excellent performance as just previously noted, we choose this routine to optimize our loss function L_2.

Table 5.1: **Performance of box-constrained optimizers**. "# best" indicates the count of how many times each routine found the smallest loss function value. "Time (s)" is the average time of each routine, in seconds, across all 30 iterations performed.

Algorithm	Package	# best	Time (s)	Reference
`bobyqa`	**minqa**	0	137.5	(Powell, 2009)
`L-BFGS-B`	**stats**	0	47.6	(Byrd et al., 1995)
`nlminb`	**stats**	26	25.5	(Fox, 1997)
`nmk`	**dfoptim**	0	22.0	(Kelley, 1999)
`Rcgmin`	**Rcgmin**	0	55.4	(Nash, 1979)
`Rvmmin`	**Rvmmin**	4	33.3	(Nash, 1979)
`spg`	**BB**	0	137.3	(Birgin et al., 2000)

5.3 Data

5.3.1 Simulations

To assess performance of phylogenetic inference with our proposed least squares criterion, we simulate nucleotide sequences using the software package PAML (Yang, 2007). A variety of scenarios are considered, in which we vary number of taxa, topology and branch lengths. Throughout this work, we will use the following nomenclature to refer to particular scenarios:

- Balanced tree (B) - all branch lengths are similar to each other
- Unbalanced Long External 1 (ULE1) - one long external branch
- Unbalanced Long External 2 (ULE2) - two long external branches
- Unbalanced Long Internal 1 (ULI1) - one long internal branch
- Unbalanced Long Internal 2 (ULI2) - two long internal branches
- Unbalanced Short Internal (USI1) - one short internal branch

Fixed topology

We start by considering estimation of only the branch lengths. That is, we simulate DNA sequence alignments along a particular topology $\boldsymbol{\tau}$ and branch length set $\mathbf{b}$, and provide $\boldsymbol{\tau}$ to the estimation procedure as known and fixed. The topologies used for simulation range in size from five to seven taxa, with various branch length patterns as described above. In particular, it has been noted that the performance of phylogenetic inference with distance-based methods is inversely related to the diameter of the tree (Atteson, 1999; Lacey and Chang, 2006; Roch, 2010), where diameter refers to the maximum distance between any two tips. Thus we include scenarios in which the true phylogeny has a large diameter (e.g. ULE1, ULE2, ULI1, ULI2). In each case, we use the JC69 model of DNA evolution.

Varying topology

Next, we consider estimating branch lengths and topologies, simultaneously. Again, we simulate DNA sequence alignments along a particular topology $\boldsymbol{\tau}$ and branch length set $\mathbf{b}$ according to the JC69 model of DNA evolution, but we do not assume that $\boldsymbol{\tau}$ is known. Because optimization over topology space is prone to falsely identify local optima as the true optimum, we focus our attention on topologies of fewer taxa (four and five), so that the entire topology space can be explored (in a reasonable amount of time). It has been noted that trees with short internal branch lengths are more difficult to infer correctly (Martyn and Steel, 2012); thus, we use the USI1 scenario, vary the sequence alignment length from 200, 400, 600 and 800 nucleotides, and compare how many times the correct topology is obtained by each method.

Inferring substitution model parameters

We also explore added model complexity, but reverting back to fixed topologies. The framework available through the use of the robust distances $\hat{e}_{kl}$ readily allows for

natural estimation of model parameters from more complex models of DNA evolution, simultaneously with branch length estimation. We demonstrate an example using the K80 model (Kimura, 1980). While the JC69 model assumes that all substitutions occur at the same rate, the K80 model allows for varying rates of certain types of substitutions. Specifically, in real data, the rate of transitions ($T \leftrightarrow C$ or $A \leftrightarrow G$) is often higher than the rate of transversions ($T, C \leftrightarrow A, G$). Thus, Kimura (1980) proposed the K80 model to parameterize the transition and transversion rates separately, as α and β respectively. The substitution rate matrix for the K80 model is thus

$$\mathbf{\Lambda} = \begin{pmatrix} -(\alpha+2\beta) & \beta & \alpha & \beta \\ \beta & -(\alpha+2\beta) & \beta & \alpha \\ \alpha & \beta & -(\alpha+2\beta) & \beta \\ \beta & \alpha & \beta & -(\alpha+2\beta) \end{pmatrix}, \tag{5.21}$$

where the row and column labels are in alphabetical order: (A, C, G, T).

Then, the ratio $\kappa = \alpha/\beta$ can be estimated along with the distance matrix. For least squares phylogenetic inference with the original loss function, distances can be adjusted for κ; thus, the form of (5.8) remains the same, using adjusted distance estimates $\hat{d}_{kl}^{(\kappa)}$. For our new loss function, we appeal to the concept of labeled distances introduced in Chapter 4, and count transitions and transversions separately. Thus, our loss function becomes

$$\sum_{k=1}^{n}\sum_{l=1}^{n}\left[\hat{e}_{kl}^{(ts)} - t_{kl}^{(ts)}(\boldsymbol{\tau}, \mathbf{b})\right]^2 + \left[\hat{e}_{kl}^{(tv)} - t_{kl}^{(tv)}(\boldsymbol{\tau}, \mathbf{b})\right]^2, \tag{5.22}$$

where "ts" designates transitions, and "tv" designates transversions. To test estimation under both loss functions, we simulate DNA sequence alignments according to $\kappa = (0.25, 0.5, 2, 4)$, and the ULI1 and ULI2 scenarios with four and five taxa trees.

Rate heterogeneity: JC69+dΓ

Finally, we consider substitution rate heterogeneity across sites. In real sequences, it is often the case that each site will have a different substitution rate, and ignoring this can have dramatic impact on phylogenetic inference (Tateno et al., 1994; Sullivan and Swofford, 2001). Thus, we simulate DNA sequence alignments with rate heterogeneity, and examine the resulting impact on least squares phylogenetic inference with the L_1 and L_2 loss functions.

While one could argue that the most realistic model would give each site its own rate of evolution, it is typically prohibitive to attempt to model rate heterogeneity with a distinct parameter per site, since this will result in far too many parameters for sequence alignments of any reasonable length. Instead, one strategy is to use a statistical distribution to model the rate variation (Yang, 2006). The most common choice is the gamma distribution, with shape and rate parameters α and β, and setting $\alpha = \beta$ so that the mean of distribution is 1 (Yang, 1993; Gu et al., 1995; Kelly and Rice, 1996). In this model, rate scalings for each site are drawn at random from this distribution, and the substitution rate matrix for each site is then multiplied by the respective values. If, for example, the substitution rate matrix is originally for the JC69 model, then this overall model is called JC69+Γ.

Unfortunately, implementing even this simplification is often too slow, for datasets greater than 10 sequences (Yang, 2006). A further simplification is to use the discrete-gamma model (Yang, 1994a), in which the substitution matrix is multiplied by a value drawn at random from a discretized gamma density (into any chosen number of categories), again defined by one parameter α which is used for both the shape and rate. For example, suppose that the chosen number of categories is four. Then, the usual gamma density is discretized into four equally probable categories. The mean values of each of the categories are then the possible rate scaling values that are drawn at random. Then, if the substitution rate matrix is originally for the JC69 model,

then this overall model is called JC69+dΓ. It is of note that as $\alpha \to \infty$, all of the mass of the distribution tends towards 1, so that the rate heterogeneity disappears.

We simulate DNA sequence alignments according to the i.e. JC69+dΓ model, with four categories. We choose $\alpha = 0.5$ and $\alpha = 2$, to simulate under high rate heterogeneity and low rate heterogeneity respectively. Using four and five taxa trees, we simulate data under the ULI1 and ULI2 scenarios, using sequence lengths of 1000 nucleotides. Our new loss function is of the same form as before, but now $\hat{e}_{kl}$ is calculated piece-wise for each rate. That is, with four rate categories, we have

$$\sum_{k=1}^{n}\sum_{l=1}^{n}\left[\left(\frac{1}{4}\sum_{m=1}^{4}\hat{e}_{kl}^{(r_m)}\right) - t_{kl}(\boldsymbol{\tau}, \mathbf{b})\right]^2, \tag{5.23}$$

where $\hat{e}_{kl}^{(r_m)}$ is the estimated robust distance with substitution rate matrix that is scaled by r_m. The values of r_m are determined by the value of the parameter α, which is estimated simultaneously with $\boldsymbol{\tau}$ and $\mathbf{b}$. We note that we experienced convergence issues with (5.23); to ameliorate this, we ran the three optimization routines `nlminb`, `spg` and `Rvmmin` simultaneously for each iteration, and chose the result that had the lowest value of (5.23), given that it converged. If none of the routines converged, then we reported a failed iteration.

Currently, there is no standard method for estimating α with the original distances. While the distance matrix can be adjusted if the value of α is known (Felsenstein, 2004), our attempts to estimate α by minimizing the loss function (5.8) over all values of α with adjusted distances $d_{kl}^{(\alpha)}$ were unsuccessful due both to convergence issues and instability of the α estimates even when the optimization algorithms had passed their internal convergence diagnostics. Also, the additional efforts used to achieve convergence when minimizing (5.23) as described above were unsuccessful in alleviating the convergence issues when using the original distances. Thus, while we use (5.23) to estimate α within the framework of our new least squares loss, we

compare this to least squares phylogenetic inference using the L_1 loss function without modification.

5.3.2 Cancer Cell Lineage Fate Maps

Here, we use phylogenetics to study the relationships between individual cells in an organism. This line of work was pioneered by Sulston et al. (1983), who tracked the lineage of all 671 cells of the nematode *C. elegans* during its embryogenesis, through physical inspection under a microscope. Such a task is only possible in *C. elegans* due to the fact that it is translucent and has a relatively small number of cells to keep track of; in higher mammals, a similar endeavour would be virtually impossible. However, relationships between cells can be inferred due to the fact that, while rare, errors in DNA replication do occur during cell division; in normal human cells, this has been estimated to occur at a rate of about 1 error per 10^9 nucleotides, in each cell division (Alberts et al., 2002).

A higher error rate would, of course, produce more substitutions and thus a higher number of informative sites from which to infer a phylogenetic tree. Recent work has taken advantage of this by studying cancer cells, which have a higher mutation rate than healthy cells, at an estimated 1 error per 10^6 nucleotides (Carlson et al., 2012). Also, interesting scientific questions can be addressed by studying the phylogenetics of cancer cells, such as whether tumors have a monoclonal origin (Frumkin et al., 2008), the variable rate of mutation at different stages of the cancer (Navin et al., 2011), and mechanisms for relapse of acute leukemia (Shlush et al., 2012).

We examine data from the work of Carlson et al. (2012), in which deep DNA sequencing of experimentally reproducing mouse tumor cells was performed; because they were reproduced on petri dishes, the true cell lineage is known. In their study, Carlson et al. (2012) performed phylogenetic inference using a Bayesian framework (Rannala and Yang, 1996; Yang and Rannala, 1997), implemented in the MrBayes software (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012), to determine

whether they could reconstruct the true cell lineage using the observed mutations found from their deep DNA sequencing. Here, we revisit their analyses using our least squares estimation procedures.

5.4 Results

5.4.1 Simulations

First, we examine branch length estimation over the correct topology, through simulation of DNA sequence alignments along topologies of five, six and seven taxa, using the JC69 model and sequence lengths of 1000 nucleotides. Results from 1000 iterations with the five taxa, ULE2 scenario are shown in Figure 5.2. We show boxplots of the normalized errors from each iteration, which is calculated as

$$\text{Normalized error} = \frac{\hat{b}_i - b_i^{\text{true}}}{b_i^{\text{true}}}, \tag{5.24}$$

for each branch b_i and each method.

For each scenario, we then show the normalized error for branch length estimates on the three shortest branches of the true tree, in Figure 5.3. We observe that there is sometimes an improvement with respect to the bias, but always an improvement with respect to the variance, which is shown in most of the branch comparisons by the narrower interquartile range using the L_2 loss compared to the L_1 loss.

Next, we examine simultaneous topology and branch length estimation. Using the USI1 scenarios with four and five taxa, we simulate DNA sequence alignment data under the JC69 model, with alignment lengths of 200, 400, 600 and 800 sites. Then, out of 1000 replications, we count how many times each estimator obtained the correct topology. Results are shown in Figure 5.4.

In every scenario, our point estimate for the percentage of topologies correctly identified is higher using the L_2 loss function. However, we note that the 95% confidence intervals within each scenario always overlap, and p-values from a two-sample

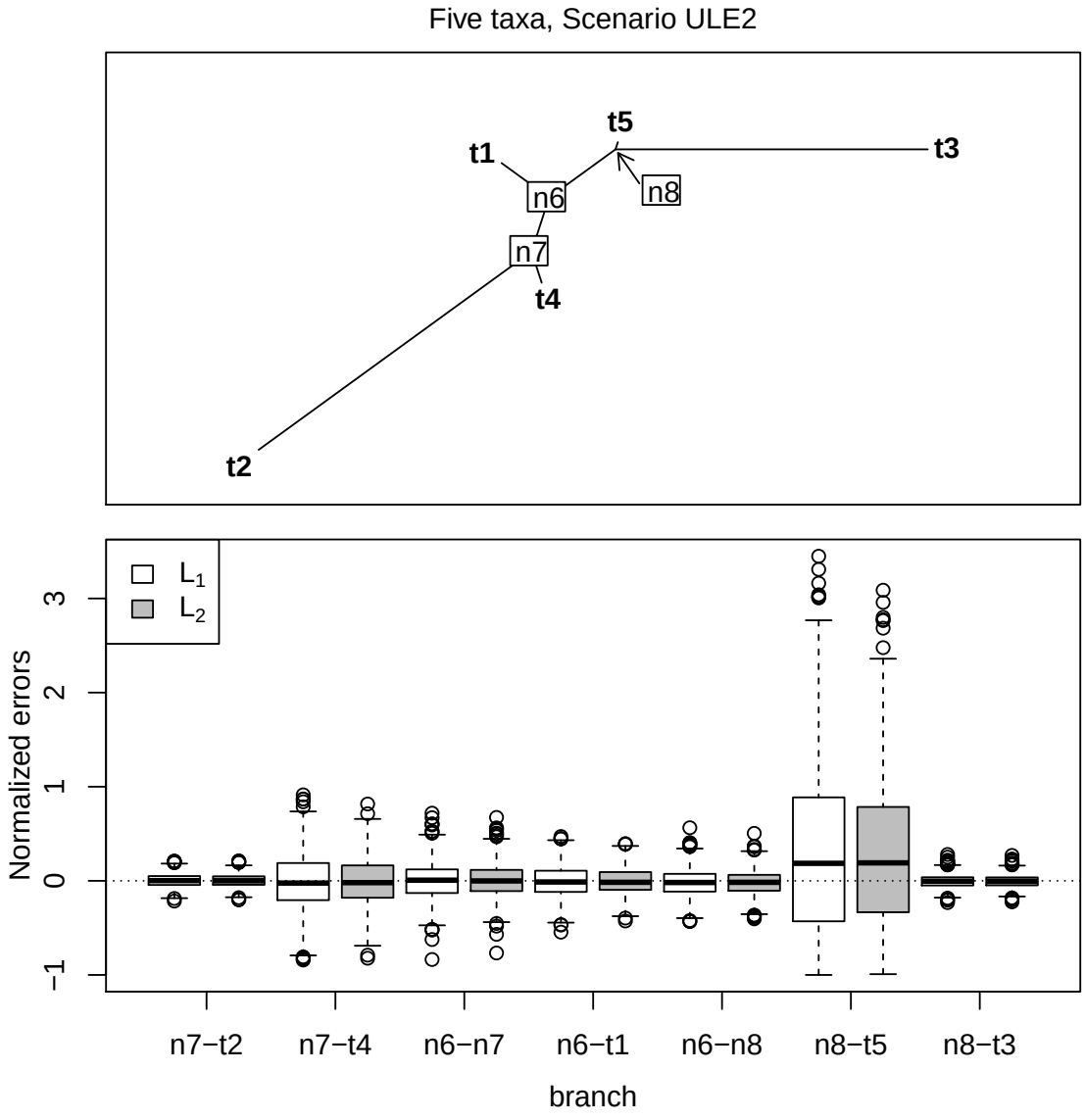

Figure 5.2: **Simulation results: five taxa ULE2**. The tree used for simulations under this scenario is shown in the top panel. For each method, normalized errors are calculated as $(\hat{b}_i - b_i^{\text{true}})/b_i^{\text{true}}$ for each branch b_i.

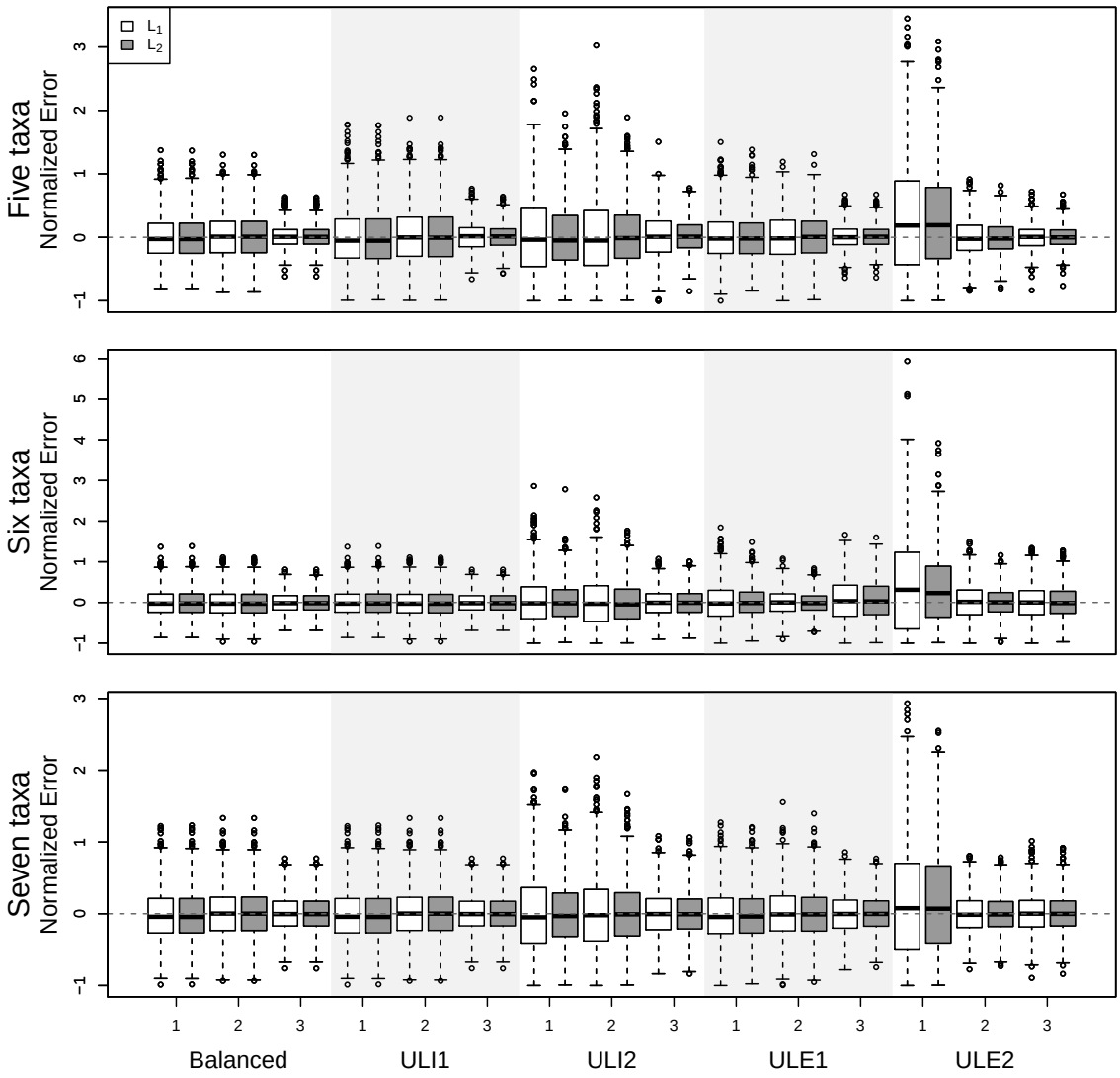

Figure 5.3: **Simulation results: fixed topology**. We report histograms of the smallest three branches (according to the true tree) for each scenario. Normalized Error is equal to $(\hat{b}_i - b_i^{\text{true}})/b_i^{\text{true}}$.

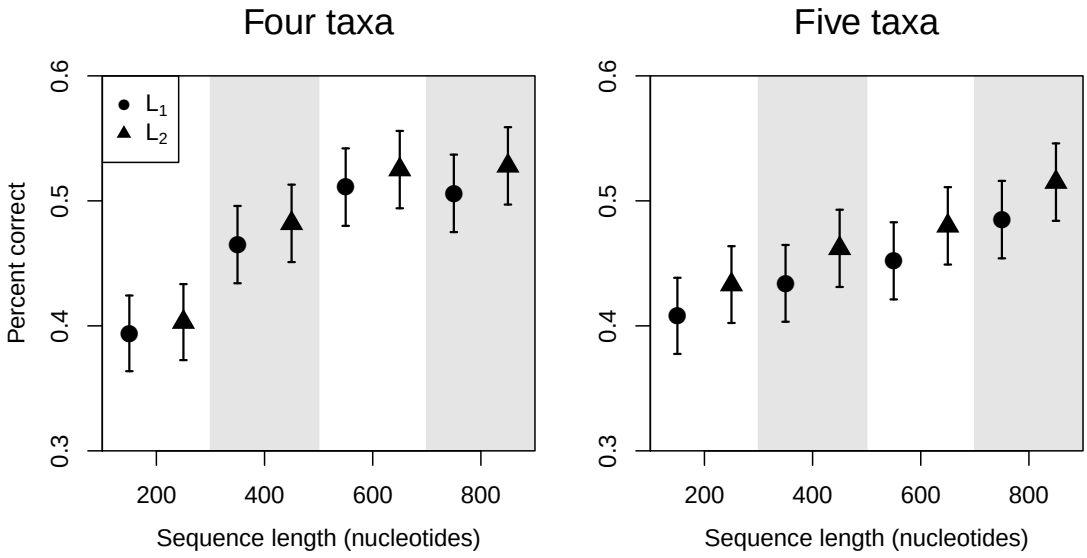

Figure 5.4: **Simulation results: topology estimation**. Using the USI1 scenario, we varied the length of the simulated sequence alignments and counted how many times each method produced the true topology. Error bars represent 95% confidence intervals. For each scenario, we simulated 1000 replicates.

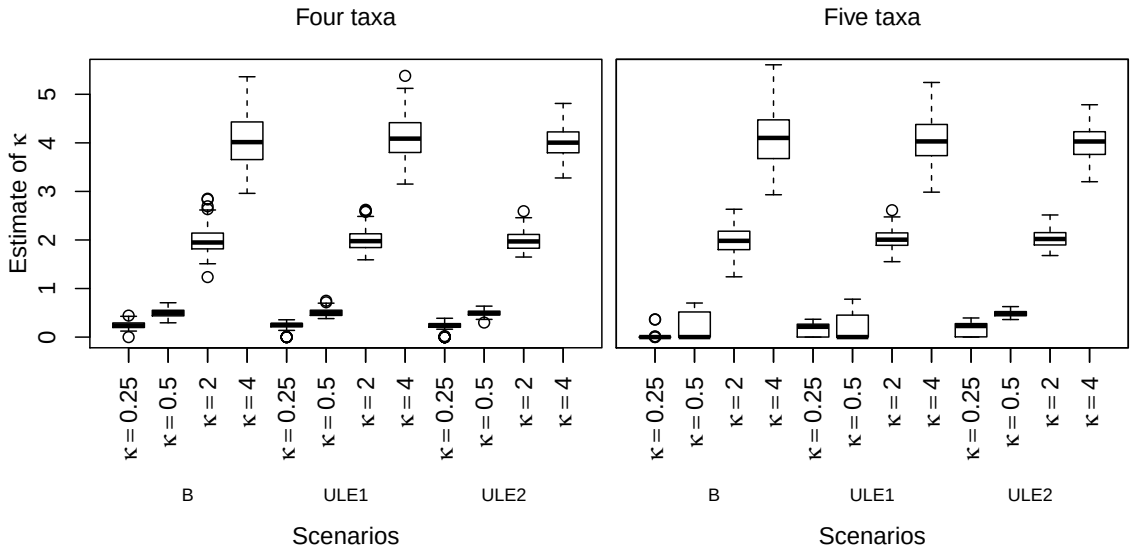

Figure 5.5: **Simulation results: K80**. Boxplots of κ estimates from 100 replications.

test of proportions are greater than 0.05 in each case; thus, we do not find a statistically significant improvement. Still, we note that the direction of the effect is the same in eight different scenarios, thus suggesting that perhaps some improvement in topology estimation is possible using the L_2 loss as compared to the original L_1 loss. While the differences are small, it appears likely that a statistically significant difference could be observed by increasing the number of simulation replicates.

Moving away from the JC69 model but returning to fixed topologies, we explore branch length optimization under the K80 model. Using $\kappa = 0.25, 0.5, 2$ and 4, we simulate DNA sequence alignment data with four and five taxa, under the B, ULE1 and ULE2 scenarios, with sequence lengths of 100 nucleotides, and 100 replications. Summaries of the κ estimates from each scenario are shown in Figure 5.5. We notice that we obtain κ estimates that are generally centered around the true values in all scenarios. However, branch length estimates (not shown) did not show conclusive improvement using the L_2 loss function compared to the L_1 loss function.

Finally, we explore branch length and substitution parameter estimation with rate heterogeneity, under the JC69+dΓ model. Using $\alpha = 0.5$ and $\alpha = 2$, we simulate DNA sequence alignment data with four and five taxa, under the ULE1 and ULE2 scenarios. We show estimates of α in Figure 5.6, from 100 simulated replicates. In addition to convergence issues with our optimization routine, we note that, in most cases, the central tendancies of our α estimates do not appear to be near the true value. Regarding branch length estimation, we observed evidence of improvement in terms of bias and variance for some, but not all, of the cases (not shown). Therefore, we cannot conclusively state that branch length estimation is currently improved by using the L_2 loss as compared to the L_1 loss. However, if we could estimate α with less bias and variance than we currently do, it is likely that the branch length estimates would improve accordingly.

5.4.2 Data Analysis: Mouse Tumor Cell Evolution

To investigate cellular evolution within a mouse tumor, Carlson et al. (2012) created a controlled cell lineage by allowing a single mouse tumor cell to reproduce on Petri dishes to produce a total of 15 cells. Using Bayesian phylogenetic inference (Rannala and Yang, 1996; Yang and Rannala, 1997), implemented in the MrBayes software (Ronquist and Huelsenbeck, 2003; Ronquist et al., 2012), Carlson et al. (2012) reproduced the known cell lineage with reasonable, but not perfect, accuracy.

In their work, Carlson et al. (2012) used only the variable sites to create their phylogenetic reconstructions. Thus, we proceed similarly, by creating a sequence alignment of only the variable sites throughout the tumor cell genome. This leads to an alignment of 592 nucleotides in length. We focus on the region of the tree for which their reconstruction was the most erroneous, with the cells labeled as 2, 4, 5, 8, 9 and 10, as can be observed in Figure 1 of the study by Carlson et al. (2012). Since the true tree is known, we can assess performance of our least squares estimators by calculating their respective bootstrap support values on the true tree, where sites from

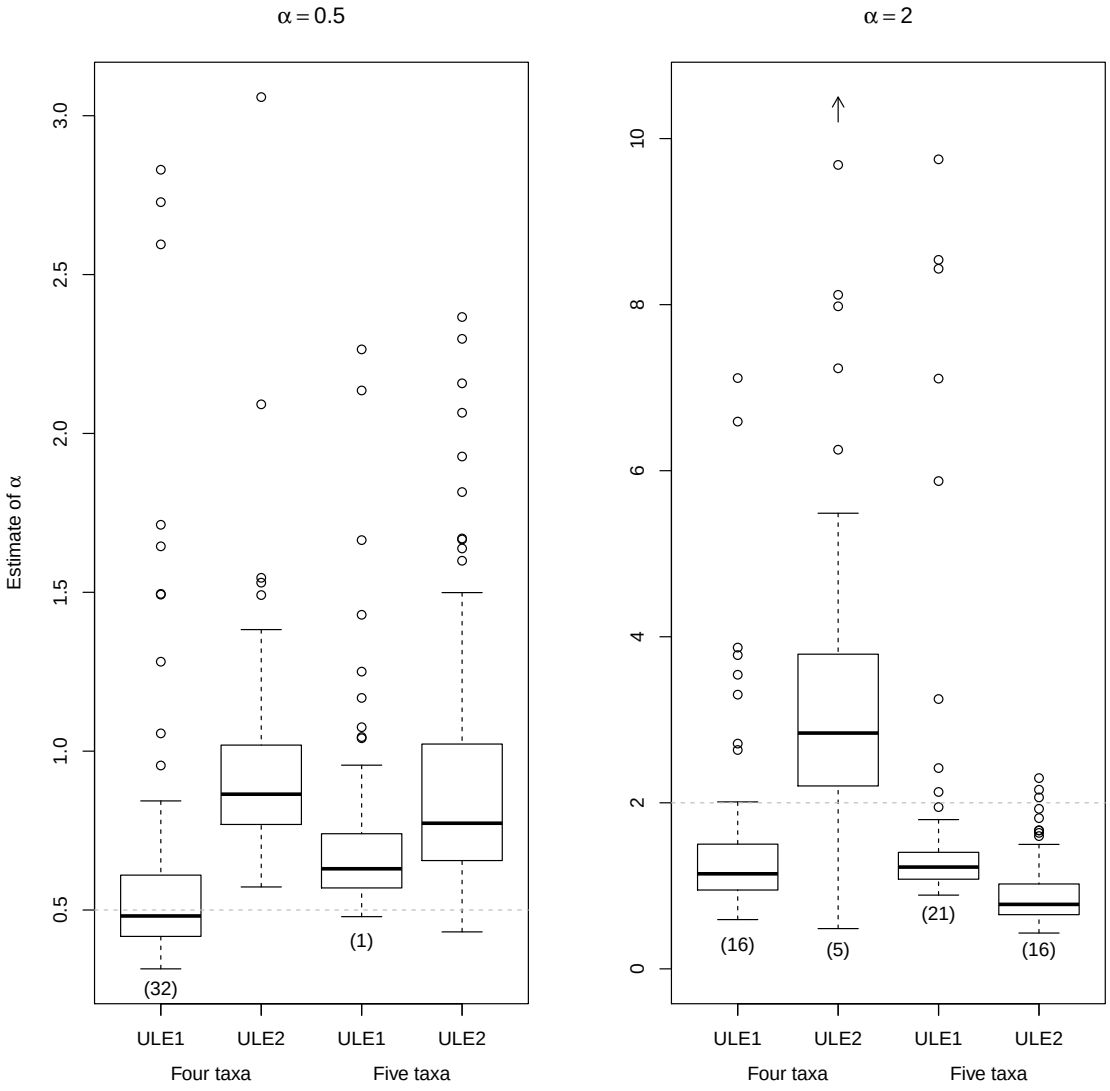

Figure 5.6: **Estimates of** α. Numbers in parentheses indicate how many iterations converged to the upper boundary of the box constraint, which are otherwise excluded on this plot. Upward arrow on the four taxa ULE2 scenario indicates three additional outliers that are above the upper limit of the plot, at values of 14.4, 26.3 and 43.9. In each scenario, we simulated 100 replications.

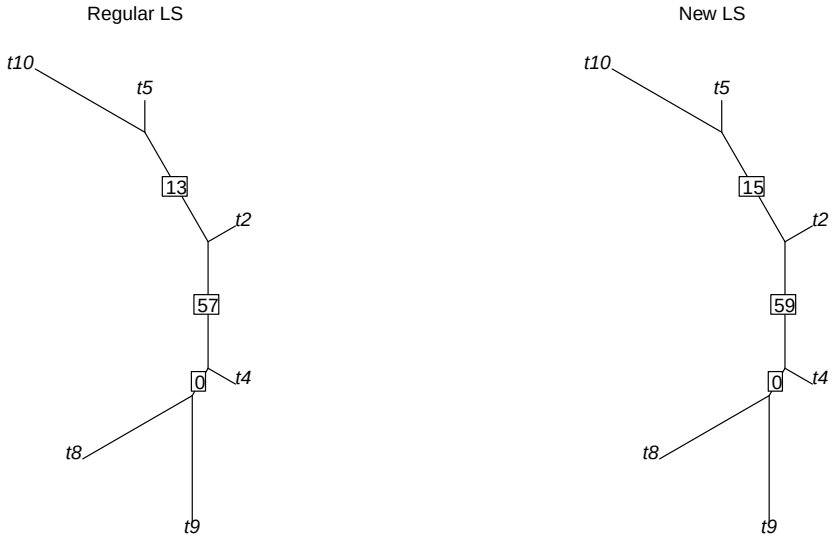

Figure 5.7: **Mouse tumor cell lineages**. We calculate bootstrap support for each bipartition on the true phylogeny, using both least squares estimators. For each case, $B = 500$.

the original sequence alignment are sampled at random, with replacement, and then bootstrap estimates of the tree are obtained. The bootstrap support values represent the proportion of times that each bipartition was represented in the bootstrap trees. Results are shown in Figure 5.7. With $B = 500$, we observe that using the L_2 loss results in estimation that favors the true tree more than estimation using the L_1 loss, although the differences are very small.

5.5 Discussion

Here, we propose a modification to least squares phylogenetic inference, in which we use a new loss function. We find that for branch length estimation under the simple JC69 model of DNA evolution, gains can be achieved in terms of bias, variance and MSE, as compared to the regular least squares estimator. For estimating topology, we find that our new least squares estimator correctly identifies the true topology

more frequently than the regular least squares estimator in every case, and while a statistically significant difference is not observed at $\alpha = 0.05$ with 1000 simulated replications, it appears likely that statistically significant differences could be obtained with a greater number of replications.

One drawback of our use of the L_2 loss function is that it is slower than the original least squares estimator using the L_1 loss function. Since regular least squares phylogenetic inference can be accomplished with simple matrix operations as an ordinary least squares problem, solutions are nearly instantaneous, especially for the small number of taxa in our experiments. While our new least squares estimator still performs on the order of seconds for problems of five to seven taxa, the difference becomes even more stark when dealing with more taxa, and with optimization over topologies; even with seven taxa, an exhaustive search over all possible unrooted bifurcating would give 945 topologies to search over. A real data problem would then call for bootstrap estimates of support for each split, which means that a full analysis using our method would take roughly 87.5 days, as opposed to < 1 day using the regular least squares estimator.

Even for the cases in which we fix the topology to the truth, we have limitations with regard to the number of taxa that our new estimator can currently handle. In the simplest JC69 case, branch length estimation using the L_2 loss slows down considerably with seven or more taxa, at about 80 seconds per iteration as opposed to 7.8 seconds per iteration for five taxa. Even more problematic is that convergence issues become more unruly; while the balanced case for seven taxa has no convergence problems, the more complex scenarios have a substantial proportion of failed convergences. In our current simulation study, we simply discarded these scenarios and continued until we obtained 1000 iterations with convergence. In addition to raising a concern with selection bias in our simulation study, it is even more worrisome that with real data, if convergence is not achieved, we cannot then just turn to our next simulated dataset as we do in our simulation study.

A potential solution to the issues of both speed and convergence is the development of a more appropriate optimization routine to minimize L_2 loss. Currently, we use the `nlminb` routine, and while it was the best performer among all those that we tested, we note that it was not specifically designed for this optimization problem. For example, Felsenstein (1997) proposed and developed an optimization routine specifically for least squares phylogenetic inference. A similar development of an optimization routine specific to our loss function would likely result in substantial gains both in terms of speed and convergence.

In addition to the development of an optimization routine specific to our loss function, one other possible solution in practice would be to first infer a tree with a fast algorithm such as Neighbor-joining (Saitou and Nei, 1987), use this as a starting point, and then perform a tree search using standard strategies such as nearest-neighbor interchanges, subtree pruning and regrafting, and tree bisection and reconnection (Felsenstein, 2004). Here, however, in our effort to demonstrate this proof-of-principle, we wanted to ensure that we obtained the true optimum at each iteration; an exhaustive search was thus necessary to achieve this aim.

Another issue arises from the fact that we must use box-constrained optimization for our new least squares estimator using L_2, since $L_2 \to 0$ as each $t_{kl}(\cdot, \cdot) \to \infty$. In our simulation studies, the usage of upper constraints has successfully produced well-behaved estimates of branch lengths, as we showed. However, it is fair to ask how one might know what to choose as upper constraint for a real data problem. Here, we choose an upper constraint of 1, which is well above the length of any of our branch lengths from the trees that we use in our simulations. Clearly, this knowledge may not be as readily available for a real data problem. On the other hand, since this is purely a computational issue, we do not feel that any statistical validity would be lost if one were to encounter the need to change the constraints *ad hoc* until the best local optimum that does not involve any branches equalling ∞ is obtained.

The framework of our new least squares estimator does have the potential for

increased flexibility, in terms of estimation under various models. We demonstrated that we can incorporate the appropriate parameters into the loss function for both the K80 and JC69+dΓ models, which cannot be done explicitly using standard evolutionary distances. Our results for estimation of κ under the K80 model are promising. However, whether this translates to better branch length estimation is still uncertain. For estimation of α under the JC69+dΓ model, we observed that our L_2 loss function produced estimates that tended to be biased, have large variance and even converge to the upper boundary of the box constraint. One puzzling aspect is that as $\alpha \to \infty$, the model becomes JC69 without any rate heterogeneity, so it seems that even when estimates of α converged to the upper boundary, our resulting branch length estimates should at least enjoy the same advantage over those from the L_1 loss function that we observed in the first set of simulations from Section 5.4.1. This did not appear to be the case. Further studies could aim to resolve this issue.

As we saw in Chapter 4, it can be of interest to base phylogenetic inference on synonymous substitutions, if we want to avoid the footprint of selective pressures. Currently, distance-based inference is the only choice when one wishes to distinguish between synonymous and nonsynonymous substitutions; the incorporation of labeled substitutions has not yet been demonstrated in the likelihood setting. Within our new framework proposed here, it is a natural extension to estimate a phylogeny based on synonymous substitutions. This may be desired if the aim is to infer a phylogeny on a group of taxa that is known to be under convergent selective pressures, in order to represent the true evolutionary history as opposed to one that is influenced by selection.

We have shown a proof of principle in this study, that least squares phylogenetic inference can be improved in terms of bias, variance and MSE using our proposed loss function. We also demonstrate the added modeling flexibility under this framework. However, there are some computational issues that need to be resolved before our new method can become more widely available for use, which may be the focus of

future study

Chapter 6

FUTURE DIRECTIONS

6.1 Finer distinction between Recombination and Convergent Evolution

6.1.1 Convergent Evolution as the Null Hypothesis

In Chapter 4, we proposed the synonymous Dss statistic for recombination detection that avoids false positives due to convergent evolution. This led us to the ability to identify whether tree incongruence was due to recombination or convergent evolution. However, we could only do so indirectly; that is, with each test, our actual null hypothesis is that there is neither recombination nor convergent evolution. It is only by comparing the results from the original Dss statistic to that of the synonymous Dss statistic that we are able to make any conclusions about convergent evolution. Yet, we do not actually have an ability to assign statistical significance to this conclusion.

To obtain such significance assessment would require that we formulate our hypothesis test with the null hypothesis of either "recombination but no convergent evolution" or "convergent evolution but no recombination," and then test for the one that is not present under this null hypothesis. For example, with our synonymous Dss statistic, we currently detect recombination while avoiding the signal of convergent evolution. However, our Type I error rate of the synonymous Dss statistic was not truly calibrated to our desired $\alpha = 0.05$ for the convergent evolution scenario; it was calibrated for the original null hypothesis of the absence of both recombination and convergent evolution.

Furthermore, while the simulation of a recombination signal by using incongruent trees for each part of a sequence alignment is well-justified in accordance to the true

mechanism of recombination, how to properly simulate convergent evolution is less clear. Here, we make single nucleotide changes on two different taxa to substitute differing amino acids into the same amino acid state. While the outcome of our approach appears to be the same as that of convergent evolution, there are nuances that we may be ignoring, such as the possibility that specific substitution patterns could be likely through convergent evolution. Therefore, the development of a reasonable model for convergent evolution may be an important next step. This would, in turn, enable us to establish a stable type I error rate of our synonymous Dss statistic under the null hypothesis of "convergent evolution but no recombination."

6.1.2 Using Different Distances Within Each Window

Alternatively, we also propose an additional modification to the Dss statistic. Currently, the Dss statistic uses one distance measure for each window. That is, the original Dss statistic uses standard evolutionary distances for each window, the synonymous Dss statistic uses synonymous distances for each window, and the nonsynonymous Dss statistic uses nonsynonymous distances for each window. It may be of benefit to examine topologies for each window that are based on synonymous distances and nonsynonymous distances, separately. That is, for a particular window, we would obtain $\hat{\tau}_{\text{Dss.syn}}$ and $\hat{\tau}_{\text{Dss.nonsyn}}$, using labeled distances as described in Chapter 4.

Then, we can set our null hypothesis to include "recombination but no convergent evolution." Under this null hypothesis, we would expect $\hat{\tau}_{\text{Dss}}$ and $\hat{\tau}_{\text{Dss.syn}}$ to be similar, both for windows that do contain a recombination breakpoint and those that do not. On the other hand, for windows in which convergent evolution has occurred, $\hat{\tau}_{\text{Dss}}$ and $\hat{\tau}_{\text{Dss.syn}}$ should be dissimilar. Our test statistic could then be based on a measure of differences between topologies (Felsenstein, 2004, Chapter 30). That is, if convergent evolution has not occurred, then the difference between topologies should be small for each window across the sequence alignment, and therefore the test statistic based on

these differences should be small. On the other hand, if there is convergent evolution, then the difference between topologies should be large, particularly in the region(s) where convergent evolution has occurred; therefore, the test statistic based on these differences should be large, in the corresponding windows.

6.2 Further Improvements to Least Squares Phylogenetic Inference

6.2.1 Improving optimization of the least squares criterion

In Chapter 5, we discussed limitations due to the fact that an exhaustive tree search is time consuming. While one potential strategy is to use the neighbor-joining tree as the starting tree, we believe that it may be even more beneficial to employ other strategies. With current least squares phylogenetic inference, recall that the goal is to find the solution to

$$\underset{\boldsymbol{\tau},\mathbf{b}}{\operatorname{argmin}} \sum_{k=1}^{n}\sum_{l=1}^{n}\left[\hat{d}_{kl} - t_{kl}(\boldsymbol{\tau},\mathbf{b})\right]^2, \tag{6.1}$$

where $\boldsymbol{\tau}$ is the tree topology, and $\mathbf{b}$ is the set of branch lengths. The most straightforward way to do this is to find $\hat{\mathbf{b}}$ that minimizes the criterion for any particular $\boldsymbol{\tau}$, and then use nearest-neighbor interchanges or some other heuristic algorithm to move through tree space to search for the optimal topology, estimating $\mathbf{b}$ on each one. However, this strategy may use more time than is needed on each candidate topology. Specifically, it has been pointed out in the case of maximum likelihood that branch length estimation does not need to be fully optimized on every candidate tree along the search, especially those which are highly unlikely to be the true tree (Salter and Pearl, 2001; Guindon and Gascuel, 2003). Thus, an analogous solution may be available in the case of our new least squares phylogenetic estimator.

A first step towards answering this question might be to examine the landscape of the L_2 loss function for a given set of branch lengths, over all candidate topologies for a given set of taxa. That is, suppose we have a sequence alignment of seven taxa. There

are thus 945 unrooted bifurcating topologies to consider. While optimizing the L_2 loss over all topologies while estimating branch lengths would be very time-consuming, it would be relatively fast to simply calculate the L_2 loss over all 945 topologies for a given set of branch lengths (on the order of seconds). If the sequence alignment was simulated from a known tree, or if the true evolutionary history was otherwise known, we could then examine how the true topology fared in this experiment.

Of course, this would be influenced by the choice of branch lengths; if they are near the true branch lengths, then clearly the L_2 loss in this experiment would be close to its global minimum and we would practically obtain the true least squares solution. On the other hand, if the chosen branch lengths are very far from the true branch lengths, then the L_2 loss might not be minimized at the true topology in this experiment. However, it might be interesting to determine the relative value of L_2 for the true topology compared to all other topologies, for any given set of branch lengths. For example, if the true topology tends to be within the best 5% of all topologies in terms of L_2 for any given set of branch lengths, then this would lead us to a more optimal search strategy: first, determine the top 5% of topologies for L_2 with any given set of branch lengths, and then narrow the search down to this 5% when estimating branch lengths.

Additionally, we note that when optimizing the L_2 loss over branch lengths for a particular topology, we are using a pre-prescribed optimization routine (`nlminb`). We believe that we could obtain better performance by writing an optimization routine that is specific to our optimization problem. An analogous proposition for the original least squares estimator using L_1 loss was put forth by Felsenstein (1997). Thus, we believe that similar gains could be obtained in this manner for least squares phylogenetic estimation with the L_2 loss.

6.2.2 Constraint by penalty

As an alternative to using Box constraints to avoid the issue of $t_{kl}(\boldsymbol{\tau}, \mathbf{b}) \to \infty$, we also propose the use of a penalized loss function. That is, consider

$$L_3(\mathbf{Y}, \mathbf{b}, \boldsymbol{\tau}, \boldsymbol{\theta}) = \sum_{k=1}^{n} \sum_{l=1}^{n} \left(\hat{e}_{kl}(\boldsymbol{\tau}, \mathbf{b}, \boldsymbol{\theta}) - t_{kl}(\boldsymbol{\tau}, \mathbf{b})\right)^2 + \sum_{j} b_j. \tag{6.2}$$

Then, while the L_2 loss function was minimized as $t_{kl}(\boldsymbol{\tau}, \mathbf{b}) \to \infty$, this would not be true with the L_3 loss. We aim to explore option in future studies.

6.2.3 Inferring labeled phylogenies

As discussed in Chapter 4, phylogenies based on synonymous substitutions are more likely to uncover the true evolutionary history, if convergent selective pressures are present. Our new distances $\hat{e}_{kl}$ allow for coherent estimation of labeled distances, which can thus be used to infer such phylogenies. However, it is important to note that branch lengths under labeled distances are defined as the expected number of labeled substitutions per site, as opposed to the total expected number of substitutions per site. Thus, we propose a strategy in which first a topology is estimated using synonymous distances. Then, once the synonymous topology is obtained, we can then estimate branch lengths using unlabeled distances, to obtain branch lengths on the standard scale of expected number of substitutions per site.

BIBLIOGRAPHY

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88:669–679, 1993.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Molecular Biology of the Cell. Garland Science, 2002.

C. Althouse, S. Patterson, P. Fedorka-Cray, and R.E. Isaacson. Type 1 fimbriae of *Salmonella enterica* serovar typhimurium bind to enterocytes and contribute to colonozation of swine in vivo. Infection and Immunity, 71:6446–6452, 2003.

M. Anisimova, R. Nielsen, and Z. Yang. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics, 164:1229–1236, 2003.

K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica, 25:251–278, 1999.

P. Awadalla. The evolutionary genomics of pathogen recombination. Nature Reviews, 4:50–60, 2003.

F. Ball and R. Milne. Simple derivations of properties of counting processes associated with Markov renewal processes. Journal of Applied Probability, 42:1031–1043, 2005.

A.J. Bäumler, R.M. Tsolis, and F. Heffron. Fimbrial adhesins of *Salmonella typhimurium*. role in bacterial interactions with epithelial cells. Advances in Experimental Medicine and Biology, 412:149–158, 1997.

E.G. Birgin, J.M. Martínez, and M. Raydan. Nonmonotone spectral gradient methods on convex sets. SIAM Journal on Optimization, 10:1196–1211, 2000.

J.D. Boddicker, N.A. Ledeboer, J. Jagnow, B.D. Jones, and S. Clegg. Differential binding to and biofilm formation on, HEp-2 cells by *Salmonella enterica* serovar Typhimurium is dependent upon allelic variation in the *fimH* gene of the *fim* gene cluster. Molecular Microbiology, 45:1255–1265, 2002.

M.J. Box. A new method of constrained optimization and a comparison with other methods. The Computer Journal, 8:42–52, 1965.

P.S. Bridges. Vertebral arthritis and physical activities in the prehistoric southeastern United States. American Journal of Physical Anthropology, 93:83–93, 1994.

P.R. Burton, J. Bowden, and M.D. Tobin. Epidemiology and Genetic Epidemiology. Chapter in Handbook of Statistical Genetics. Wiley, 2007.

R.H. Byrd, P. Lu, J Nocedal, and C.Y. Zhu. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing, 16:1190–1208, 1995.

J.H. Camin and R.R. Sokal. A method for deducing branching sequences in phylogeny. Evolution, 19:311–326, 1965.

P.S. Caplan, L.M. Freedman, and T.P. Connelly. Degenerative joint disease of the lumbar spine in coal miners - a clinical and X-ray study. Arthritis & Rheumatism, 9:693–702, 1966.

C.A. Carlson, A. Kas, R. Kirkwood, L.E. Hays, B.D. Preston, S.J. Salipante, and M.S. Horwitz. Decoding cell lineage from acquired mutations using arbitrary deep sequencing. Nature Methods, 9:78–82, 2012.

L.L Cavalli-Sforza and A.W.F. Edwards. Phylogenetic analysis: models and estimation procedures. Evolution, 21:550–570, 1967.

J.A. Cavender. Taxonomy with confidence. Mathematical Biosciences, 40:271–280, 1978.

J.T. Chang. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Mathematical Biosciences, 134:189–215, 1996.

P.A. Christin, G. Besnard, E.J. Edwards, and N. Salamin. Effect of genetic convergence on phylogenetic inference. Molecular Phylogenetics and Evolution, 62: 921–927, 2012.

Y.C. Chuang, K.C.Wang, Y.T. Chen, C.H. Yang, S.C. Men, C.C. Fan, L-H Chang, and K-S Yeh. Identification of the genetic determinants of *Salmonella enterica* serotype Typhimurium that may regulate the expression of the type 1 fimbriae in response to solid agar and static broth culture conditions. BMC Microbiology, 8: 126, 2008.

C. Clegg and K.T. Hughes. FimZ is a molecular link between sticking and swimming in *Salmonella enterica* serovar Typhimurium. Journal of Bacteriology, 184:1209–1213, 2002.

J.M. Coffin, S.H. Hughes, and H.E. Vamus. Retroviruses. Cold Spring Harbor Laboratory Press, 1997.

E.L. Cohn, E.J. Maurer, T.E. Keats, R.G. Dussault, and P.A. Kaplan. Plain film evaluation of degenerative disk disease at the lumbosarcal junction. Skeletal Radiology, 26:161–166, 1997.

T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. Introduction to Algorithms. MIT Press, 2009.

M.K. Cowles. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Statistical Computing, 6:101–111, 1996.

H. Cramér. Mathematical Methods of Statistics. Princeton University Press, 1945.

K.A. Crandall, C.R. Kelsey, H. Imamichi, H.C. Lane, and N.P. Salzman. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Molecular Biology and Evolution, 16:372–382, 1999.

E.R. Dempster and I.M. Lerner. Heritability of threshold characters. Genetics, 35: 212–236, 1950.

C.J. DeRousseau. Aging in the musculoskeletal system of rhesus monkeys: II. Degenerative joint disease. American Journal of Physical Anthropology, 67:177–184, 1985.

A.J. Drummond, M.A. Suchard, D. Xie, and A. Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29:1969–1973, 2012.

A.E. Duncan, R.J. Colman, and P.A. Kramer. Longitudinal study of radiographic spinal osteoarthritis in a macaque model. Journal of Orthopaedic Research, 29: 1152–1160, 2011.

A.E. Duncan, R.J. Colman, and P.A. Kramer. Sex differences in spinal osteoarthritis in humans and rhesus monkeys (*Macaca mulatta*). Spine, 15:915–922, 2012.

A.W.F. Edwards and L.L. Cavalli-Sforza. Phenetic and Phylogenetic Classification. Systematics Association Publications, 1964.

B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences, 93:13429–13434, 1996.

D.S. Falconer and T.F.C. Mackay. Quantitative Genetics. Benjamin Cummings, 1996.

J.S. Farris. Methods for computing Wagner trees. Systematic Zoology, 19:83–92, 1970.

J. Felsenstein. PHYLIP - phylogeny inference package (Version 3.2). Cladistics, 5: 164–166, 1989.

J.F. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology, 27:401–410, 1978.

J.F. Felsenstein. Evoluationary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution, 17:368–376, 1981.

J.F. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39:783–791, 1985.

J.F. Felsenstein. An alternating least squares approach to inferring phylogenies from pairwise distances. Systematic Biology, 46:101–111, 1997.

J.F. Felsenstein. Inferring Phylogenies. Sinauer Associates, 2004.

J.F. Felsenstein. Using the quantitative genetic threshold model for inferences between and within species. Philosophical Transactions of the Royal Society, 360:1427–1434, 2005.

J.F. Felsenstein. A comparative method for both discrete and continuous characters using the threshold model. The American Naturalist, 179:145–156, 2012.

J.F. Felsenstein and G.A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. Molecular Biology and Evolution, 13:93–104, 1996.

J.F. Felsenstein and H. Kishino. Is there something wrong with the bootstrap on phylogenies. Systematic Biology, 43:193–200, 1993.

R.A. Fisher. On an absolute criterion for fitting frequency curves. Messenger of Mathematics, 41:155–160, 1912.

R.A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh, 52:399–433, 1918.

W.M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology, 20:406–416, 1971.

W.M. Fitch and E. Margoliash. Construction of phylogenetic trees. Science, 155: 279–284, 1967.

J.L. Foulley, D. Gianola, and S. Im. Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters. Theoretical and Applied Genetics, 73:870–877, 1987.

P. Fox. The Port Mathematical Subroutine Library, Version 3. AT&T Bell Laboratories, Murray Hill, NJ, 1997.

D. Frumkin, A. Wasserstrom, S. Itzkovitz, T. Stern, A. Harmelin, R. Eilam, G. Rechavi, and E. Shapiro. Cell lineage analysis of a mouse tumor. Cancer Research, 68:5924–5931, 2008.

J.W. Frymoyer, A. Newberg, M.H. Pope, D.G. Wilder, J. Clements, and B. MacPherson. Spine radiographs in patients with low-back pain: an epidemiological study in men. The Journal of Bone and Joint Surgery, 66:1048–1055, 1984.

J.E. Galán. Molecular genetic bases of *Salmonella* entry into host cells. Molecular Microbiology, 20:263–271, 1997.

D. Gianola. Heritability of polychotomous characters. Genetics, 93:1051–1055, 1979.

D. Gianola. Theory and analysis of threshold characters. Journal of Animal Science, 54:1079–1096, 1982.

A.R. Gilmour, R. Thompson, and B.R. Cullis. Average information REML, an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics, 51:1440–1450, 1995.

D.S. Gladstein. Efficient incremental character optimization. Cladistics, 13:21–26, 1997.

N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution, 11:725–736, 1994.

P.A. Goloboff, J.S. Farris, and K.C. Nixon. TNT, a free program for phylogenetic analysis. Cladistics, 24:774–786, 2008.

M.J. Gonzales, J.M. Dugan, and R.W. Shafer. Synonymous-non-synonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN). Bioinformatics, 18:886–887, 2002.

N.C. Grassly and E.C. Holmes. A likelihood method for the detection of selection and recombination using nucleotide sequences. Molecular Biology and Evolution, 14:239–247, 1997.

E.A. Groisman and H. Ochman. How *Salmonella* became a pathogen. Trends in Microbiology, 5:343–349, 1997.

X. Gu and W.H. Li. A general additive distance with time-reversibility and rate variation among nucleotide sites. Proceedings of the National Academy of Sciences, 93:4671–4676, 1996.

X. Gu, Y.X. Fu, and W.H. Li. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Molecular Biology and Evolution, 12: 546–557, 1995.

S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology, 52:696–704, 2003.

S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihoode phylogenies: assessing the performance of PhyML 3.0. Systematic Biology, 59:307–321, 2010.

P. Guttorp. Stochastic Modeling of Scientific Data. Chapman & Hall, 1995.

Jarrod Hadfield. Mcmcglmm course notes, 2011.

J.D. Hadfield. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. Journal of Statistical Software, 33(2):1–22, 2010.

A.G. Hadjipavlou, J.W. Simmons, M.H. Pope, J.T. Necessary, and V.K. Goel. Pathomechanics and clinical relevance of disc degeneration and annular tear: a point-of-view review. American Journal of Orthopedics, 28:561–571, 1999.

D.A. Harville and R.W. Mee. A mixed-model procedure for analyzing ordered categorical data. Biometrics, 40:393–408, 1984.

M. Hasegawa and M. Fujiwara. Relative efficiencies of the maximum likelihood, maximum parsimony and neighbor joining methods for estimating protein phylogeny. Molecular Phylogenetics and Evolution, 2:1–5, 1993.

M. Hasegawa and H. Kishino. Confidence limits on the maximum-likelihood estimate of the Homonoid tree from mitochondrial DNA sequences. Evolution, 43:672–677, 1989.

M. Hill, G. Tachedjian, and J. Mak. The packaging and maturation of the HIV-1 Pol proteins. Current HIV Research, 3:73–85, 2005.

S. Ho. The molecular clock and estimating species divergence. Nature Education, 1: 1, 2008.

K.E. Holt, N.R. Thomson, J. Wain, G.C. Langridge, R. Hasan, Z.A. Bhutta, M.A. Quail, H. Norbertczak, D. Walker, M. Simmonds, B. White, N. Bason, K. Mungall, G. Dougan, and J. Parkhill. Psuedogene accumulation in the evolutionary histories of *Salmonella enterica* serovars paratyphi a and typhi. BMC Genomics, 10:36, 2009.

I Höschele. Estimation of breeding values and variance components with quasi-continuous data. PhD thesis, Universität Hohenheim, Germany, 1986.

J.P. Huelsenbeck. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Molecular Biology and Evolution, 12:843–849, 1995a.

J.P. Huelsenbeck. The performance of phylogenetic methods in simulation. Systematic Biology, 44:17–48, 1995b.

J.P. Huelsenbeck, F. Ronquist, R. Nielsen, and J.P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294:2310–2314, 2001.

D. Husmeier and F. Wright. Probabilistic divergence measures for detecting interspecies recombination. Bioinformatics, 17:S123–S131, 2001.

D. Husmeier and F. Wright. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Molecular Biology and Evolution, 20:315–337, 2003.

A. Jacquard. Logique du calcul des coefficients d'identité entre deux individus. Population (French Edition), 21(4):751–776, Jul-Aug 1966.

L. Jin and M. Nei. Limitations of the evolutionary parsimony method of phylogenetic analysis. Molecular Biology and Evolution, 7:82–102, 1990.

V.A. Johnson, F. Brun-Vezinet, B. Clotet, B. Conway, R.T. D'Aquila, L.M. Demeter,

D.R. Kuritzkes, D. Pillay, J.M. Schapiro, A. Telenti, and D.D. Richman. Drug resistance mutations in HIV-1. Topics in HIV Medicine, 11:215–221, 2003.

M.D. Jones, M.J. Pais, and B. Omiya. Bony overgrowths and abnormal calcifications about the spine. Radiology Clinics of North America, 26:1213–1234, 1988.

T.H. Jukes. Transitions, transversions, and the molecular evolutionary clock. Journal of Molecular Evolution, 26:87–98, 1987.

T.H. Jukes and C.R. Cantor. Mammalian Protein Metabolism, volume III. Academic Press, 1969.

R.D. Jurmain and L. Kilgore. Skeletal evidence of osteoarthritis: a paleopathological perspective. Annals of Rheumatic Diseases, 54:443–450, 1995.

C.T. Kelley. Iterative Methods for Optimization. SIAM, 1999.

C. Kelly and J. Rice. Modeling nucleotide evolution: a heterogeneous rate analysis. Mathematical Biosciences, 133:85–109, 1996.

L.I. Kerttula, W.S. Serlo, O.A. Tervonen, E.L. Pääkkö, and H.V. Vanharanta. Post-traumatic findings of the spine after earlier vertebral fracture in young patients: clinical and MRI study. Spine, 25:1104–1108, 2000.

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

D.I. Kisiela, S. Chattopadhyay, S.J. Libby, J.E. Karlinsey, F.C. Fang, V. Tchesnokova, J.J. Kramer, V. Beskhlebnaya, M. Samadpour, K. Grzymajlo, M. Ugorski, E.W. Lankau, R.I. Mackie, S. Clegg, and E.V. Sokurenko. Evolution of *Salmonella enterica* virulence via point mutations in the fimbrial adhesion. PLoS Pathogens, 8: e1002733, 2012.

A.G. Kluge and J.S. Farris. Quantitative phyletics and the evolution of anurans. Systematic Zoology, 18:1–32, 1969.

C.J. Knusel, S. Goggel, and D. Lucy. Comparative degenerative joint disease of the vertebral column in the medieval monastic cemetery of the Gilbertine priory of St. Andrew, Fishergate, York, England. American Journal of Physical Anthropology, 103:481–495, 1997.

P.A. Kramer, L.L. Newell-Morris, and P.A. Simkin. Spinal degenerative disk disease (ddd) in female macaque monkeys: epidemiology and comparison with women. Journal of Orthopaedic Research, 20:399–408, 2002.

M.K. Kuhner and J.F. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution, 11: 459–468, 1994.

L.Y. Kwan and R.E. Isaacson. Identification and characterization of a phase-variable nonfimbrial *Salmonella typhimurium* gene that alters O-antigen production. Infection and Immunology, 66:5725–5730, 1998.

M.R. Lacey and J.T. Chang. A signal-to-noise analysis of phylogenetic estimation by neighbor-joining: Insufficiency of polynomial length sequences. Mathematical Biosciences, 199:188–215, 2006.

K. Lange. Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health. Springer, second edition, 2002.

K. Lange, J. Westlake, and M.A. Spence. Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet, 39(4):485–491, 1976.

K. Lange, R. Cantor, S. Horvath, M. Perola, C. Sabatti, J. Sinsheimer, and E. Sobel. MENDEL version 4.0: A complete package for the exact genetic analysis of discrete

traits in pedigree and population data sets. American Journal of Human Genetics, 69(supplement):504, 2001.

J.S. Lawrence. Disc degeneration: its frequency and relationship to symptoms. Annals of the Rheumatic Diseases, 28:121–138, 1969.

P. Lemey, I. Derdelinckx, A. Rambaut, K. Van Laethem, S. Dumont, S. Vermeulen, and E. Van Wijngaerden. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. Journal of Virology, 79:11981–11989, 2005.

P. Lemey, A. Rambaut, A.J. Drummond, and M.A. Suchard. Bayesian phylogeography finds its roots. PLoS Computational Biology, 5:e1000520, 2009.

S. Li, D.K. Pearl, and H. Doss. Phylogenetic tree construction using Markov chain Monte Carlo. Journal of the American Statistical Association, 95:493–508, 2000.

M.F. Luo, P.J. Boettcher, L.R. Schaeffer, and J.C.M. Dekkers. Bayesian inference for categorical traits with an application to variance component estimation. Journal of Dairy Science, 84:694–704, 2001.

T. Margush and F.R. McMorris. Consensus n-trees. Bulletin of Mathematical Biology, 43:239–244, 1981.

I. Martyn and M. Steel. The impact and interplay of long and short branches on phylogenetic information content. Journal of Theoretical Biology, 314:157–163, 2012.

C.A.P. Matos, D.L. Thomas, D. Gianola, R.J. Tempelman, and L.D. Young. Genetic analysis of discrete reproductive traits in sheep using linear and nonlinear models: I. estimation of genetic parameters. Journal of Animal Science, 75:76–87, 1997.

B. Mau and M.A. Newton. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. Journal of Computational and Graphical Statistics, 2:122–131, 1997.

G. McGuire and F. Wright. TOPAL 2.0: improved detection of mosaic sequences within muliple alignments. Bioinformatics, 16:130–134, 2000.

G. McGuire, F. Wright, and M.J. Prentice. A graphical method for detecting recombination in phylogenetic data sets. Molecular Biology and Evolution, 14:1125–1131, 1997.

C.D. Michener and R.R. Sokal. A quantitative approach to a problem in classification. Evolution, 11:130–162, 1957.

J.A. Miller, C. Schmatz, and A.B. Schultz. Lumbar disc degeneration: correlation with age, sex, and spine level in 600 autopsy specimens. Spine, 13:173–178, 1988.

I. Milne, F. Wright, G. Rowe, D.F. Marshall, D. Husmeier, and G. McGuire. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. Bioinformatics, 20:1806–1807, 2004.

V.N. Minin and M.A. Suchard. Counting labeled transitions in continuous-time Markov models of evolution. Journal of Mathematical Biology, 56:391–412, 2008.

V.N. Minin, K.S. Dorman, F. Fang, and M.A. Suchard. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics, 21:3034–3042, 2005.

T. Miyata and T. Yasunaga. Molecular evolution of mRNA: a method for estimating evolutionary rates of synoymous and amino acid substitutions from homologous nucleotide sequences and its applications. Journal of Molecular Evolution, 16:23–36, 1980.

I. Mizstal, D. Gianola, and J.L. Foulley. Computing aspects of a nonlinear method of sire evaluation for categorical data. Journal of Dairy Science, 72:1557–1568, 1989.

G.W. Moore, M. Goodman, and J. Barnabas. An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular datasets. Journal of Theoretical Biology, 38:423–457, 1973.

S. Muse and B. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substituion rates, with application to the chloroplast genome. Molecular Biology and Evolution, 11:715–724, 1994.

J.C. Nash. Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation. Adam Hilger, Institute of Physics Publications, Bristol, 1979.

P.J. Naughton, G. Grant, S. Bardocz, E. Allen-Vercoe, M.J. Woodward, and A. Pusztai. Expression of type 1 fimbriae (SEF 21) of *Salmonella enterica* serotype Enteritidis in the early colonisation of the rat intestine. Journal of Medical Microbiology, 50:191–197, 2001.

N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W.R. McCombie, J. Hicks, and M. Wigler. Tumor evolution inferred by single-cell sequencing. Nature, 472:90–95, 2011.

R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics, 148:929–936, 1998.

J.D. O'Brien, V.N. Minin, and M.A. Suchard. Learning to count: robust estimates for labeled distances between molecular sequences. Molecular Biology and Evolution, 26:801–814, 2009.

J. Ødegård, T.H.E. Meuwissen, B. Heringstad, and P. Madsen. A simple algorithm to estimate genetic variance in an animal threshold model using Bayesian inference. Genetics Selection Evolution, 42:29, 2010.

D. Penny, M.D Hendy, and M.A. Steel. Progress with methods for constructing evolutionary trees. Trends in Ecology and Evolution, 7:73–79, 1992.

D. Posada and K.A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences, 98:13757–13762, 2001.

D. Posada and K.A. Crandall. The effect of recombination on the accuracy of phylogenetic estimation. Journal of Molecular Evolution, 54:396–402, 2002.

M.J.D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Centre for Mathematical Sciences, University of Cambridge, Report No. DAMTP 2009/NA06, 2009.

A. Rambaut, D. Posada, K.A. Crandall, and E.C. Holmes. The causes and consequences of HIV evolution. Nature Reviews. Genetics, 5:52–61, 2004.

B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution, 43:304–311, 1996.

C.R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 37:81–89, 1945.

H. Riihimaki, T. Mattsson, A. Zitting, G. Wickstrom, and K. Hanninen. Radiographically detectable degenerative changes of the lumbar spine among concrete reinforcement workers and house painters. Spine, 15:114–119, 1990.

S. Roch. Toward extracting all phylogenetic information from matrices of evolutionary distances. Science, 327:1376–1379, 2010.

F. Rodriguez, J.F. Oliver, A. Marin, and J.R. Medina. The general stochastic model of nucleotide substitutions. Journal of Theoretical Biology, 142:485–501, 1990.

F. Ronquist. Fast fitch-parsimony algorithms for large data sets. Cladistics, 13: 387–400, 1998.

F. Ronquist and J.P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19:1572–1574, 2003.

F. Ronquist, M. Teslenko, P. van der Mark, D.L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M.A. Suchard, and J.P. Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, 61:539–542, 2012.

N. Saitou and T. Imanishi. Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum likelihood, minimum evolution, and neighbor joining methods of phylogenetic tree construction in obtaining the correct tree. Molecular Biology and Evolution, 6:514–525, 1989.

N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4:406–425, 1987.

L.A. Salter and D.K. Pearl. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Systematic Biology, 50:7–17, 2001.

D. Sankoff. Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics, 28:35–42, 1975.

D.I. Scaduto, J.M. Brown, W.C. Haaland, D.J. Zwickl, D.M. Hillis, and M.L. Metzker. Source identification in two criminal cases using phylogenetic analysis of HIV-1

DNA sequences. Proceedings of the National Academy of Sciences, 107:21242–21247, 2010.

M.H. Schierup and J. Hein. Consequences of recombination on traditional phylogenetic analysis. Genetics, 156:879–891, 2000.

J.R. Schott. Matrix Analysis for Statistics. Wiley, New York, 1997.

A.H. Schultz. The Life of Primates. Weidenfeld and Nicholson, 1969.

L.I. Shlush, N. Chapal-Ilani, R. Adar, N. Pery, Y. Maruvka, A. Spiro, R. Shouval, J.M. Rowe, M. Tzukerman, D. Bercovich, S. Izraeli, G. Marcucci, C.D. Bloomfield, T. Zuckerman, K. Skorecki, and E. Shapiro. Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. Blood, 120:603–612, 2012.

G.R. Shorack. Statistics with Probability. Unpublished, 2006.

L.R. Shore. On osteo-arthritis in the dorsal intervertebral joints. British Journal of Surgery, 22:833–849, 1935.

P.H.A. Sneath. The application of computers to taxonomy. Journal of General Microbiology, 17:201–226, 1957.

R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 38:1409–1438, 1958.

D.A. Sorensen, S. Andersen, D. Gianola, and I. Kornsaard. Bayesian inference in threshold models using Gibbs sampling. Genetics Selection Evolution, 27:229–249, 1995.

D.A. Sorensen, D. Gianola, and I.R. Korsgaard. Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. Acta Agriculturae Scandinavica. Section A, Animal Science, 48:222–229, 1998.

A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22:2688–2690, 2006.

K.F. Stock, O Distl, and I Hoeschele. Influence of priors in Bayesian estimation of genetic parameters for multivariate threshold models using Gibbs sampling. Genetics Selection Evolution, 39:123–137, 2007.

J. Sullivan and D.L. Swofford. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution patterns are violated? Systematic Biology, 50:723–729, 2001.

J.E. Sulston, E. Schierenberg, J.G. White, and J.N. Thomson. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Developmental Biology, 64:64–119, 1983.

T.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82:528–550, 1987.

Y. Tateno, N. Takezaki, and M. Nei. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Molecular Biology and Evolution, 11:261–277, 1994.

S. Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences, 17:57–86, 1986.

J.K. Tinker, L.S. Hancox, and S. Clegg. FimW is a negative regulator affecting type 1 fimbrial expression in *Salmonella enterica* serovar typhimurium. Journal of Bacteriology, 183:435–442, 2001.

B. Vernon-Roberts and C.J. Pirie. Degenerative changes in the intervertebral discs of the lumbar spine and their sequelae. Rheumatology and Rehabilitation, 16:13–21, 1977.

T. Videman and M.C. Battie. The influence of occupation on lumbar degeneration. Spine, 24:1164–1168, 1999.

T. Videman, M. Nurminen, and J.D. Troup. Lumbar spinal pathology in cadaveric material in relation to history of back pain, occupation, and physical loading. Spine, 15:728–740, 1990.

D.R. Vieites, M. Min, and D.B. Wake. Rapid diversification and dispersal during periods of global warming by plethodontid salamanders. Proceedings of the National Academy of Sciences, 104:19903–19907, 2007.

P.J. Waddell and M.A. Steel. General time-reversible distances with unequal rates across sites: mixing gamma and inverse gaussian distributions with invariant sites. Molecular Phylogenetics and Evolution, 8:398–414, 1997.

D.B. Wake, M.H. Wake, and C.D. Specht. Homoplasy: from detecting pattern to determining process and mechanism of evolution. Science, 331:1032–1035, 2011.

M.S. Waterman. On the similarity of dendrograms. Journal of Theoretical Biology, 73:789–800, 1978.

S. Wright. Coefficients of inbreeding and relationship. American Naturalist, 56(645): 330–338, Jul - Aug 1922.

S. Wright. An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics, 19:506–536, 1934.

Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when

substitution rates differ over sites. Molecular Biology and Evolution, 10:1396–1401, 1993.

Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution, 39:306–314, 1994a.

Z. Yang. Estimating the pattern of nucleotide substitution. Journal of Molecular Evolution, 39:105–111, 1994b.

Z. Yang. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Journal of Molecular Evolution, 43:329–342, 1994c.

Z. Yang. Computational Molecular Evolution. Oxford University Press, 2006.

Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution, 24:1586–1591, 2007.

Z. Yang and S. Kumar. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Molecular Biology and Evolution, 13:650–659, 1996.

Z. Yang and R. Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution, 19:908–917, 2002.

Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular Biology and Evolution, 14:717–724, 1997.

Z. Yang and B. Rannala. Molecular phylogenetics: principles and practice. Nature Reviews Genetics, 13:303–314, 2012.

Z. Yang, R. Nielsen, N. Goldman, and A.K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics, 155:431–449, 2000a.

Z. Yang, W.J. Swanson, and V.D. Vacquier. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Molecular Biology and Evolution, 17:1446–1455, 2000b.