SCALABLE PHYLOGENETIC ANALYSIS AND FUNCTIONAL INTERPRETATION OF GENOMES WITH COMPLEX EVOLUTIONARY HISTORIES

By

Hussein El Abbass Hejase

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science and Engineering – Doctor of Philosophy

2017

ABSTRACT

SCALABLE PHYLOGENETIC ANALYSIS AND FUNCTIONAL INTERPRETATION OF GENOMES WITH COMPLEX EVOLUTIONARY HISTORIES

By

Hussein El Abbass Hejase

Phylogenomics involves the inference of a genome-scale phylogeny. A phylogeny is typically inferred using sequences from multiple loci across a set of genomes of multiple organisms by reconstructing gene trees and then reconciling them into a species phylogeny. Many studies have shown that evolutionary processes such as gene flow, incomplete lineage sorting, recombination, selection, gene duplication and loss have shaped our genomes and played a major role in the evolution of a diverse array of metazoans, including humans and ancient hominins, mice, bacteria, and butterflies. The aforementioned evolutionary processes are primary causes of gene tree discordance, which introduce different loci in a genome that exhibit local genealogical variation (i.e. gene trees differing from each other and the species phylogeny in terms of topology and/or branch length).

In this dissertation, we develop a method for fast and accurate inference of phylogenetic networks using large-scale sequence data. The advent of high-throughput sequencing technologies has brought about two main scalability challenges: (1) dataset size in terms of the number of taxa and (2) the evolutionary divergence of the taxa in a study. We explore the impact of both dimensions of scale on phylogenetic network inference and then introduce a new phylogenetic divide-and-conquer method which we call FastNet. We show using synthetic and empirical data spanning a range of evolutionary scenarios that FastNet outperforms the state-of-the-art in terms of accuracy and computational requirements.

Furthermore, we develop methods that use better and more accurate phylogenies to functionally interpret genomes. One way to study and understand the biological function of genomes is through association mapping, which pinpoints statistical associations between genotypic and phenotypic characters while modeling the relatedness between samples to avoid generating spurious inferences. Many methods have been proposed to perform association mapping while accounting for sample

relatedness. However, the state of the art predominantly utilizes the simplifying assumption that sample relatedness is effectively fixed across the genome. Recent studies have shown that sample relatedness can vary greatly across different loci within a genome where gene trees could differ from each other and the species phylogeny. Thus, there is an imminent need for methods to account for local genealogical variation in functional genomic analyses. We address this methodological gap by introducing two methods, Coal-Map and Coal-Miner, which account for sample relatedness locally within loci and globally across the entire genome. We show through simulated and empirical datasets that these newly introduced methods offer comparable or typically better statistical power and type I error control compared to the state-of-the-art.

This dissertation is dedicated to Ale and Layla. Thank you for your continued support.

ACKNOWLEDGEMENTS

Throughout my graduate academic studies, which involved research work with brilliant professors and high-end class experience, I have been exposed to plenty of personality build up experience and was fortunate to work in a healthy research and academic environment. This dissertation would not have been completed without the help and support from many people. I have been privileged to have Dr. Kevin Liu as my advisor. I joined Dr. Liu's lab at a time where I had almost lost interest in pursuing my graduate degree due to a variety of reasons beyond my control. Dr. Liu helped me in a variety of aspects along the way such as becoming a well-rounded researcher, refining my writing and presentation skills, attention to detail, professional advice, laboratory resources, and career aspirations. Dr. Liu gave me a valuable advice that I believe is an indispensable skill to become an expert in a specific field of study and that is to READ!

I also would like to thank my committee members Jin Chen, Shin-Han Shiu, and Yanni Sun for their contributions along the way. I am grateful to Dr. Sun who introduced me to research when I was an undergraduate junior at Michigan State University. Additionally, a big thanks to Dr. Shiu for his valuable insights along the way. Finally, I would like to thank Dr. Chen whom I had learned a lot from, especially in his network biology class, and had valuable research conversations with him.

A word of thanks to Christina Chan and Leslie Kuhn whom I had the privilege to work with them. A word of appreciation to our collaborators Gregory Bonito, Patrick Edger, and Natalie Vande Pol whom I enjoyed researching alongside them where our collaborative work resulted in numerous journal and conference publications.

I was fortunate to have been raised by hard-working parents Ale and Layla who provided me with the steppingstone to become the person I am today. They provided all the requirements in order to excel in this journey and encouraged me towards attaining my future goals.

I also have been privileged to have my siblings Jose, who has my back every time I need him, Charifa, who provides me with encouragement and energizes me every single day, and Ahmad,

who finds numerous ways to entertain and put a smile on my face. A special thanks to my dear wife Maya, who have filled my life with happiness and love from the moment I met her, for her prayer and encouragement along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

KEY TO ABBREVIATIONS

AM Association Mapping

AUROC Area Under Receiver Operating Characteristic

FPR False Positive Rate

GWA Genome-Wide Association

HMM Hidden Markov Model

ILS Incomplete Lineage Sorting

LMM Linear Mixed Model

MDC Minimize Deep Coalescences

MWEC Minimum-Weight Edge Cover

PCA Principal Component Analysis

RF Robinson-Foulds

ROC Receiver Operating Characteristic

TPR True Positive Rate

CHAPTER 1

INTRODUCTION AND CONTRIBUTIONS

A phylogeny represents the evolutionary history for a set of organisms. Phylogenies are used in a range of applications, which includes studying pathogens (Grenfell *et al.*, 2004), representing the relationships between species in the tree of life, studying speciation and extinction (Grenier & Weatherbee, 2001), studying the spread of antibiotic resistance between species (Ochman *et al.*, 2000; Thomas & Nielsen, 2005), identifying genes and non-coding RNAs in newly sequenced genomes (Kellis *et al.*, 2003), and reconstructing ancestral genomes (Ma, 2011). The aforementioned examples are a subset of many applications that have driven researchers to design methods for phylogeny reconstruction.

Most phylogenetic reconstruction methods assume that the underlying biological data follows a tree-like structure. However, it is known that in some areas (Consortium, 2012; Green *et al.*, 2010; Liu *et al.*, 2015) this assumption does not always hold due to evolutionary processes such as horizontal gene transfer in prokaryotes (Ochman *et al.*, 2000; Thomas & Nielsen, 2005) and hybridization in eukaryotes (Consortium, 2012; Green *et al.*, 2010; Liu *et al.*, 2015), which necessitates going beyond trees (Bapteste *et al.*, 2013). These evolutionary processes result in reticulate evolutionary histories and are best modeled using a phylogenetic network, which accounts for vertical and non-vertical evolutionary processes.

1.1 Contributions

This dissertation addresses two main issues. First, we develop accurate and fast phylogenetic network inference methods for large-scale biological sequence datasets. We propose to study and understand how species are related using whole genome sequencing data for many dozens of taxa, which is captured using a phylogenetic network. For any system of interest (i.e. humans), present-day populations arose through complex evolutionary histories that involved mutation, recombination, ancestral polymorphism, and gene flow. Using a phylogenetic network, we are able

to understand the evolutionary causes underpinning the biological system under study and thereby develop novel therapeutics and methods of disease detection. The advent of high-throughput sequencing technologies has brought about two main scalability challenges: (1) dataset size in terms of the number of taxa and (2) the evolutionary divergence of the taxa in a study. The impact of both dimensions of scale and the scalability limits of phylogenetic network inference methods are largely unknown. In chapter 3, we show that current state-of-the-art phylogenetic network inference methods lag well behind the scope of current phylogenomic studies, which introduces the critical need for new algorithmic development to address this methodological gap. The methodological gap remains: how can phylogenetic networks be accurately and efficiently inferred using genomic sequence data involving many dozens of taxa? In chapter 4, we address this gap by introducing a divide-and-conquer method which we call FastNet. Using synthetic and empirical data spanning a range of evolutionary scenarios, we demonstrate that FastNet outperforms state-of-the-art methods in terms of computational efficiency and topological accuracy. The second issue addressed in this dissertation is how to utilize more accurate phylogenies to improve the functional interpretation of genomes. One of the most widely used approaches for the functional interpretation of genomes is association mapping (AM), which examines the relationship between genotype and phenotype to recover statistical associations pointing to the underlying genetic causes of phenotypic traits. Previous studies (Price *et al.*, 2010a; Devlin & Roeder, 1999; Marchini *et al.*, 2004; Astle & Balding, 2009) have shown that an important consideration in AM is that relatedness between sampled individuals, or sample structure, can induce spurious associations between genotypic and phenotypic characters when not properly accounted for. Furthermore, population genetic theory and empirical studies (see Edwards for a review of relevant literature) have shown that sample relatedness can vary greatly across different loci within a genome. This phenomenon, referred to as local genealogical variation, is commonly encountered in many genomic datasets, which introduces the need for new AM methods to better account for local variation in sample relatedness within genomes. We address this issue and improve AM methods by utilizing more accurate phylogenies to model sample relatedness and local genealogical variation across genomes. In chapters 5 and 6, we introduce

two methods, Coal-Map and Coal-Miner, that utilize linear mixed models in the context of AM to study the relationship between genotype and phenotype. We apply these methods on simulations, incorporating a range of evolutionary scenarios, and empirical datasets, involving bacteria, mice, plants, and butterflies, and show their performance advantage against state-of-the-art AM methods.

1.2 Organization

This document is outlined as follows: In chapter 2, we describe background information for the research presented in this dissertation. In chapter 3, we present a performance study to explore the scalability limits of phylogenetic network inference methods. In chapter 4, we introduce FastNet, which is a scalable phylogenetic network inference method, that deals with large-scale datasets. In chapter 5, we introduce Coal-Map, which models local genealogical variation across genomes using a fixed effects model. In chapter 6, we introduce Coal-Miner, which adds more modeling contributions (i.e. multiple effects instead of a fixed effects model, detection of candidate loci) relative to Coal-Map. Finally, in chapter 7, we conclude by summarizing our presented work and highlighting future research directions.

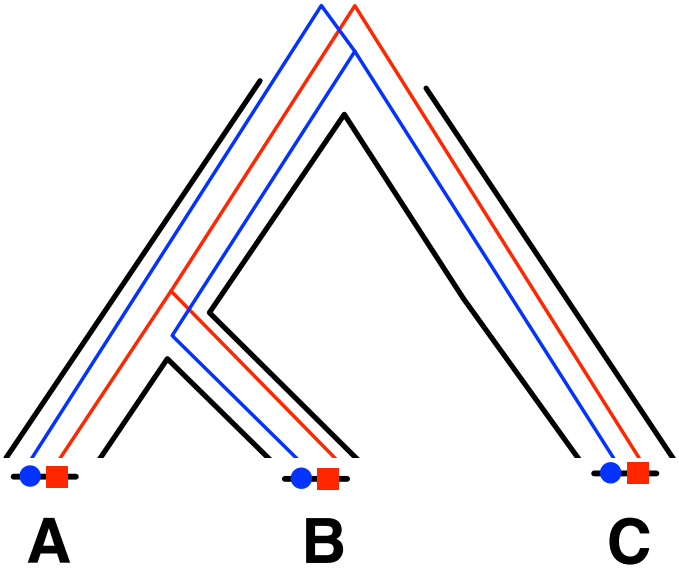
CHAPTER 2

BACKGROUND AND RELATED WORK

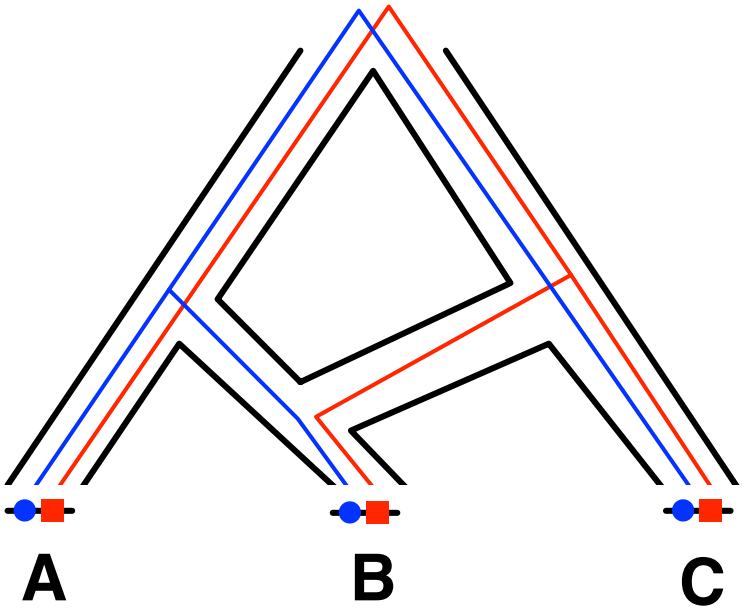
2.1 Phylogenomics

Phylogenomics aims to reconstruct a genome-scale phylogeny or the evolutionary history of organisms by analyzing their genomes (Delsuc *et al.*, 2005). Today phylogenies are often reconstructed using computational analysis of genomic sequence data. Many organisms across the Tree of Life has a phylogeny, or evolutionary history, which cannot be represented as a tree, where a branching event reflects strict bifurcating and/or multifurcating speciation/splitting and subsequent genetic isolation of the resulting species/populations. In these cases, the phylogeny takes the more general form of a directed acyclic graph known as a phylogenetic network (Maddison, 1997). Traditional approaches for phylogenetic inference (Bryant & Moulton, 2004; Edwards, 2009; Schliep, 2009) involved concatenating a set of sequence alignments from a set of genes into one multiple sequence alignment, and then use an inference method to infer a phylogeny. One issue that arises from such an approach is that it does not account for local genealogical heterogeneity. On the other hand, new approaches for phylogenetic inference (Yang & Warnow, 2011; Degnan & Rosenberg, 2009; Nakhleh, 2013; Yu *et al.*, 2014, 2013a) involve inferring a gene tree for each gene sequence alignment and then reconciling those discordant gene trees using a summary-based method into a species phylogeny. An advantage of using such an approach over concatenation methods is that it accounts for local genealogical heterogeneity. This summary-based approach is only one of many approaches that have been proposed to tackle this problem. Other approaches include the simultaneous inference of gene trees and species phylogenies (Bryant *et al.*, 2012). As we scan across a genome, the topology and/or branch lengths of gene trees change, and the species phylogeny inferred could be different than some of the observed gene trees. This is mainly due to different parts of the genome exhibiting different evolutionary histories. Some of the primary causes of this local genealogical heterogeneity are ILS/genetic drift, hybridization/gene flow, as well as other

evolutionary processes such as recombination, natural selection, and gene duplication/loss. ILS occurs when lineages from two genetically isolated populations coalesce at a time more ancient than their most recent common ancestral population, and is known to play a crucial role in the evolution of much of the Tree of Life (Edwards, 2009). Under neutral evolution, genetic drift - the outcome of purely stochastic inheritance over successive populations - can cause ILS; other factors contributing to the maintenance of ancestral polymorphisms and ILS include balancing selection. On the other hand, gene flow is a process by which genetic material is exchanged between different populations and/or species existing at the same point in time. In this dissertation, we focus on inference methods that account for ILS/genetic drift and hybridization/gene flow. Figure 2.1 displays an example of incongruent local genealogies evolving within a species phylogeny where this incongruence is due to ILS or hybridization. When we have both ILS and hybridization occurring simultaneously as depicted in Figure 2.2, the picture gets more complicated and introduces the need for inference methods that could distinguish the signals due to ILS or hybridization.



(a) ILS



(b) Hybridization

Figure 2.1: **Illustration of two gene trees growing inside a species phylogeny.** (a) The red gene tree agrees with the topology of the species tree while the blue gene tree disagrees with the species tree. When we run into ILS, we have a collection of gene trees that disagree with one another and with the species tree. (b) Hybridization is another factor that causes gene tree incongruence. In this case, there is gene flow occurring from one species to another. By tracing the lineages of the hybrid species B, they either coalesce with the left or right parent of species A and C, respectively.

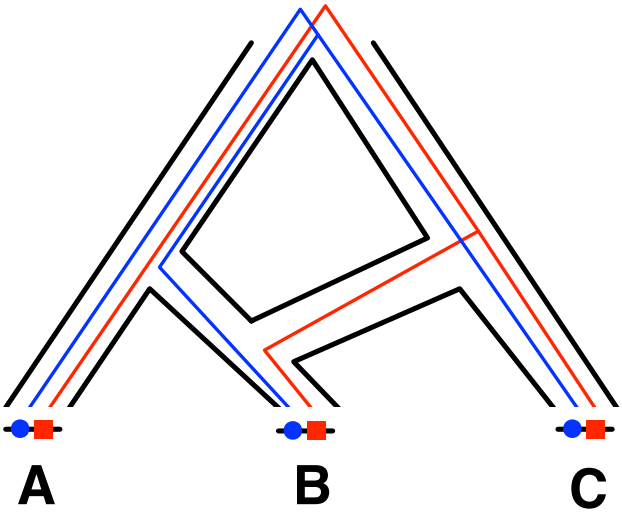


Figure 2.2: **Illustration of both ILS and hybridization acting simultaneously.** The topology of the blue gene tree agrees with the red gene tree. In this illustration, hybridization is happening but ILS is conflicting with the signal of hybridization.

2.1.1 Phylogenetic study

A traditional approach for phylogeny reconstruction requires four design choices: data, evolutionary model, computational method, and support measure. One of the requirements in a phylogenetic study is an evolutionary model. We'll focus here on sequence evolution models dealing with nucleotide substitution. Many models of DNA sequence evolution have been introduced (Jukes & Cantor, 1969; Felsenstein, 1981; Kimura, 1980; Tavaré, 1986). These models differ based on the number of parameters that describe the rate of a nucleotide changing to another. Most of the methods assume that the sites have evolved down the edges of a tree under a Markov process. Each site evolving down an edge in a tree is annotated using a substitution matrix. One of the simplest models of DNA sequence evolution is the Jukes-Cantor mutation model (Jukes & Cantor, 1969), which makes three assumptions. First, the evolution of each site is identically and independently distributed. Second, the initial state at the root is randomly assigned. Third, a site changes a state with equal probability to any of the remaining three states on an edge. The Jukes-Cantor model uses a single parameter to describe the rate of a nucleotide changing to another and has the following

transition matrix:

$$\begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix} \tag{2.1}$$

The base frequencies π_A, π_C, π_G, and π_T are $\frac{1}{4}$. One could have two possible changes where either a nucleotide changes to another or it does not change. The probabilities of these transitions are:

$$\begin{aligned} P(t)_{ii} &= \frac{1}{4} + \frac{3}{4}e^{-\mu t} \\ P(t)_{ij} &= \frac{1}{4} - \frac{1}{4}e^{-\mu t} \end{aligned} \tag{2.2}$$

The Generalized Time-Reversible model (GTR) (Tavaré, 1986) is a more complex model that is time-reversible with six transition rate parameters to describe the rate of a nucleotide changing to another $r_{ij} = a$, b, c, d, e, and f and four equilibrium base frequency parameters π_A, π_C, π_G, and π_T, which describe the frequency of each base at each site. r_{ij} and the base frequencies form the rates in the transition matrix:

$$\begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix} \tag{2.3}$$

Another design choice in a phylogenetic study is the computational method used for phylogeny inference. These computational methods fall into two categories: distance-based or sequence-based. A distance-based approach transforms the multiple sequence alignment into pairwise distances (similarities), and then use the distance matrix to construct the tree. An example of methods that use this approach are UPGMA (Michener & Sokal, 1957) and neighbor-joining (Saitou & Nei, 1987). For sequence-based methods, one needs to define an optimization criterion such as maximum parsimony (Felsenstein, 1978; Sober, 1983) or maximum likelihood (Felsenstein, 1981) and then infer a tree that optimizes the criterion. The principle of likelihood requires to find $P(S|H)$ where S is the set of sequences that are compared and H is the phylogenetic tree and

substitution model. Hence, we need to find the likelihood of obtaining the observed sequences given a tree based on a substitution model.

2.1.2 A maximum likelihood approach

The general formula for reconstructing evolutionary histories from sequence data or gene trees was introduced by Maddison (1997):

$$L(\Psi|S) = \prod_{s \in S} \Big[\sum_{T} [P(s|T) \cdot P(T|\Psi)] \Big] \tag{2.4}$$

Given a set of sequence alignments from multiple loci S and we are interested in the species tree Ψ and its maximum likelihood, then one wants to maximize the above entity $L(\Psi|S)$. $L(\Psi|S)$ is the product over all loci simply because of the assumption that the loci are independent. If we observe the sequence alignment of every locus, one could sum over all possible gene tree topologies. The two terms $P(s|T)$ and $P(T|\Psi)$ represent the probability of sequences given a tree and the probability of a gene tree given a species tree, respectively. To account for ILS or ILS and hybridization, then $P(T|\Psi)$ has to be derived. An alternative scenario is to reconstruct the gene trees first, which factors out completely this issue of sequences and integration over gene trees, and then focus completely on gene trees and its probability (see equation 2.5).

$$L(\Psi|G) = \prod_{g \in G} P(g|\Psi) \tag{2.5}$$

2.1.3 Coalescent theory

Population genetics involves the study of genetic variation within a population whereas phylogenetics uses this genetic variation between a number of taxa (i.e. species or populations) to infer the evolutionary relationships between them. In the pre-genomic era, each taxon was represented using a single sequence while the availability of data constrained the number of genes into a single gene; the end-goal was to infer the evolutionary history for that gene. In the post-genomic era, given the advances in high-throughput sequencing technologies, we can do much more now such as

sample multiple number of individuals within each taxon and sequence more genes. This requires modeling the relationships within a population using population genetics techniques as well as modeling what happens between populations using phylogenetics.

The most simple population genetics model is the Wright-Fisher model. Given N diploid individuals where each individual has two copies of each gene (i.e. a population of $2N$ gene copies), the Wright-Fisher model assumes non-overlapping generations where at each generation each gene is randomly assigned with-replacement a gene from the previous generation. Therefore, the probability that two gene copies come from the same gene in the previous generation is $\frac{1}{2N}$. Therefore, in every generation, we have a chance of $\frac{1}{2N}$ to coalesce. By tracing the sampled lineages backwards through generations, they follow a geometric distribution where (1) the number of generations since two genes first shared a common ancestor is $\frac{1}{2N}$ and (2) the number of generations since at least two genes in a sample of k shared a common ancestor is $\frac{k(k-1)}{4N}$. The disadvantage of using the Wright-Fisher model is that it ignores many important aspects such as mutation, recombination, selection, population structure, and many other factors.

The coalescent model provides a framework for studying patterns of genetic variation, which involves the complex interplay of different evolutionary processes that shaped the genomes. The coalescent model traces gene lineages moving backwards in time and coalescing to a lineage, known as the most recent common ancestor (MRCA). The basic coalescent model captures genetic drift and assumes that there is no recombination, natural selection, gene flow, ILS, and population structure. However, extensions to the coalescent model have been introduced (Hein *et al.*, 2004) to include any of the aforementioned evolutionary or demographic models. The structure of the tree generated by the coalescent model is determined by coalescent times t_i and how lineages are selected to merge at each coalescent event. In the basic coalescent model, lineages are selected to merge randomly, which is the result of random selection of parents. The probability that two alleles chosen at random coalesced t generations ago is the following:

$$P(T = t) = (1 - \frac{1}{2N_e})^{t-1}(\frac{1}{2N_e}) \tag{2.6}$$

where t is the time the coalescent event occurs and N_e is the effective population size.

So far, we've considered the coalescent process within a single population. A phylogenetic tree consists of many populations followed throughout evolutionary time. The goal is to apply the coalescent model across the phylogeny. The basic assumption is that events that occur in one population are independent of what happens in other populations within the phylogeny. More specifically, given the number of lineages entering and leaving a population, coalescent events within populations are independent of one another. It is also important to recall an assumption we "inherit" from our population genetics model: all pairs of lineages are equally likely to coalesce within a population. When discussing gene tree distributions, there are two cases of interest: the gene tree topology distribution and the joint distribution of topologies and branch lengths.

2.1.4 P(g|Ψ) under the coalescent

Degnan & Salter (2005) gave the mass probability function of a gene tree topology g for a given species tree with topology Ψ and a vector of branch length λ:

$$P_{\Psi,\lambda}(g) = \sum_{h \in H_\Psi(g)} \frac{w(h)}{d(h)} \prod_{b=1}^{n-2} \frac{w_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}(\lambda_b) \tag{2.7}$$

where $H_\Psi(g)$ denotes the set of all coalescent histories that explain a gene tree g appearing inside a species tree with topology Ψ and a vector of branch length λ, $u_b(h)$ denotes under coalescent history h the number of incoming lineages into branch λ_b, $v_b(h)$ denotes the number of exiting lineages from branch λ_b, $w_b(h)$ denotes the number of possible ways the coalescent events could have occurred consistently with the gene tree, $d_b(h)$ denotes the number of sequences of coalescences that give the number of coalescent events specified by h, and $p_{u_b(h)v_b(h)}(\lambda_b)$ denotes the probability of u_b lineages entering the branch and v_b lineages exiting the branch over a branch of length λ_b.

2.1.5 The multi-species coalescent

In the most simplest case (see Figure 2.3), the above formula 2.7 gets reduced to these three values:

$$
\begin{aligned}
P[((HC)G)] &= 1 - \frac{2}{3}e^{-(T_2-T_1)/N} \\
P[((HG)C)] &= \frac{1}{3}e^{-(T_2-T_1)/N} \\
P[((CG)H)] &= \frac{1}{3}e^{-(T_2-T_1)/N}
\end{aligned}
\tag{2.8}
$$

where T_2 - T_1 is the length of the branch that separates the most recent common ancestors in terms of the number of generations and N is the population size or the width of the branch. In phylogenetics, researchers don't focus on ILS because if the sampling is done in such a way that the branch T_2 - T_1 is long and if enough genes are sampled, then the gene with the highest frequency is indicative of the species tree. One caveat here is that this is only true for the three species case and under certain circumstances the gene tree with the highest probability is not necessarily the species tree when the species phylogeny goes beyond three species (Degnan & Rosenberg, 2006).

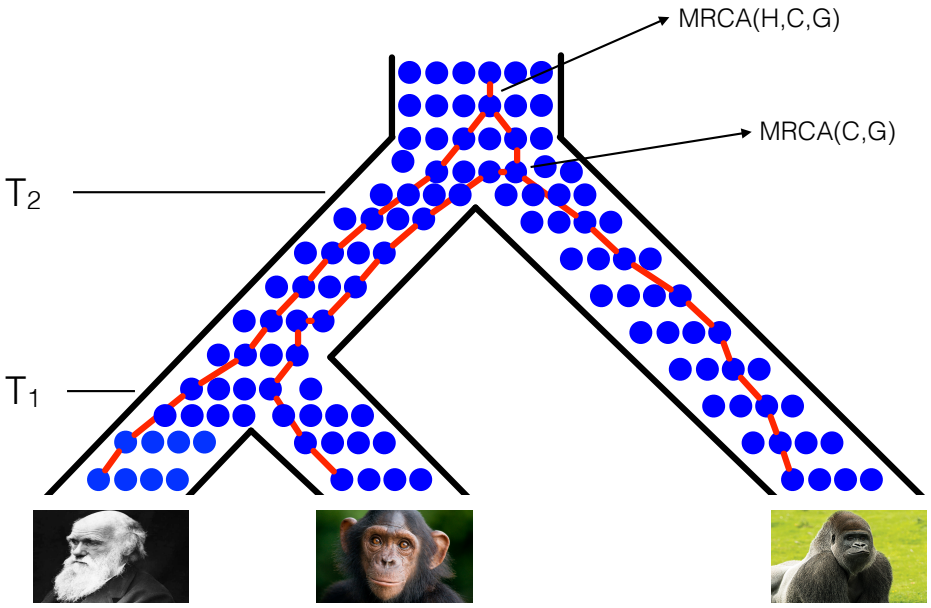


Figure 2.3: **A three species phylogeny containing human (H), chimpanzee (C), and gorilla (G) along with one red embedded gene tree.** There are three possible gene tree topologies that could appear inside this species tree: ((HC)G), ((HG)C), and ((CG)H).

2.1.6 Phylogenetic networks

Traditionally, it has been assumed that the relationship at the interspecific (across species) level takes the form of a phylogenetic tree. In a tree-like structure, the lineages are independent of one another and are arranged in a strictly vertical descent fashion. Although a phylogenetic tree is appropriate for many groups of taxa, it is inadequate for others. For example, evolutionary processes such as horizontal gene transfer, which is prevalent in bacteria (Hao & Golding, 2004; Kurland *et al.*, 2003), and hybrid speciation, which is common in many groups of organisms (Ellstrand *et al.*, 1996; Baack & Rieseberg, 2007; Noor & Feder, 2006) such as plants, fish, and frogs cannot be modeled using phylogenetic trees. These events, which are also known as reticulate events, introduce edges that connect nodes from different branches of a tree (Moret *et al.*, 2004). This creates a structure that takes the general form of a directed acyclic graph known as a phylogenetic network (Maddison, 1997).

Here, we consider the more general problem of inferring rooted species phylogenies that are directed phylogenetic networks. A directed phylogenetic tree is a special case of a directed phylogenetic network which contains no reticulation nodes (and edges). An unrooted topology can be obtained from a directed tree by ignoring edge directionality. A phylogenetic network is defined as follow:

Let N be a phylogenetic network. N is an ordered pair (G, f) such that $G = (V, E)$ is a directed acyclic graph with $V = r \cup V_N \cup V_T \cup V_L$ where:

- r is the root of the network N: $indegree(r) = 0$
- V_N is the set of reticulation nodes in N: $\forall\ v$ in V_N, $indegree(v) = 2$ and $outdegree(v) = 1$
- V_T is the set of tree nodes in N: $\forall\ v$ in V_T, $indegree(v) = 1$ and $outdegree(v) \geq 2$
- V_L is the set of leaves in N: $\forall\ v$ in V_L, $indegree(v) = 1$ and $outdegree(v) = 0$

f is the leaf-labeling function which maps every leaf node in V_L to N. Figure 2.4 shows an example of a phylogenetic network.

An important observation in a phylogenetic network is that it induces a set of trees, which provide a way to characterize a network topology based upon the set of encoded tree topologies. A primary complication is that different trees in a genome commonly exhibit local genealogical incongruence (i.e. gene trees can differ from each other and the species phylogeny in terms of topology and/or branch length), which is the result of complex evolutionary processes such as ILS, hybridization, and horizontal gene transfer. In Figure 2.4, two trees T_1 and T_2 are contained within the phylogenetic network N. These trees describe the relatedness between sampled individuals and could be captured and understood using the coalescent model.

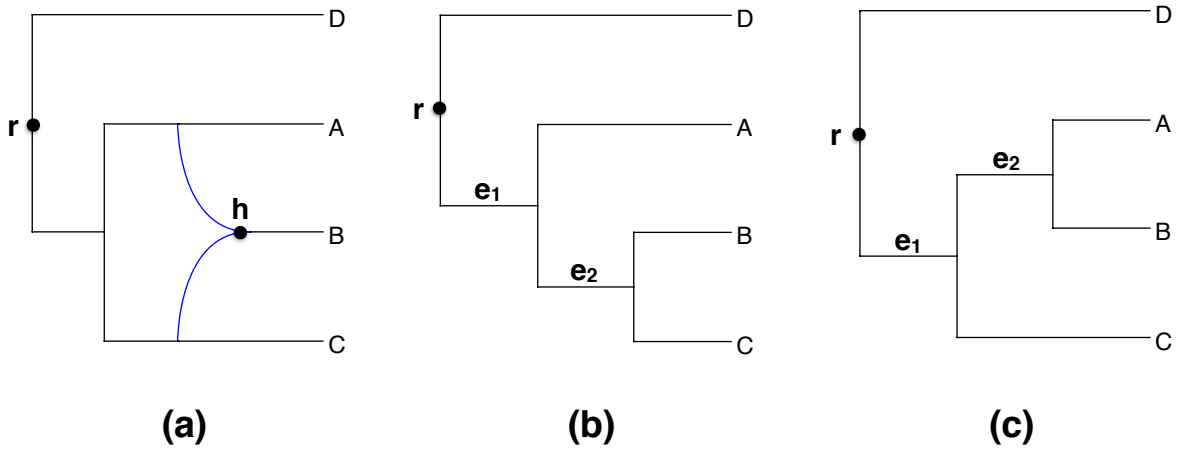


Figure 2.4: **A phylogenetic network N and its corresponding two induced gene trees (T_1 and T_2).** The network shown in (a) and the trees shown in (b) and (c) are rooted at node r. h is the reticulation node. The leaf labels of N are $\{A, B, C, D\}$.

2.1.7 The multi-species network coalescent

$$
\begin{aligned}
L(\Psi, \Gamma|G) &= \prod_{i=1}^{m} p(G_i|\Psi, \Gamma) \\
p(G_j|\Psi, \Gamma) &= \sum_{h \in H_\Psi(G_j)} p(h|\Psi, \Gamma) \\
p(h|\Psi, \Gamma) &= \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} \Gamma[b, j]^{u_b(h)} p_{u_b(h) v_b(h)}(\lambda_b)
\end{aligned}
\tag{2.9}
$$

Yu *et al.* (2014) derived the density function when the gene trees are given with their topology only. Suppose that for every locus, a gene tree G_i has been inferred, so that the input is $G =$

$G_1, G_2, ..., G_m$. The inference problem (under maximum likelihood) consists of finding the pair (Ψ^*, Γ^*) that maximizes the likelihood function $L(\Psi, \Gamma|G)$. Given a phylogenetic network Ψ and a gene tree G_i, an element that is central to computing gene tree distributions is the set of coalescent histories denoted by $H_\Psi(G_j)$. The aforementioned probabilistic approaches have been noted to have high computational requirements, and model likelihood calculations were found to be a major performance bottleneck (Hejase & Liu, 2016b; Yu *et al.*, 2014). For this reason, pseudo-likelihood approximations to full model likelihood calculations have been proposed. Yu *et al.* (2014) introduced the pseudo-likelihood of phylogenetic network Ψ and inheritance probabilities Γ given a set of gene trees G:

$$L(\Psi, \Gamma|G) = \prod_{\{X,Y,Z\}\subseteq\mathbb{X}} f(\rho(XY|Z, G), \rho(XZ|Y, G), \rho(YZ|X, G)|\Psi, \Gamma) \tag{2.10}$$

where $\mathbb{X}$ is the set of taxa, $XY|Z$ $XZ|Y$ and $YZ|X$ are binary triples with $X, Y, Z \in \mathbb{X}$, G is the set of gene trees, and ρ is the number of times a binary triple is induced by G. A maximum pseudo-likelihood approach seeks Ψ^* and Γ^* that maximize equation 2.10.

2.1.8 Simulation study

In a simulation study, we sample a set of gene trees from a model phylogeny (Hudson, 2002). Using a model of DNA sequence evolution (Jukes & Cantor, 1969; Felsenstein, 1981; Kimura, 1980), a set of sequences are evolved down the edges of each gene tree. This generates a sequence alignment for each gene tree.

The summary-based inference procedure of species phylogenies involves two steps. First, the local gene trees are inferred for each sequence alignment (Price *et al.*, 2009, 2010b). The second step involves the reconciliation of local gene trees into a species phylogeny using a computational multi-locus method (Yu *et al.*, 2014, 2013a). The inferred species phylogeny is then compared to the true species phylogeny (i.e. ground truth) to evaluate how accurate is the method under study.

The pipeline for summary-based inference of species networks, which is the focus of this dissertation, is defined as follow:

- **Input:** A set of gene trees G_i from a set of sample organisms n, and a criterion ω.
- **Output:** A phylogenetic network N, which models G_i, and is optimal under ω.

We evaluate the performance of phylogenetic network inference methods using two criteria. The first evaluation criterion involves computing the accuracy of the method. Accuracy is computed by comparing the inferred with the model phylogeny using distance-based measures such as Minimum-Weight Edge Cover (MWEC) or tripartition distance (Nakhleh *et al.*, 2003). The MWEC measure compares the similarity between the sets of trees induced by the inferred and model networks using RF distance (Robinson & Foulds, 1981). The RF distance is the sum of the number of false positive bipartitions (bipartitions found in inferred species network but not in true species network) and false negative bipartitions (bipartitions found in true species network but not in inferred species network). A bipartition is a split of a set of taxa where each edge of an unrooted phylogenetic tree represents one split. The tripartition distance counts the proportion of tripartitions that are not shared between the inferred and model networks. Assume we have a phylogenetic network N with L as a set of leafs. For a node u in N, we refer to an ancestor v of u as a strict ancestor if all the paths from root of N to u contain v. Otherwise we refer to v as a non-strict ancestor of u. The tripartition of node u is defined as $(A(u), B(u), C(u))$ where $A(u)$ is $s \in L$ for u is a strict ancestor of s, $B(u)$ is $s \in L$ for u is not a strict ancestor of s, and $C(u)$ is $s \in L$ for u is not an ancestor of s. The second evaluation criterion is the computational requirements of the method, which is measured in terms of running CPU time and memory usage. We note here that other evaluation criteria exist such as the inferred branch lengths accuracy and the estimated model parameters (i.e. substitution rates, recombination rates).

2.2 Association Mapping

For any species of interest (i.e. humans), present-day populations arose through complex evolutionary histories that involved mutation, recombination, ancestral polymorphism, natural selection, and gene flow. Using statistical models, researchers can begin to understand the evolutionary causes underpinning the biological system under study and potentially has translational value for

developing novel therapeutics and methods of disease detection. There are many scenarios where complex evolution occurs. For example, horizontal gene transfer involves the transfer of genetic material across species (i.e. Neanderthals and early modern humans (Green *et al.*, 2010)). Horizontal gene transfer is well documented in bacteria and has been shown to have an extensive role in the spread of antibiotic resistance between species (Thomas & Nielsen, 2005; Ochman *et al.*, 2000). Another example involves the evolution of pathogenicity in yeasts belonging to *Candida*, the most common fungal pathogens in humans, which is the result of genetic and evolutionary processes such as gene duplication and horizontal gene transfer (Moran *et al.*, 2011). Additionally, adaptive gene flow between species of mice is known to have a role in resistance and susceptibility to warfarin (Song *et al.*, 2011; Rost *et al.*, 2004), a widely used anticoagulant rodenticide. Some of the introgressed regions that contribute to warfarin resistance have orthologs in humans with cardiovascular functions. These orthologs have provided great assistance in the personalization of warfarin therapy, which involves a drug used in the prevention of blood clots (Song *et al.*, 2011). Finally, another application of complex evolution involves understanding the human origins. For example, a previous study (Huerta-Sánchez *et al.*, 2014) has shown that the adaptation of Tibetans to high-altitude is the result of introgression of DNA segments from Denisovan individuals into humans. The aforementioned are four disparate examples of complex evolution of genomes and physical traits. The fundamental commonality between these examples is that they arose through a process that could be captured and described using a single model unifying phylogenetic and population genetic stochastic processes. One fundamental question from the aforementioned examples that needs to be addressed is the biological function of the genetic variation. A widely used approach for understanding the biological function of genetic variation are GWA studies. GWA studies are approaches that recover statistical associations pointing to underlying genetic causes of target phenotypic variation by examining the relationship between genotype and phenotype.

2.2.1 Association mapping success stories

One application of AM studies is to identify the underlying genetic variants contributing to a disease, which provides an aid in developing prevention techniques and therapeutic strategies (Wang *et al.*, 2011; Bush & Moore, 2012). One example of an early success of GWA studies is the identification of Complement Factor H and its association with age-related macular degeneration (AMD) (Edwards *et al.*, 2005; Haines *et al.*, 2005). So far, AM methods have been successful in identifying genetic variations related to different diseases such as heart disorders (McPherson *et al.*, 2007), obesity (Scuteri *et al.*, 2007), risk of type 2 diabetes (Zeggini *et al.*, 2007), prostate cancer (Yeager *et al.*, 2007), and Crohn's disease (Barrett *et al.*, 2008). More recently, AM studies have begun to examine more divergent populations (Kang *et al.*, 2010; Porter *et al.*, 2017; Hejase *et al.*, 2017b) with complex evolutionary origins due to their non-tree-like evolutionary histories.

2.2.2 Sample relatedness

In AM studies, the relatedness between sampled individuals introduces a complex sample structure that confounds the association analysis and generates spurious results if not accounted for (Price *et al.*, 2010a; Devlin & Roeder, 1999). This sample relatedness can be due to more distant relationships from population subdivision (Marchini *et al.*, 2004), as well as less distant relationships such as family relationships and cryptic relatedness (Astle & Balding, 2009). Additionally, the sample relatedness is the result of population-level processes such as genetic admixture, ILS, hybridization, sequence mutation, gene duplication and loss, and recombination. These processes typically introduce variation among local genealogies, which can also differ from the global sample structure measured across all genomic loci. Figure 2.5 illustrates the effect of sample relatedness and local genealogical variation at the genomic sequence level.

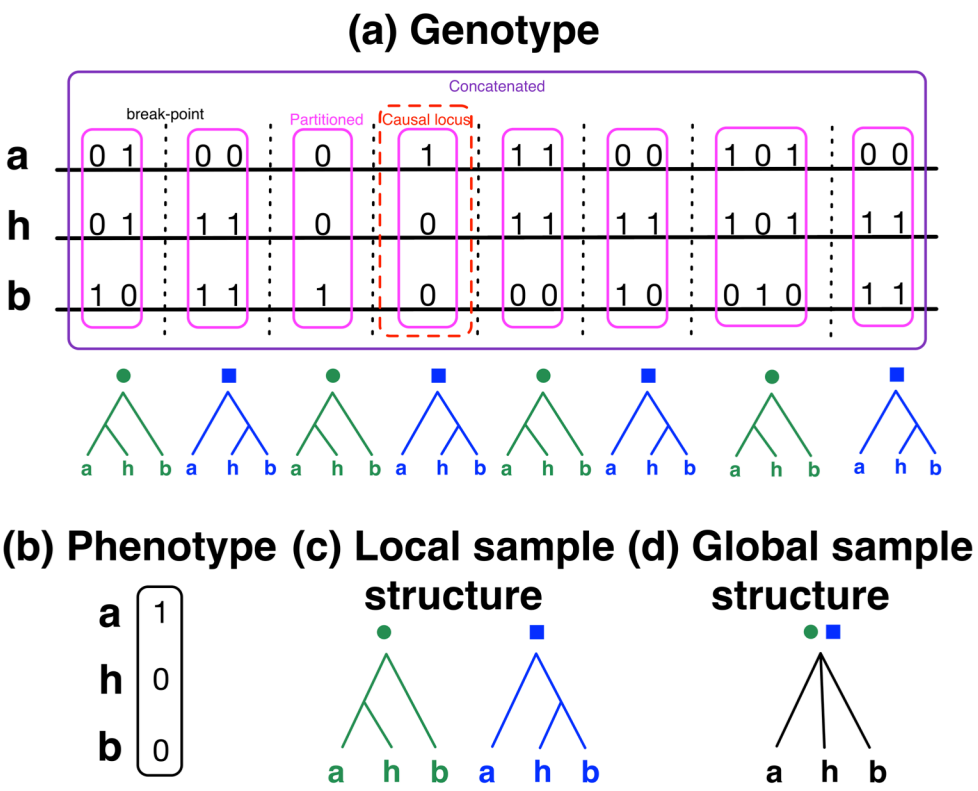


Figure 2.5: **A sequence level view of local genealogical variation.** (a) An illustration of three haploid genomes (a, h, and b) sampled from three populations using the model phylogeny in Figure 5.1. The sequence data is simulated under an infinite sites model with ancestral and derived alleles represented as 0 and 1, respectively. Each genomic locus has an evolutionary history represented using a gene tree (green or blue). The magenta box (local partition) represents a set of genomic sites that share the same evolutionary history and are separated from other local partitions using breakpoints. (b) A binary vector encoding the trait for each sample. AM methods scan every site across a multiple sequence alignment and evaluate its statistical association with the phenotype to identify the underlying genetic causes (i.e. causal locus) behind variation in the corresponding phenotype of interest. (c) Each genomic locus has an evolutionary history or a local sample structure represented as one of two possible gene trees: green or blue. (d) Global sample structure, measured across all loci, is represented using a star tree and differs in terms of topology from the local sample structure.

2.2.3 Methods

AM studies should account for and model the evolutionary relatedness (i.e. sample structure) between samples to avoid generating spurious and incorrect results (Price *et al.*, 2006). A range of methods, which differ in terms of their complexity, have been proposed to account for sample

relatedness. Non-parametric approaches such as genomic control (Devlin & Roeder, 1999) are used for controlling the inflation of test statistics. Genomic control computes an inflation factor based on the degree of sample relatedness, which is then used to correct for association statistics. Other approaches include highly parameterized models that use fixed effects or a mixture of fixed and random effects. EIGENSTRAT (Price *et al.*, 2006) utilizes a fixed effects model and uses PCA to infer population structure in genetic data. From an n by m genotypic matrix X where n is the number of SNPs and m is the number of individuals, an m by m covariance matrix ϕ is computed. The top k principal components are defined as the top k eigenvectors of ϕ (i.e. k eigenvectors of the k largest eigenvalues). Using the top k principal components as covariates, EIGENSTRAT corrects for population structure using the following:

$$X_{ij,adjusted} = X_{ij} - \alpha_i a_j \tag{2.11}$$

where i =1 to n, j =1 to m, α_i is the regression coefficient, and a_j is the axis of variation. After genetic and phenotypic adjustment based on the top principal components using equation (2.11), EIGENSTRAT applies a χ^2 association analysis between each genetic site and the phenotype.

Other AM approaches include mixed models such as EMMA (Kang *et al.*, 2008), EMMAX (Kang *et al.*, 2010), and GEMMA (Zhou & Stephens, 2012), which utilize a combination of fixed and random effects. These approaches typically infer a kinship matrix that represents the marker similarity or additive relationship between samples to describe their relatedness. Mixed models represent the phenotype Y as a function of fixed ($X\beta$) and random ($u + \epsilon$) effects:

$$\begin{aligned} Y &= X\beta + u + \epsilon \\ Var(u) &= {\sigma_g}^2 K \end{aligned} \tag{2.12}$$

where X represents the genotype of the candidate marker and additional covariates (fixed effects), β is the coefficients of fixed effects, u represents the heritable component of random variation and ϵ, which follows a normal distribution, represents the non-heritable component of random variation. In order to account for confounding factors such as population structure, mixed models represent the variation of the random effect u according to a distribution, which follows

a kinship matrix K, where K represents the pairwise genotypic similarity between individuals. Finally, other methods such as STRUCTURE (Pritchard *et al.*, 2000) and ADMIXTURE (Alexander *et al.*, 2009) utilize subpopulation clustering to infer sample structure.

The common assumption between all the aforementioned methods is that they either explicitly or implicitly assume that the evolutionary history of all genomic loci are effectively fixed. This assumption does not always hold due to different loci in a genome exhibiting local genealogical variation where gene trees differ from one another and from the species phylogeny. Local genealogical variation is due to the complex interplay of different evolutionary processes such as genetic admixture, ILS, hybridization, sequence mutation, gene duplication and loss, and recombination. Current methods only account for global sample relatedness, which is measured across all loci. There is an imminent need for methods that model local and global sample relatedness to capture local genealogical variation that shaped our genomes.

CHAPTER 3

A SCALABILITY STUDY OF PHYLOGENETIC NETWORK INFERENCE METHODS USING MULTI-LOCUS SEQUENCE DATA

Interspecific gene flow involves the transfer of genetic material between species and has been shown to have played a major role in the evolution of multiple species such as humans (Green *et al.*, 2010; Reich *et al.*, 2010), mice (Liu *et al.*, 2015), and butterflies (Consortium, 2012). Due to gene flow and other evolutionary processes, such as horizontal gene transfer in prokaryotes and hybridization in eukaryotes, an important question that arises is to what extent can this relationship not be represented using a tree, but instead a different structure that takes the general form of a directed acyclic graph also known as a phylogenetic network. Phylogenetic networks are inferred using multi-locus sequence data. For example, a concatenated analysis is an approach that uses multi-locus sequence data to infer a single phylogeny. The most common concatenated approach makes use of traditional phylogenetic inference methods and only accounts for sequence mutation and gene flow, where all local genealogical discordance is ascribed to gene flow, which poses a limitation because such methods do not handle the complex interplay of different evolutionary processes such as recombination, incomplete lineage sorting (ILS), and gene duplication and loss, which cause local genealogical variation. In contrast to a concatenated approach, other phylogenetic inference methods have been introduced to handle a combination of the aforementioned evolutionary processes. Among the phylogenomic approaches are methods that perform inference that accounts for sequence mutation, ILS, and gene flow (Durand *et al.*, 2011; Yu *et al.*, 2014, 2013a; Yu & Nakhleh, 2015). In practice, the sequence inputs to these methods are filtered to mitigate the impact of other evolutionary processes on downstream phylogenetic inference (Gatesy & Springer, 2013; Edwards, 2009).

These inference methods account for a broad set of evolutionary processes, in particular ILS, which is known to play a crucial role in the evolution of much of the Tree of Life (Edwards, 2009). The inference methods of Yu *et al.*, Yu *et al.*, and Yu & Nakhleh which are implemented in

PhyloNet (Than *et al.*, 2008) are among the more widely used methods in the multi-locus inference category. These methods perform heuristic search for optimization problems that are suspected to be NP-hard (Yu *et al.*, 2014, 2013a; Yu & Nakhleh, 2015). Thanks to rapid advances in genome sequencing and related biotechnologies (Metzker, 2010), large-scale phylogenetic studies involving many dozens of genomes or more are now common (see Yang & Rannala for a survey). These developments pose two primary scalability challenges: (1) the number of taxa in a study, and (2) sequence divergence, which reflects the evolutionary divergence of the taxa in a study.

We note that, for the special case of phylogenetic tree inference from phylogenomic data, recent studies have examined these scalability challenges (Mirarab *et al.*, 2014c,b,a), including evolutionary scenarios involving gene flow (Davidson *et al.*, 2015; Leaché *et al.*, 2014), and proposed new methods for large-scale analysis (Mirarab *et al.*, 2014c; Mirarab & Warnow, 2015; Chifman & Kubatko, 2014). In contrast, for phylogenetic network inference methods, the limits of scalability on inputs with more than a few dozen taxa as well as performance at these limits have yet to be established. Much is unknown about the scalability of these methods: what are their computational requirements, and what is their accuracy on large-scale inputs with dozens of taxa or more? The primary open question regarding these methods concerns their accuracy, especially concerning the workaround of analyzing subsets of the full set of taxa. Does subset-based inference result in less accuracy than a single combined analysis of the full set of input sequences, and, if so, how much is accuracy affected? To resolve these open questions, we conducted a study (Hejase & Liu, 2016b) that evaluated state-of-the-art phylogenetic network inference methods on both simulated and empirical datasets. To our knowledge, our study is the first to address these open questions and provide scalability guidelines for state-of-the-art phylogenetic network inference methods.

We chose representative methods from the category of multi-locus methods. We evaluated four optimization methods implemented in PhyloNet (Than *et al.*, 2008): (1) a probabilistic method (Yu *et al.*, 2014) that maximizes the likelihood of a coalescent-based model given the local gene trees including branch lengths, (2) a related probabilistic method (Yu *et al.*, 2014, 2012) that

performs maximum likelihood estimation using local gene tree topologies without branch lengths, (3) a probabilistic method (Yu & Nakhleh, 2015) that uses pseudo-likelihood approximations to full model likelihood calculations, and (4) a parsimony-based inference method (Yu *et al.*, 2013a) that seeks a phylogeny which minimizes the number of deep coalescences necessary to explain the observed local genealogies (as proposed by Maddison). We refer to the aforementioned methods as MLE-length, MLE, MPL, and MP, respectively. We further evaluated the performance of SNaQ (Solís-Lemus & Ané, 2016), which combines the use of pseudo-likelihoods under a coalescent-based model with quartet-based concordance analysis. All five methods infer local gene trees as part of a methodological pipeline; we used FastTree (Price *et al.*, 2009, 2010b), a method for inferring maximum likelihood phylogenetic trees under a substitution-only model, for this purpose. In the simulation study, we focus on the simpler case of search among phylogenetic networks with the correct number of reticulations, which is one in all model conditions. The more general case of search among network hypotheses with differing reticulations necessitates the use of model selection techniques to balance model fit versus complexity, and is suspected to be more difficult for this reason (Yu *et al.*, 2014, 2012). We compared these methods based on three performance measures: (1) computational time, (2) memory usage, and (3) topological accuracy. For the latter measure, we compared the inferred phylogeny against the model phylogeny on simulated datasets; we used the tripartition distance (Nakhleh *et al.*, 2003) to compare all edges in the two phylogenies. The empirical datasets consisted of positive controls based on past studies of natural mouse populations.

3.1 Simulation study

3.1.1 Generation of random model trees using r8s

Random model trees were generated using r8s version 1.7 (Sanderson, 2003). The following script was used to simulate random birth-death model trees for 5, 6, 7, 9, 10, 15, 20, 25, and 30 taxa:

```
begin rates;
```

```
simulate diversemodel=bdback seed=<integer random seed> nreps=20
ntaxa=<5 or 6 or 7 or 9 or 10 or 15 or 20 or 25 or 30> T=0;
describe tree=0 plot=chrono_description;
end;
```

Twenty random model trees (replicates) for each model condition were generated using r8s. Using a custom script, the branches of each random model tree were scaled by a factor x so that the model height phylogeny is h. We examined two different h settings: a height of 1 was used throughout the study, except for experiments involving inferred gene trees where a height of 5 was used. This range of heights correspond to moderate to high levels of ILS based on the classification scheme of Vachaspati & Warnow (2015).

3.1.2 Generation of random model networks using ms

We added a single reticulation to each random model tree using the following procedure: (1) choose a random time unit t such that $0.01 \leq t \leq \frac{h}{4}$, and (2) add unidirectional migration, with a rate of five between two taxa or subpopulations such that migration occurs from $t - 0.01$ to $t + 0.01$. A single outgroup was added for each model network at coalescent time $1.5h$. We simulated 1000 gene trees for each random model network using ms (Hudson, 2002). The following ms command was used to generate the model network:

```
ms <number of taxa> <number of gene trees> -T -I k n_1 n_2 ...  n_k
-ej t i j -em t_1 i j 5.0 -em t_2 i j 0
```

The -T parameter outputs the gene trees that represent the history of the sampled taxa. The -I parameter is followed by k that represents the number of subpopulations. In our simulation study, we used k = <number of taxa>. (n_1 n_2 ... n_k) is a list of integers that represent the number of taxa sampled for each subpopulation. We sampled one taxon per subpopulation. The -ej parameter

specifies to move all lineages in subpopulation i to subpopulation j at time t. The first -em parameter sets migration at time t_1 from subpopulation j to subpopulation i to five. The second -em parameter sets migration at time t_2 from subpopulation j to subpopulation i to zero.

3.1.3 Simulation of sequences using seq-gen

The gene trees output by ms were used as input to seq-gen (Rambaut & Grassly, 1997), a sequence evolution program, which can simulate the evolution of sequences according to a finite-sites model. For each local genealogy simulated by ms, we simulated DNA sequence evolution using the Jukes-Cantor mutation model (Jukes & Cantor, 1969). The total length of the simulated sequences were 1000 kb distributed equally across all the local genealogies (1000 bp per local genealogy). The following command was used to simulate the evolution of sequences:

```
seq-gen -mHKY -l 1000 -s <0.02 or 0.04 or 0.08 or 0.16 or 0.32 or
0.64> <genetreefile >seqfile
```

The -mHKY parameter specifies the Jukes-Cantor mutation model. The -s parameter scales the branch lengths such that θ per base pair is < 0.02 or 0.04 or 0.08 or 0.16 or 0.32 or 0.64 >. The -l parameter specifies the length of a sequence in base pairs.

3.1.4 Species phylogeny inference

A single pipeline with two stages was used to infer a species phylogeny. The first stage consists of obtaining gene trees, where either true gene trees were used or FastTree was used to infer gene trees using the sequence alignments for the loci. The second stage uses the gene trees from the first stage to infer a species phylogeny.

3.1.4.1 Gene tree inference

Local gene tree inference using FastTree (Price *et al.*, 2009, 2010b) under the Jukes-Cantor model was used to infer the maximum-likelihood gene tree for each sequence alignment generated by seq-gen. Using the APE package (Paradis *et al.*, 2004) in R, the inferred gene trees from FastTree were rooted based on the outgroup. After rooting each inferred gene tree, the outgroup was dropped. The branch lengths of the inferred gene trees were scaled by FastTree in terms of expected number of substitutions. Using a custom script, we converted the branch lengths from expected number of substitutions to coalescent time using equation (3.1) in Hein *et al.* (2004).

3.1.4.2 Reconciliation of local gene trees into species phylogeny

We examined two different local gene tree input settings: the true gene trees generated by ms were used throughout the study, except for one experiment involving inferred gene trees. The following methods which are part of the PhyloNet package (Than *et al.*, 2008) were used to infer a species phylogeny:

- **MLE-length**: infers a species phylogeny under maximum likelihood (Yu *et al.*, 2014) using the topologies and branch lengths of the gene trees.
- **MLE**: infers a species phylogeny under maximum likelihood (Yu *et al.*, 2014) using the topologies of the gene trees.
- **MPL**: infers a species phylogeny under maximum pseudo-likelihood (Yu & Nakhleh, 2015) using the topologies of the gene trees.
- **MP**: infers a species phylogeny using a parsimony-based method (Yu *et al.*, 2013a) under the MDC criterion using the topologies of the gene trees.

The following is a sample NEXUS script file that was used to execute the PhyloNet commands:

```
#NEXUS
BEGIN TREES;
TREE gt1 = gene tree 1 in rich newick format
TREE gt2 = gene tree 2 in rich newick format
...
...
...
TREE gt1000 = gene tree 1000 in newick format
END;

BEGIN PHYLONET;
InferNetwork_ML (all) 1 -bl;
InferNetwork_ML (all) 1;
InferNetwork_MP (all) 1;
InferNetwork_MPL (all) 1;
END;
```

The commands located in the TREES block contain the gene trees. The commands located in the PHYLONET block contain the inference methods and parameters used to infer a species network. The InferNetwork_ML command infers a species network with one reticulation using maximum likelihood (Yu *et al.*, 2014). The -bl parameter specifies the use of branch lengths of gene trees in the inference. In the absence of -bl, only the topologies of gene trees are used in the inference. The InferNetwork_MP command infers a species network with one reticulation using a parsimony-based method (Yu *et al.*, 2013a) under the MDC criterion. The InferNetwork_MPL command infers a species network with one reticulation using maximum pseudo-likelihood (Yu & Nakhleh, 2015). We opted to assess the performance of these inference methods on the true number of reticulations by constraining the search to networks with only a single reticulation, which is following the practice

of prior simulation studies (Yu *et al.*, 2014; Solís-Lemus & Ané, 2016). Thus, our performance comparison focused on the simpler case of search among phylogenetic networks with the correct number of reticulations. Our findings therefore provide a bound on the performance of the methods in our study, since more complex networks are anticipated to present even greater scalability challenges.

We further used SNaQ (Solís-Lemus & Ané, 2016) to infer a species network. The following is a sample script used to execute the SNaQ commands:

```
d=readTrees2CF(<gene trees filename>);
T=readTopology(<starting topology filename>);
snaq!(T, d, hmax=1, <output filename>, outgroup=<outgroup name>);
```

The gene trees are summarized as quartet concordance factors using the readTrees2CF function. The readTopology reads the tree used as a starting point for the search. The starting tree was estimated using the MDC criterion. The snaq! command estimates a network using the input quartet concordance factors d and starting from tree T. hmax specifies the number of reticulations. outgroup specifies the outgroup taxon used to root the inferred network.

3.1.5 Performance measures

We measured the accuracy of the inferred species networks using two distance-based methods. A tripartition-based measure (Nakhleh *et al.*, 2003), which finds the proportion of tripartitions that are not shared between two networks, was used to compute the distance between the inferred and true phylogenetic networks. The following PhyloNet NEXUS script was used to compute the tripartition-based measure between two networks:

```
#NEXUS
BEGIN NETWORKS;
```

```
Network net1 = network in rich newick format
Network net2 = network in rich newick format
END;
BEGIN PHYLONET;
Cmpnets net1 net2 -m tri;
END;
```

An alternative method known as the splits distance, which is based on RF distance (Robinson & Foulds, 1981), was used to compute the distance between the tree-based edges of the inferred and true species networks. The splits distance counts the number of false positive bipartitions (bipartitions found in the inferred species network but not in true species network) and false negative bipartitions (bipartitions found in true species network but not in inferred species network). We used a custom R script to compute the splits distance.

We further evaluated the computational requirements of the inference methods, which was measured in terms of CPU runtime and memory usage. Each analysis was run on a 2.5 GHz Intel Xeon E5-2670v2 processor with 128 GiB of main memory.

3.2 Empirical study

We used mouse genomic sequence data sampled from natural *Mus* populations that were collected from previous studies (Liu *et al.*, 2014; Staubach *et al.*, 2012; Yang *et al.*, 2011; Song *et al.*, 2011; Yang *et al.*, 2009; Keane *et al.*, 2011). The collected samples represent 92 haploid mouse genomes (see Table A.2) that are either wild or wild-derived samples. The procedure that was used to generate the sequence data is described in the study of Liu *et al.* (2015). The sequences were filtered to 414,376 SNPs that were genotyped across all samples.

Datasets were constructed from the empirical samples using the following sampling procedure. For each dataset, we randomly selected one sample from each of the following mouse species or subspecies: *Mus musculus musculus*, *Mus musculus domesticus*, *Mus musculus castaneus*, *Mus*

spretus, *Mus spicilegus*, and *Mus macedonicus*. The sampling was repeated twenty times to obtain twenty datasets.

Recombination change gene tree topologies across different regions of a multiple genome alignment. We sampled 6 species (*Mus musculus musculus*, *Mus musculus domesticus*, *Mus musculus castaneus*, *Mus spretus*, *Mus spicilegus*, and *Mus macedonicus*) and used their multiple genome alignment as input to RecHMM, a hidden Markov model-based method (Westesson & Holmes, 2009), to identify recombination breakpoints. We computed local phylogeny switching breakpoints, which resulted in 3013 recombination-free genomic regions. FastTree using Generalized Time-Reversible model (Tavaré, 1986) was used to infer the gene tree for each recombination-free genomic region resulting in 3013 gene trees. We used rat (rn5 assembly from UCSC) as an outgroup to root each gene tree generated by FastTree.

MLE-length, MP, and SNaQ were used to infer species networks with zero or one reticulations. For inferred networks with zero reticulations, we measured the topological distance between inferred trees using the Robinson-Foulds distance. For inferred networks with one reticulation, the tripartition distance was used to compute the topological distance between inferred networks. We further compared the reticulations inferred by the inference methods to previous studies which detected two cases of gene flow: one among the *Mus musculus* subspecies (Staubach *et al.*, 2012), and the other between *Mus musculus domesticus* and *Mus spretus* (Liu *et al.*, 2015). Inference accuracy was evaluated by computing the proportion of replicates for which the inferred phylogeny was consistent with either of the two known instances of gene flow. Finally, we compared the inferred networks to the consensus *Mus* phylogeny proposed by Guénet & Bonhomme (2003).

3.3 Results

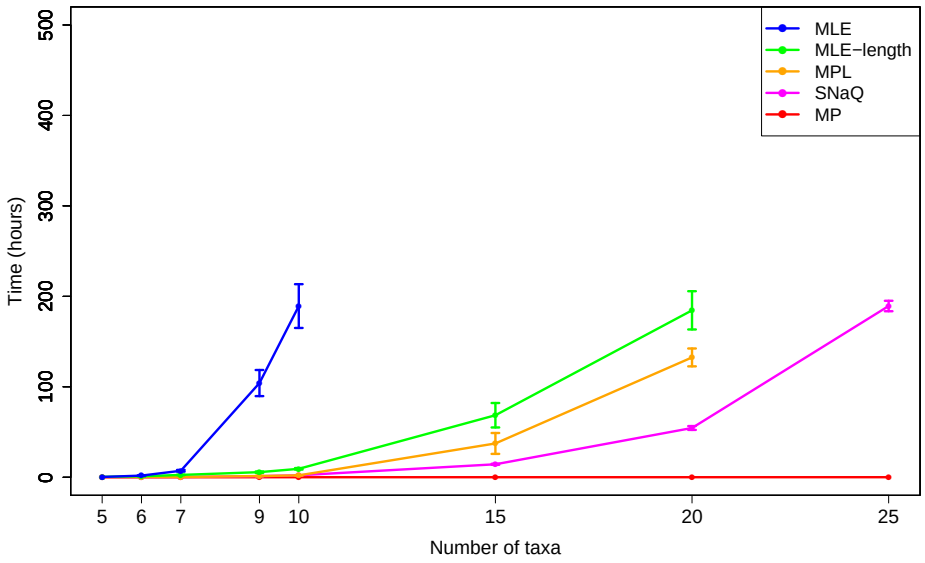
We began by assessing the effect of dataset size on computational time and memory requirements. We focused on the multi-locus methods since they were the most accurate in our experiments (see below). For all the multi-locus methods aside from MP, runtime became infeasible on datasets with more than 15 taxa. Of the full-likelihood approaches, MLE-length was faster than MLE. For

the pseudo-likelihood-based approaches, SNaQ was faster than MPL. We observed that SNaQ was the fastest among all the probabilistic multi-locus inference methods. The largest dataset for which MLE-length and SNaQ completed within ten days of runtime had 20 and 25 taxa, respectively. MLE-length and SNaQ required over ten days to complete on datasets with 25 and 30 taxa, respectively. MLE was slower compared to MLE-length, MPL, and SNaQ. On inputs with fewer than eight taxa, MLE analyses required total runtime of less than a day and approximately two gigabytes of main memory (Figure 3.1). As the number of taxa grew to nine, total runtime and main memory usage of MLE increased rapidly to around four days and ten gigabytes, respectively. Using MLE, the largest datasets for which analyses completed within a week of total runtime had nine taxa. The next largest datasets in our study had ten taxa; analyses of these datasets using MLE did not complete within a week but instead required around eight days of runtime and 12 gigabytes of main memory. We also attempted analyses of 15 taxon datasets using MLE; none of these analyses finished after multiple weeks of runtime.

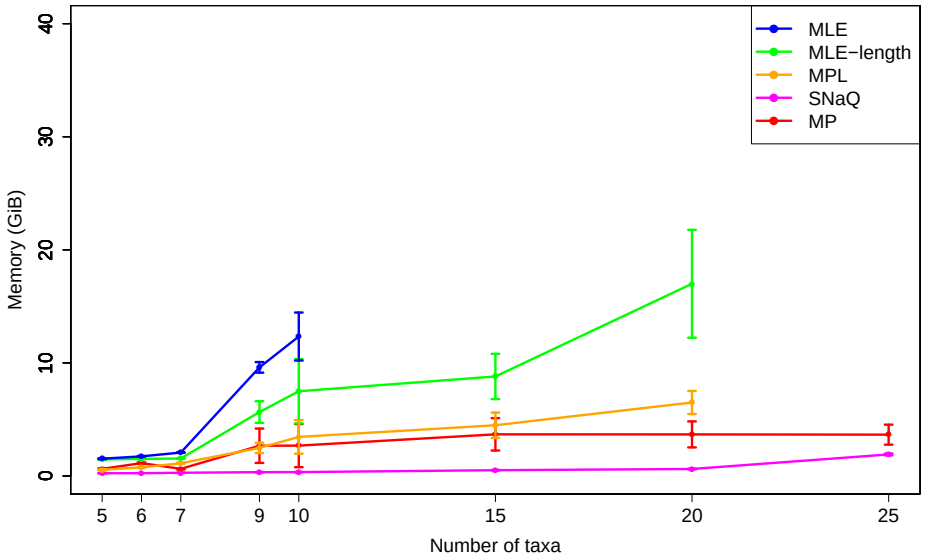
We additionally explored datasets with 40, 50, and 100 taxa on the following methods: MLE, MLE-length, MPL, and SNaQ; none of these analyses completed after ten days of runtime. The full-likelihood approaches (MLE and MLE-length) had runtime and memory usage that were higher than the pseudo-likelihood-based approaches (MPL and SNaQ). For all methods aside from MP, runtime developed super-linearly as dataset size increased. This growth in runtime is similarly observed in previous performance studies (Solís-Lemus & Ané, 2016; Yun, 2014; Yu *et al.*, 2013b), which suggest an increase in runtime as sampled dataset sizes grow.

Relative to runtime performance, the main memory requirements of the different methods contrasted to a greater degree (Figure 3.1). On datasets with more than seven taxa, the full-likelihood probabilistic approaches showed a super-linear growth in memory usage, which is similar to its performance in terms of runtime. The memory requirements of these full-likelihood methods are anticipated to increase rapidly on datasets with more than a couple of dozen taxa. MLE-length for the most part had smaller memory requirements than MLE. Interestingly, MP, MPL, and SNaQ had memory usage that was steady at around a couple of gigabytes on datasets

with up to 25 taxa. MP and MPL had memory usage beneath five gigabytes on datasets with up to 20 taxa. SNaQ's memory usage was consistent at around one gigabyte as dataset size increased from 5 to 20 taxa, and increased by only a few gigabytes as dataset size increased from 20 to 25 taxa. Overall, SNaQ's memory usage was the smallest among all multi-locus methods across all the dataset sizes examined in this performance study.



(a) Time



(b) Memory

Figure 3.1: **The computational requirements (time and memory) of multi-locus methods across different number of taxa.** The model conditions had dataset sizes ranging from five to twenty-five taxa. Results are shown for MLE, MLE-length, MPL, SNaQ, and MP analyses using true gene trees as input. The summary statistics (average and standard error) of the (a) main memory used (GiB) and (b) CPU runtime (hours) are reported across twenty replicates. The analysis of MLE on 15 taxa, MLE-length on 25 taxa, and MPL on 25 taxa did not complete after ten days of runtime.

We next examined the topological accuracy on simulations where the dataset scale increased in two ways: number of taxa and sequence divergence. We evaluated the performance of phylogenetic

network inference methods using the tripartition distance. On dataset sizes smaller than seven taxa, the probabilistic multi-locus methods returned higher accuracy compared to a parsimony-based approach. As seen in Figure 3.2, the full-likelihood approaches (MLE and MLE-length) were more accurate than the pseudo-likelihood-based approaches (MPL and SNaQ). Overall, the methods fell into four categories: (1) MLE-length was the most accurate, (2) MLE was the second most accurate, (3) SNaQ and MPL were the third most accurate, and (4) MP was the least accurate among all the multi-locus methods. Note that, for each replicate, the same set of gene trees was provided to each multi-locus method as input.

Overall, the topological error increased as the number of taxa increased (Figure 3.2). The topological error was generally the highest on the largest datasets in our study. We note here three exceptions to this observation: (1) MP was less accurate than the other methods across all dataset sizes where the topological error did not demonstrate a consistent pattern as dataset size increased, (2) MLE, MLE-length, MPL, and SNaQ's topological error decreased as dataset size increased from six to seven taxa, and (3) SNaQ's topological error decreased as dataset size increased from 20 to 25 taxa.

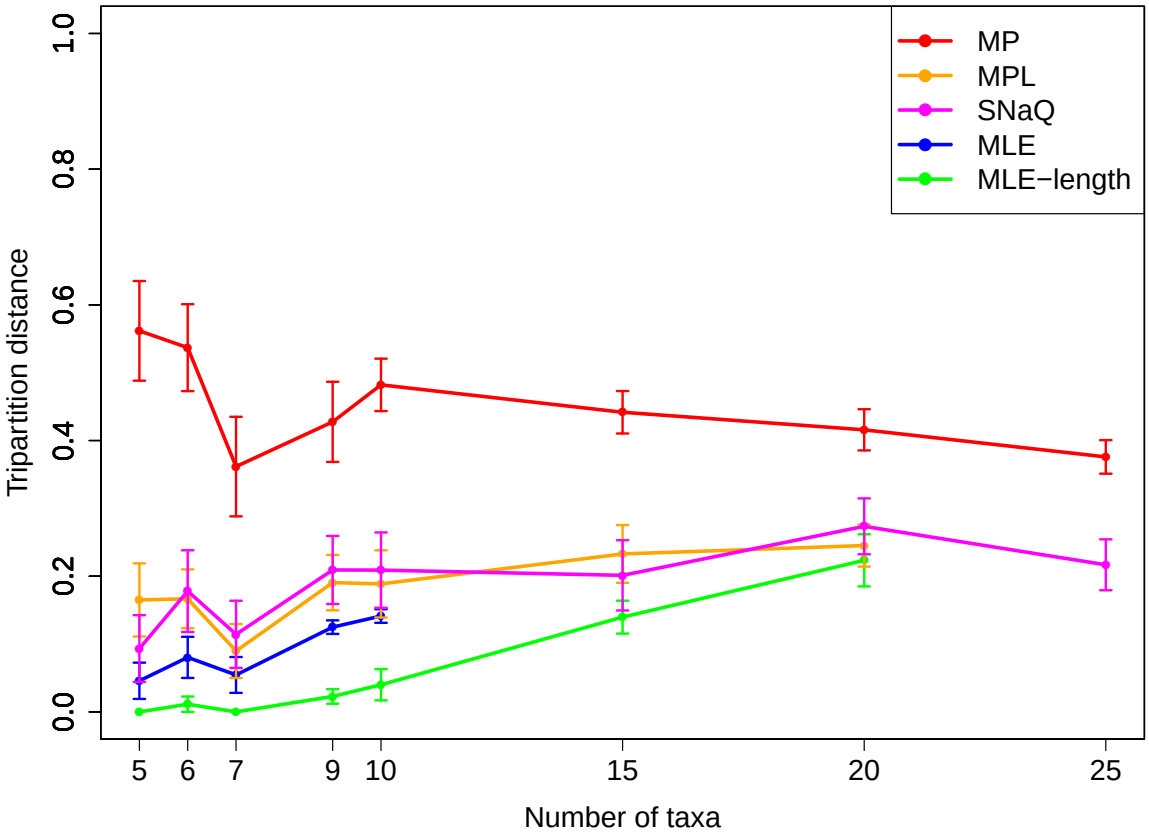


Figure 3.2: **The impact of dataset size on the topological error of multi-locus methods.** Results are shown for eight model conditions where the number of taxa ranged from five to twenty-five. The performance of five multi-locus methods (MP, MPL, SNaQ, MLE, and MLE-length) are reported using true gene trees as input. Average and standard error of the tripartition distance are reported for twenty replicates.

Furthermore, we evaluated the performance of the most accurate multi-locus inference method (MLE-length) as dataset size increased when inferred instead of true gene trees were given as input. As seen in Figure 3.3, the topological error of MLE-length increased as the number of taxa increased from five to ten.

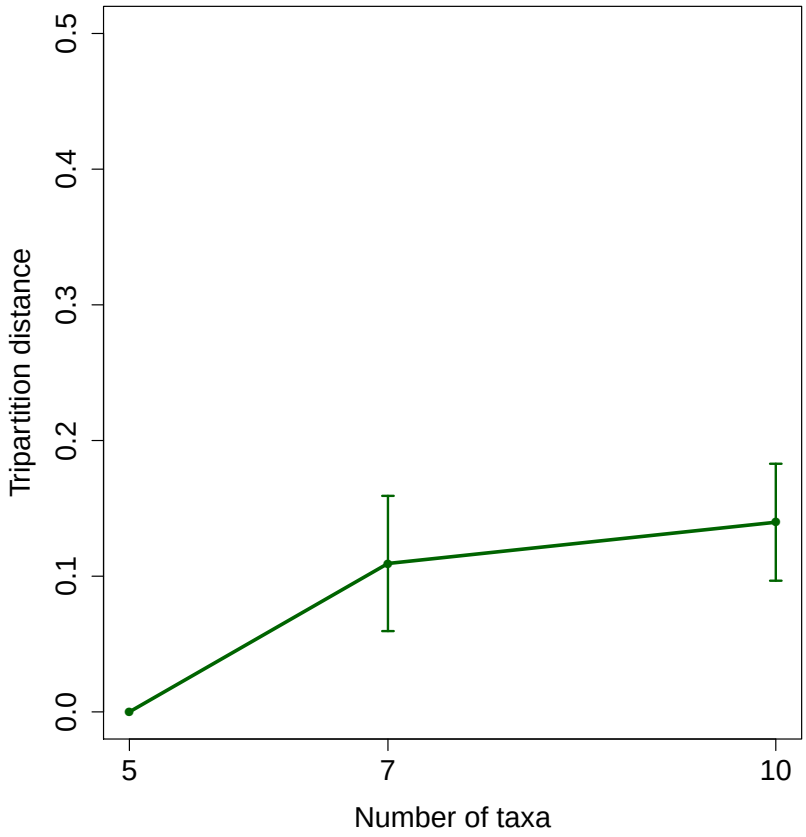


Figure 3.3: **The impact of dataset size on the topological error of MLE-length using inferred gene trees as input.** We assessed the performance of MLE-length to characterize the accuracy of multi-locus inference methods since MLE-length was generally more accurate than MLE, SNaQ, MPL, and MP. Results are shown for three model conditions where the number of taxa ranged from five to ten with θ of 0.08. Average and standard error of the tripartition distance between the inferred and model networks are reported for twenty replicates.

Compared to the effect of increasing the number of taxa, increases in the other dimension of scale – sequence divergence – similarly increased the average topological error of the most accurate multi-locus inference method (MLE-length) (see Figure A.1). As the sequence divergence increased from θ = 0.02 to 0.64, the topological error as measured by the tripartition distance increased from 0.08 to 0.6 with one minor exception; at the lowest level of sequence divergence (θ = 0.02), MLE-length was less accurate compared to the model condition with θ of 0.04. It is noteworthy that the model condition with the highest level of sequence divergence (θ = 0.64) exhibited the highest topological error in our simulation study.

We compared the most accurate multi-locus inference method (MLE-length) to two concatena-

tion methods which include NeighborNet (Bryant & Moulton, 2004) and the least squares method of Schliep (Schliep, 2009), which we refer to here as SplitsNet. We ran the concatenation methods using their default settings. The splits distance was used to evaluate the topological error, which quantifies the proportion of bipartitions that differ between the model and inferred phylogenies. As shown in Figure A.2, the three methods fell into three categories based on their topological accuracy: (1) MLE-length was the most accurate, (2) NeighborNet was the second most accurate, and (3) SplitsNet was the least accurate method. These results suggest that concatenation methods are less accurate than multi-locus inference methods. We also observed an increase in the topological error across all methods as the number of taxa increased from five to ten.

Our performance study utilized empirical samples from natural populations of *Mus musculus* subspecies and sister species (*Mus spretus*, *Mus spicilegus*, and *Mus macedonicus*). Prior studies detected gene flow between the *Mus musculus* subspecies (Staubach *et al.*, 2012) and between *Mus musculus domesticus* and *Mus spretus* (Song *et al.*, 2011; Liu *et al.*, 2015). We focused our comparison on the most accurate methods from each category of multi-locus methods: MLE-length from the full likelihood methods, SNaQ from the pseudo-likelihood-based methods, and MP. We omitted the concatenation methods from our comparison since they were among the least accurate of all methods in our simulation study.

At a coarse level, probabilistic inference using MLE-length was able to accurately detect gene flow in the empirical datasets. Specifically, the model selection criterion used by MLE-length consistently chose solutions with gene flow (i.e. phylogenetic networks with one reticulation) as opposed to solutions without gene flow (i.e. phylogenetic trees).

As shown in Table A.1, all of the methods inferred an identical species tree topology when constrained to infer a solution involving zero reticulations. For the inferred phylogenetic networks, greater topological similarity was observed among phylogenies inferred using the same method as opposed to phylogenies inferred using different methods. Furthermore, greater topological agreement was observed when solutions were constrained to have no gene flow, as opposed to solutions involving gene flow. Based on intra-method comparison of inferred networks, the greatest topo-

logical agreement was observed among MLE-length, followed by SNaQ, and then MP. Topological comparison of the different methods (i.e. MLE-length compared to MP, MLE-length compared to SNaQ, and MP compared to SNaQ) yielded topological distances which were the highest observed in our empirical study, and comparable disagreement was observed between the different pairs of methods as measured by average topological distance.

We further evaluated whether the methods detected known instances of interspecific and inter-subspecific gene flow: the former involving gene flow between *Mus musculus domesticus* and *Mus spretus* and the latter involving gene flow between the *Mus musculus* subspecies. MP, SNaQ, and MLE-length inferred a phylogenetic network consistent with gene flow between *Mus musculus domesticus* and *Mus spretus* in 12, 0, and 15 replicates, respectively (out of 20 replicates in total); the three methods inferred a network consistent with inter-subspecific gene flow among the *Mus musculus* subspecies in 0, 17, and 3 replicates, respectively.

We also compared the phylogenetic networks inferred by MLE-length, SNaQ, and MP to a consensus *Mus species* tree obtained from prior literature studies (Guénet & Bonhomme, 2003). The bipartitions in the consensus tree were consistently inferred by the different methods (see Figure A.3 for the consensus tree). Compared to MP, the probabilistic multi-locus methods more frequently inferred reticulations that were consistent with known interspecific/intersubspecific gene flow; however, the methods largely disagreed on the exact location of reticulation within the phylogeny.

3.4 Discussion

Among the multi-locus inference methods using true gene trees, MLE-length inferred phylogenetic networks with the greatest topological accuracy on all datasets examined in our study. For this reason, we focus on MLE-length in our discussion of the performance of this category of methods. The relative performance of the probabilistic methods compared to MP is consistent with previous performance studies of phylogenetic tree inference methods (Philippe *et al.*, 2011; Felsenstein, 1978). We conjecture that, as in the tree inference case, long branch attraction plays a role

in the relative performance of probabilistic and parsimony-based phylogenetic network inference methods.

On datasets with more than 25 taxa, MLE-length's computational requirements are projected to be nearing the limits of the most powerful computational clusters available to us and similar computational resources. This dataset size is the largest in our study and yet is not considered large in the context of today's phylogenomic studies. Relative to current phylogenomic studies, we note that the mutation rates explored in our study spanned a larger range of biologically realistic values compared to the dataset sizes explored in our study; our exploration of larger dataset sizes was primarily constrained by the large computational requirements of the multi-locus inference methods. We speculate that the multi-locus inference methods will be unable to analyze inputs with more than hundred taxa due to main memory requirements.

Increasing either of the two dimensions of scale – the number of taxa and sequence divergence – reduced the topological accuracy of the species phylogeny inferred by MLE-length. We speculate that one contributing factor may be the heuristic approaches necessary for analysis of NP-hard optimization problems; practical issues such as local optima in the search space can pose major challenges to the performance of these heuristics. Another contributing factor was inferred gene tree error. Increasing mutation rates reduced the accuracy of inferred gene trees, which is consistent with theoretical expectations and empirical observations about long branch attraction in other phylogenetic studies (Philippe *et al.*, 2011).

In this chapter, we conducted a performance study to highlight two scalability issues with current state-of-the-art phylogenomic network inference methods: topological accuracy and computational time/space requirements. The best performing methods in terms of topological accuracy were the probabilistic inference methods which maximize likelihood under coalescent-based models by searching among all possible phylogenetic networks for a given input set of taxa. In general, we found that topological accuracy degrades as the number of taxa increases; a similar effect was observed with increased sequence mutation rate. The improved accuracy obtained with these inference methods comes at a computational cost in terms of runtime and main memory usage,

which quickly becomes prohibitive as dataset size grows past ten taxa.

Relative to the scope of current phylogenomic studies, the state of the art of phylogenetic network inference is near or at the limits of scalability. New algorithmic development is critically needed to address this methodological gap.

CHAPTER 4

FASTNET: FAST AND ACCURATE INFERENCE OF PHYLOGENETIC NETWORKS USING LARGE-SCALE GENOMIC SEQUENCE DATA

Interspecific gene flow is an evolutionary process that results in gene tree incongruence and non-tree-like phylogenetic relationships among species. These non-tree-like relationships necessitate the use of more complex representations such as phylogenetic networks. To study and model phylogenetic networks, we need two ingredients (1) densely sampled and divergent genomic sequence data and (2) computational methods that are capable of accurately and efficiently inferring phylogenetic networks on large-scale genomic sequence datasets.

Recall from chapter 3 that the probabilistic phylogenetic network inference methods were found to be most accurate among state-of-the-art methods, and MLE-length was the most accurate method within this class. MLE was the second most accurate while MPL was third; MPL was designed to tradeoff optimization under a pseudo-likelihood-based approximation to the full likelihood for increased computational efficiency compared with full likelihood methods. However, the tradeoff netted efficiency that was well short of current phylogenomic dataset sizes: given a week of runtime in our previous study, we found that MPL was capable of analyzing datasets with 20 taxa but no larger. These results suggest that the scalability of the state of the art falls well short of that required by current phylogenetic studies, where many dozens or hundreds of divergent genomic sequences are common.

One observation obtained from chapter 3 is that phylogenetic network inference methods are faster and more accurate for smaller number of taxa and evolutionary divergence. One of the insights obtained from this observation is the following: what if we divide our problem into smaller subproblems? In doing so, we could solve each subproblem separately, thus constraining the scalability limits that arise from the number of taxa and evolutionary divergence, and then merge the solutions of these subproblems into a single solution. Interestingly, there is an algorithm design paradigm known as divide-and-conquer that could help solve this problem. Other studies (Liu

et al., 2009) have successfully applied divide-and-conquer approaches to enable scalable inference in the context of species tree estimation. In this chapter, we introduce FastNet (Hejase *et al.*, 2017b), which is a method that uses divide-and-conquer for phylogenetic network inference using large-scale datasets. Figure 4.1 shows a flowchart of the FastNet algorithm.

In this chapter, we focus on addressing two dimensions of scalability constraints. First, we focus on dataset size in terms of the number of taxa in the species phylogeny. Second, we focus on the number of reticulations in the species phylogeny. We note here that scalability limits could arise from other dimensions of constraints such as allele sampling for each taxon and evolutionary divergence. We further focus on deep and non-deep gene flow, which has been the focus of recent high-profile studies (Leaché *et al.*, 2014; Green *et al.*, 2010).

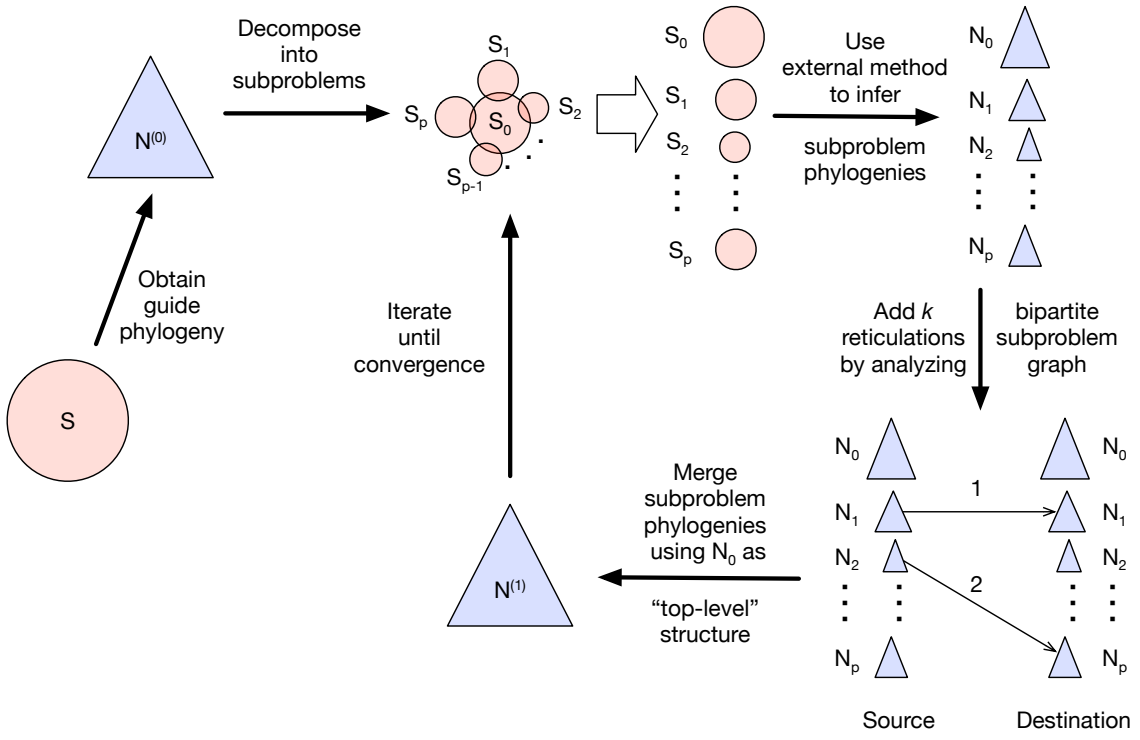


Figure 4.1: **Flowchart of the FastNet algorithm.** We first infer a guide phylogeny $N^{(0)}$ from a given input problem S. Next, we decompose the input problem into subproblems using $N^{(0)}$. We then infer phylogenies N_i on subproblems S_i where $0 \leq i \leq$ p using an external base method (i.e. MPL or MLE or MLE-length), and then proceed with gene flow detection by analyzing the subproblem phylogenies as a bipartite graph. Finally, we merge the subproblem phylogenies using the top-level structure N_0.

4.1 Method

We now describe our new divide-and-conquer algorithm, which we refer to as FastNet. Let $\mathbb{X}$ is the set of taxa. The input is a set of gene trees G from a set of independent loci and total number of reticulation nodes in the output phylogeny c_τ. The output consists of a directed phylogenetic network N where each leaf in N corresponds to a taxon $x \in \mathbb{X}$.

4.1.1 Guide phylogeny inference

The FastNet algorithm requires inferring a guide phylogeny $N^{(0)}$ for the subsequent subproblem decomposition step. The relationships in a guide phylogeny measure the evolutionary relatedness among subproblems while constraining the evolutionary divergence of taxa in a subproblem. There are two requirements for inferring a guide phylogeny: (1) a method needs to be sufficiently accurate to inform subsequent divide-and-conquer steps, and (2) a method needs to have reasonable computational requirements. We explored two approaches for guide phylogeny inference. First, we used a guide tree, where ASTRAL (Mirarab & Warnow, 2015), a state-of-the-art phylogenomic tree inference method, was used to infer a species tree $N^{(0)}$. Given that ASTRAL infers unrooted and undirected species trees, we inferred a rooted guide phylogeny by making use of an outgroup taxon, which is used to root the species tree on the leaf edge corresponding to that outgroup taxon. Another limitation of using ASTRAL in this context is that it effectively infers species trees for evolutionary scenarios lacking gene flow. The assumption of tree-like evolution is generally invalid for the computational problem that we consider. Thus, an alternative approach is to use a guide network in lieu of guide tree. For this purpose, we used a parsimony-based approach (Yu *et al.*, 2013a) to infer a guide network, and then took the guide phylogeny tree encoded in the inferred network by dropping all minor hybridization edges based on hybridization frequencies Γ.

4.1.2 Subproblem decomposition

The decomposition S induces a partitioning of the set of taxa into disjoints subsets S_i where $\cup\, S_i = S$ for $1 \leq i \leq p$. The decomposition procedure proceeds by solving an optimization problem using a greedy algorithm subject to the following two constraints: (1) maximum subproblem size c_m and (2) lower bound on the number of subproblems c_s. Apart from parameterizing our divide-and-conquer based upon a different set of optimization criteria, the decomposition algorithm is similar to the Center-Tree-i decomposition used by Liu *et al.* (2009) in the context of species tree inference. We further construct an ancestral subproblem subset S_0 where the closest taxon to the root, which was computed by the number of edges to get from the tip to the root of the guide phylogeny, is sampled for each subset S_i for $1 \leq i \leq p$.

4.1.3 Bipartite graph

We convert the set of subsets S_i for $0 \leq i \leq p$ into a bipartite graph such that the two disjoint sets contain subsets S_i for $0 \leq i \leq p$. Using Algorithm 1, we initialize the number of bipartite graph edges that connect the two disjoint sets to zero. The next bipartite graph is iteratively enumerated and is based on the previous iteration. We note here that there are $\frac{p^2-p}{2} + p$ possible enumerations in the bipartite graph per iteration. Algorithm 2 iterates through all possible enumerations in a bipartite graph, and computes a pseudo-likelihood score for each enumeration instance.

Algorithm 1 Bipartite subproblem graph enumeration

procedure InitializeBipartiteGraphEnumeration(p, k)
    for $m = 1$ **to** $\frac{p^2-p}{2} + p$ **do**
        for $i = 0$ **to** p **do**
            for $j = i$ **to** p **do**
                $B_{k,m}[i, j] = 0$ ▷ Initialize number of reticulations in bipartite subproblem graph to zero
    return (B)
procedure NextBipartiteGraphEnumeration(p, k, $best$)
    $m = 1$
    for $i = 0$ **to** p **do**
        for $j = i$ **to** p **do**
            $B_{k,m}[i, j] = B_{k-1,best}[i, j] + 1$
            m += 1
    return (B)

Algorithm 2 Calculate optimization score for bipartite subproblem graph

procedure ComputeOptimizationScoreAndInferSubproblemSolutions(p, B, k, Δ, δ)
    for $m = 1$ **to** $\frac{p^2-p}{2} + p$ **do**
        for $i = 0$ **to** p **do**
            for $j = i$ **to** p **do**
                InferSubnetwork(B, k, m, i, j, Δ_m, δ_m) ▷ Store inferred network in Δ ▷ Store inferred network likelihood in δ
        $score_m = \prod_{\substack{0 \le i \le p \\ i \le j \le p}} \delta_m[i, j, k]$ ▷ Pseudolikelihood score
    Δ_m = FindBestScore(score)
    return Δ_m

4.1.4 Gene flow detection and inferring phylogenies on subproblems

In this inference problem, we seek to detect gene flow (or lack thereof) within the subproblems and between subproblems produced from decomposition. The input includes the subproblem subsets S_i for $1 \le i \le p$, the ancestral subproblem subset S_0, and the parameter c_τ, which specifies the number of reticulation nodes in the output phylogeny. Furthermore, we include the gene trees G inferred from its corresponding sequence alignments as input rather than the sequence data itself, since we

utilize summary-based approaches to address this problem. Let G_s be the restriction of gene trees G to the set of subproblem taxa for $s \in S$. The output is a function $\Delta(i, j, c_\tau)$ that stores the inferred phylogenies in the bipartite graph, where the total number of reticulation nodes is subject to the constraint $\sum_{\substack{0 \le i \le p \\ i \le j \le p}} NumberOfReticulations(\Delta(i, j, c_\tau)) = c_\tau$.

For the base methods, we make use of existing statistical summary-based methods for phylogenetic network inference. Let Ψ be such a method which infers species networks under an evolutionary model with parameters θ. In this study, we considered three choices for Ψ which were among the most accurate methods in the previous chapter. One suitable choice which was shown to be accurate in the previous chapter are the probabilistic network inference methods that use full model likelihood calculations (i.e. MLE-length and MLE). Another suitable choice is a method that uses pseudo-likelihood approximations to the full model likelihood calculations (i.e. MPL). Let $F_{\Psi,\theta}(G_s, c_\tau)$ be the species network inferred by method Ψ under its evolutionary model θ using the subproblem gene trees input G_s and c_τ be the total number of reticulations in the output phylogeny, and let $L_{\Psi,\theta}(G_s, c_\tau)$ be the corresponding model likelihood of the network. Algorithm 3 describes details on inferring subproblem solutions, which computes a network and its likelihood for a particular instance in the bipartite graph. The inferred networks and its likelihood scores are stored in Δ and δ, respectively, and then an optimization criteria described in Algorithm 2 is used to return the best enumeration to be used in the next iteration.

Algorithm 3 Inference of subproblem solutions

static variable G ▷ Set of gene trees

procedure InferSubnetwork(B, k, m, i, j, Δ_m, δ_m)
  if ($i == j$) **then**
    if defined($\Delta[i, j, k]$) **then**
      return
    (N_{ij}, score) = $F(G_{s_i}, B_{k,m}[i, j])$ ▷ Base method $F(\cdot, \cdot)$
    ▷ G_{s_i} is restriction of G to subproblem taxa S_i
  else
    if defined($\Delta[i, j, k]$) **then**
      return
    if not defined($\Delta[i, i, k]$) **then**
      InferSubnetwork(B, k, m, i, i, Δ_m, δ_m)
    if not defined($\Delta[j, j, k]$) **then**
      InferSubnetwork(B, k, m, j, j, Δ_m, δ_m)
    N^{cherry} = ConstructCherry($\Delta[i, i, k]$, $\Delta[j, j, k]$) ▷ Returns (($\Delta[i, i, k]$:b_i, ($\Delta[j, j, k]$:b_j); ▷ where b_i and b_j are inferred using base method $F(\cdot, \cdot)$
    (N_{ij}, score) = AddReticulations(N^{cherry}) ▷ Use base method $F(\cdot, \cdot)$ to perform constrained search and add reticulations
    $\Delta_m[i, j, k] = N_{ij}$ ▷ Return value by reference to mutable cache Δ
    $\delta_m[i, j, k]$ = score ▷ Return value by reference to mutable cache δ

The procedure for detecting gene flow proceeds as follows. First, we enumerate all possible assignments $\Delta_m[i, j, c_\tau]$ for $1 \leq m \leq \frac{p^2-p}{2} + p$, $0 \leq i \leq p$, and $i \leq j \leq p$. The key idea here is that for each possible assignment $\Delta_m[i, j, c_\tau]$, we use method Ψ to infer subproblem networks $\Delta_m[i, j, c_\tau]$ and its corresponding model likelihoods $\delta_m[i, j, c_\tau]$. Then, method Ψ is used to calculate the model likelihood $P(G|N_{\Delta_m}, \theta)$, which is used as the optimization criterion. The output of the gene flow detection procedure are subproblem phylogenies Δ_m and subproblem scores δ_m that optimize the likelihood $P(G|N_{\Delta_m}, \theta)$.

For more speed, we relaxed the model likelihood calculations utilized by our base method where instead of optimizing the full model likelihood $P(G|N_\Delta, \theta)$, we optimized the product of subproblem likelihoods $\prod_{\substack{0 \leq i \leq p \\ i \leq j \leq p}} \delta_m[i, j, c_\tau]$. This calculation is an approximation since it assumes that subproblems are independent. However, we note here that these subproblems are correlated

through the top-level structure in the ancestral subproblem phylogeny N_0. Given the optimal assignment Δ, each subproblem is represented using a rooted subproblem network $F_{\Psi,\theta}(G_s, \Delta(s))$.

Algorithm 4 searches for the optimal subproblem decomposition bipartite graph. It iteratively goes through c_τ iterations, where at each iteration it enumerates a bipartite graph, computes an optimization score along with subproblem solutions, and identifies the best enumeration to be used in the next iteration.

Algorithm 4 Search for optimal subproblem decomposition graph

```
1: static variable c_τ                                          ▷ Number of reticulations
2: static variable G                                            ▷ Set of gene trees

3: procedure ExhaustiveSearchForOptimalSubproblemDecompositionGraph(G)
4:     p = GetNumberOfSubproblems(S)
5:     k = 0
6:     B_{k,m} = InitializeSubproblemDecompositionGraph(p, k)
7:     [best, Δ] = ComputeOptimizationScoreAndInferSubproblemSolutions(p, B_{k,m}, k, Δ, δ)
8:     for k = 1 to c_τ do
9:         B_{k,m} = NextBipartiteGraphEnumeration(p, k, best)
10:        [best, Δ] = ComputeOptimizationScoreAndInferSubproblemSolutions(p, B_{k,m}, k, Δ, δ)
11:    return(B_{c_τ,best}, Δ)
```

4.1.5 Merging subproblem phylogenies

The merging procedure proceeds as follow (see Algorithm 5 for more details). Let p be the number of subproblems. First, we use ancestral subproblem phylogeny N_0 (i.e. "top-level" subproblem phylogeny) and optimal assignment Δ (i.e. Δ with the maximum model likelihood score) to merge subproblem phylogenies. The "top-level" structure of the output phylogeny is resolved using the ancestral subproblem phylogeny N_0 encoded in the optimal assignment Δ. For each bottom-level subproblem phylogeny N_i for $1 \leq i \leq p$, we replace it with its corresponding leaf in N_0. Finally, we add any reticulation edges in Δ that span different subproblems such that these reticulation edges are compatible with their respective subproblem phylogenies.

Algorithm 5 Merge

```
1: static variable c_τ                                          ▷ Number of reticulations

2: procedure MERGE(B, Δ)
3:     p = GetNumberOfSubproblems(S)
4:     N = Δ[0, 0, c_τ]                                         ▷ "Top-level" subproblem phylogeny
5:     for i = 1 to p do
6:         ReplaceLeafWithSubnetwork(N, B, i, Δ[i, i, c_τ])     ▷ Replace ith taxon in "top-level" N
7:     for i = 0 to p do
8:         for j = i + 1 to p do
9:             AddCompatibleReticulations(N, B, i, j, Δ[i, j, c_τ])
10:    return(N)
```

4.2 Performance study

Below we describe the steps used in the performance study.

4.2.1 Simulation of model networks

We simulated random model trees using r8s version 1.7 (Sanderson, 2003) for 15, 20, 25, and 30 taxa. Twenty random model tree replicates were generated where the height of each tree replicate was scaled to five. One, two, three, or four reticulations were added for each tree replicate using the following two settings. In the first setting, which we refer to as deep gene flow, we select two taxa (i.e. two leaf tips), and then add unidirectional migration occurring from 0 to t with a rate of 5.0 between the two selected taxa. For the second setting, which we refer to as non-deep gene flow, we choose a random time unit t, and then add unidirectional migration, with a rate of 5.0 between two subpopulations such that migration occurs from t - 0.01 to t + 0.01, where at least the sender or recipient populations must contain at least two taxa. After generating a random model network, an outgroup was added at coalescent time 15. 1000 gene trees were simulated for each random model network using ms (Hudson, 2002). In our simulation study, we sampled one allele per taxon.

4.2.2 Simulation of the evolution of DNA sequences

The evolution of sequences was simulated using seq-gen (Rambaut & Grassly, 1997), which takes the gene trees generated by ms as input and simulates the evolution of sequences according to a finite-sites model. Using the Jukes-Cantor mutation model (Jukes & Cantor, 1969), we simulated the evolution of DNA sequences for each local genealogy generated by ms. The simulated sequence had a total length of 1000 kb, which was equally distributed across all local genealogies (1000 bp per local genealogy).

4.2.3 Gene tree inference

FastTree (Price *et al.*, 2009, 2010b) was used for local gene tree inference. We used the Jukes-Cantor model (Jukes & Cantor, 1969) to infer the maximum-likelihood gene tree for each sequence alignment generated by seq-gen (Rambaut & Grassly, 1997). The inferred gene trees were rooted using the outgroup.

4.2.4 Performance measures

We evaluated the inference methods using multiple criteria. The first evaluation criterion involved computing the topological accuracy of the method under study. Topological accuracy was computed by comparing the inferred phylogeny with the model phylogeny using the tripartition distance (Nakhleh *et al.*, 2003), which counts the proportion of tripartitions that are not shared between the inferred and model network. The second evaluation criterion is the computational requirements of the method, which was measured in terms of CPU runtime.

4.3 Yeast dataset

We used genomic sequence data from a yeast dataset that was studied by Salichos & Rokas (2013). A recent study has highlighted historical gene flow between some of the populations in this study (Yu & Nakhleh, 2015). The collected sample information contained 23 yeast genomes along with 1070 genes. Using the approach of Yu *et al.* (2011), we rooted the gene trees under the

MDC criterion along with the species tree inferred using concatenation in Salichos & Rokas (2013). The rooted gene trees were used as input to FastNet, with MPL as a base method, to infer species networks with one or two reticulation nodes where slope analysis was used for model selection.

4.4 Mosquito dataset

We reanalyzed the mosquito dataset of Neafsey *et al.* (2015) using FastNet. After filtering and processing, we collected 5099 genes across 18 mosquito genomes. A recent study (Fontaine *et al.*, 2015) has highlighted gene flow by analyzing six members in the *Anopheles gambiae* complex that include *Anopheles gambiae*, *Anopheles arabiensis*, *Anopheles coluzzii*, *Anopheles quadriannulatus*, *Anopheles merus*, and *Anopheles melas*. This study reported a phylogenetic network containing gene flow between members of the *gambiae* complex, where gene tree incongruence was due to introgression and ILS. Using the approach of Yu *et al.* (2011), we rooted the gene trees under the MDC criterion along with the species tree inferred using concatenation in Neafsey *et al.* (2015). The rooted gene trees were used as input to FastNet, with MPL as a base method, to infer species networks with one or two reticulation nodes where slope analysis was used for model selection.

4.5 Results

We report the performance boost of FastNet relative to other leading phylogenetic inference methods. Table 4.1 reports the performance boost of FastNet, using MLE-length as a base method, relative to the base method itself using model conditions that varied dataset sizes from 15 to 20 taxa with a single reticulation and deep gene flow. In comparison to MLE-length, FastNet, using true gene trees as input, was faster by 49 and 114 hours for the 15 and 20 taxa model conditions, respectively. The runtime performance boost of FastNet relative to MLE-length was significant (using t-test) where multiple test correction was performed using the approach of Benjamini & Hochberg (1995). We observed similar performance improvement in terms of accuracy where FastNet was significantly more accurate than MLE-length in terms of tripartition distance by 0.103 and 0.195 for the 15 and 20 taxa model conditions, respectively. We further explored the

performance of FastNet relative to the base method MLE-length using inferred gene trees as input. We observed that FastNet was faster than MLE-length by 15 and 43 hours for the 15 and 20 taxa model conditions, respectively. Furthermore, FastNet was more accurate than MLE-length by 0.231 and 0.195 for the 15 and 20 taxa model conditions, respectively. We observed that the relative runtime performance boost of FastNet relative to MLE-length decreased when inferred instead of true gene trees were used. On the other hand, the relative accuracy performance boost increased using inferred instead of true gene trees.

Number of taxa	**Gene Trees**	**Topological distance**			**Runtime in hours**		
		Avg	**SE**	**q value**	**Avg**	**SE**	**q value**
15	True	0.103	0.021	8.8×10^{-4}	49.401	6.862	9.1×10^{-7}
15	Inferred	0.231	0.002	1.3×10^{-4}	15.433	2.045	6.7×10^{-7}
20	True	0.195	0.024	6.1×10^{-5}	114.337	14.650	3.3×10^{-7}
20	Inferred	0.195	0.005	5.8×10^{-5}	43.166	7.256	1.7×10^{-5}

Table 4.1: **Average distances and runtimes for the performance boost of FastNet (with MLE-length as a base method) over the base method itself using model conditions containing 15 or 20 taxa.** The topological distance between the inferred and model phylogenies was measured using the tripartition distance. The model conditions involved model phylogenies that contained one reticulation node with deep gene flow. True or inferred gene trees were used as input to FastNet and MLE-length. Average ("Avg") and standard errors ("SE") for the performance improvement of topological distances and runtimes are listed (n = 20). A one-sided t-test comparing the performance advantage of FastNet over the boosted method (MLE-length) for the evaluation criteria (i.e. topological distance and runtime) was conducted. Corrected q-values are reported where multiple test correction was performed using the approach of Benjamini & Hochberg (1995).

Table 4.2 reports the performance boost of FastNet, using MPL as a base method, relative to the base method itself using deep gene flow model conditions that varied the number of taxa from 15 to 30 and number of reticulations from one to four. FastNet was significantly more accurate than MPL on all model conditions. The largest performance improvement in terms of tripartition distance of FastNet relative to MPL was observed on model conditions using true gene trees containing 30 taxa and four reticulations, with a tripartition distance improvement of 0.413. In terms of running time, FastNet was significantly faster than MPL across all model conditions, with the largest performance improvement of 35 hours observed on dataset size of 25 taxa with three reticulations on model conditions using true gene trees. On the other hand, the smallest performance improvement of

2.8 hours was observed on dataset size of 15 taxa with one reticulation on model conditions using true gene trees. Furthermore, we explored the relative performance boost of FastNet to MPL when inferred instead of true gene trees were used. We observed that the topological accuracy performance boost of FastNet over the base method decreased while the runtime performance boost increased in comparison to the analyses that used true gene trees.

Number of taxa	Gene trees	Number of reticulations	Topological distance			Runtime in hours		
			Avg	SE	q value	Avg	SE	q value
15	True	**1**	0.087	0.036	3.3×10^{-2}	2.820	0.307	7.2×10^{-5}
15	Inferred	**1**	0.071	0.021	1.2×10^{-2}	3.810	0.481	7.7×10^{-5}
20	True	**2**	0.346	0.036	1.1×10^{-5}	9.630	0.07	1.11×10^{-2}
20	Inferred	**2**	0.134	0.017	1.4×10^{-2}	15.095	1.710	6.9×10^{-6}
25	True	**3**	0.281	0.024	7.9×10^{-5}	35.586	5.577	8.5×10^{-4}
30	True	**4**	0.413	0.001	8.8×10^{-12}	30.284	6.508	2.8×10^{-2}

Table 4.2: **Average distances and runtimes for the performance boost of FastNet (with MPL as a base method) over the base method itself using model conditions that varied the number of taxa and reticulations.** The model conditions involved model phylogenies that contained one, two, three, and four reticulations with dataset sizes of 15, 20, 25, and 30 taxa, respectively, with deep gene flow. Table layout and description are otherwise similar to Table 4.1.

The purpose of FastNet is to boost existing methods, much like previous tree-based divide-and-conquer methods. Using true gene trees as input, we evaluated the performance boost of FastNet using MLE-length or MPL as base methods compared to the base methods themselves on model conditions with non-deep gene flow (Tables B.1 and B.2). For FastNet with MLE-length as a base method, we observed a boost in terms of accuracy (as measured by the tripartition distance) of 0.066 and 0.070 for the 15 and 20 taxon datasets, respectively. As for runtime, the performance boost ranged from 34 to 71 hours. FastNet, using MLE-length as a base method, was significantly more accurate and faster than the boosted method. Similarly, we compared the performance of FastNet, using MPL as a base method, relative to the base method itself on model conditions with non-deep gene flow. We observed a performance boost that ranged from 0.015 to 0.166 and 2 to 8 hours for the accuracy (as measured by the tripartition distance) and runtime, respectively, on model conditions where dataset size ranged from 15 to 20 taxa. The performance advantage of FastNet (with MPL as a base method) over the boosted method was significant on all model conditions with one minor exception; on the smallest dataset size of 15 taxa with a single reticulation, FastNet

improved upon the base method by 0.015 with this performance improvement being not significant.

Table 4.3 reports the absolute performance of FastNet, using MLE as a base method, on model conditions with dataset sizes of 15 and 20 taxa. We note here that the base method MLE did not complete on these datasets after one week of runtime. The absolute topological error in terms of tripartition distance of FastNet was 0.034 and 0.075 on the 15 and 20 taxon datasets, respectively, while the absolute runtime of FastNet ranged from one to eight hours.

Number of taxa	**Topological distance**		**Runtime in hours**	
	Avg	**SE**	**Avg**	**SE**
15	0.034	0.012	1.292	0.265
20	0.075	0.027	8.167	1.847

Table 4.3: **Average distances and runtimes for the performance of FastNet (with MLE as a base method) using model conditions containing 15 or 20 taxa.** The topological distance between the inferred and model phylogenies was measured using the tripartition distance. The model conditions involved model phylogenies that contained one reticulation node with deep gene flow. Inferred gene trees were used as input to FastNet. Average ("Avg") and standard errors ("SE") of topological distances and runtimes are listed (n = 20). The base method (MLE) was unable to finish on each dataset after one week of runtime.

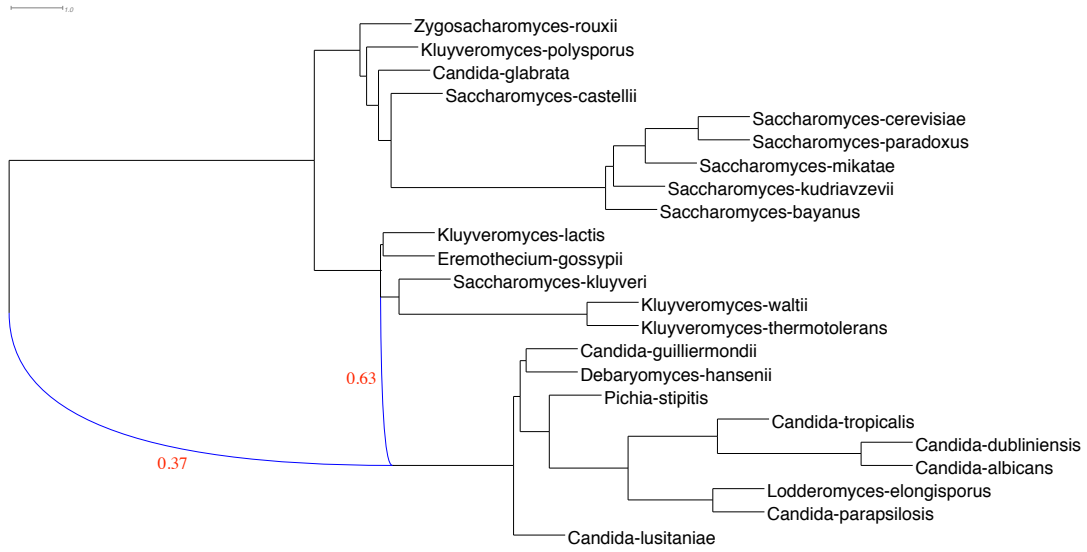
We evaluated the performance of FastNet across different number of loci (i.e. number of loci varied from 100 to 1000) using dataset size of 20 taxa. We observed that as we increased the number of loci from 100 to 1000, the topological distance as measured by the tripartition distance decreased from 0.094 to 0.075.

Number of loci	**Topological distance**	
	Avg	**SE**
100	0.094	0.028
200	0.078	0.024
1000	0.075	0.027

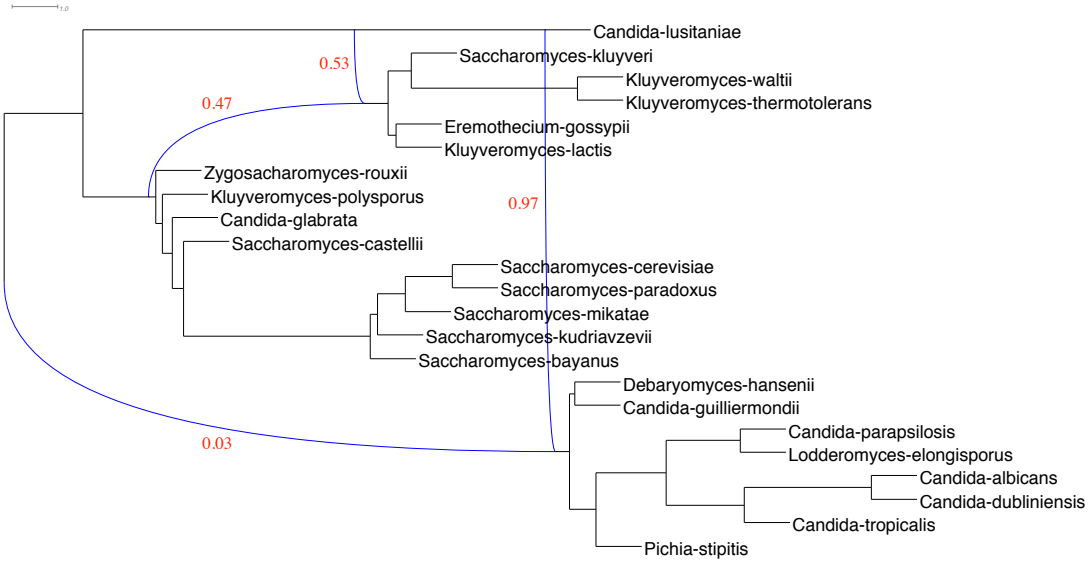
Table 4.4: **Topological error of FastNet (with MLE as a base method) on single reticulation node model conditions where the number of loci per replicate dataset ranged between 100 and 1000.** The model conditions consisted of dataset size of 20 taxa with deep gene flow. The topological accuracy of each inferred phylogeny with respect to the model phylogeny was evaluated using the tripartition distance. Inferred gene trees were used as input to FastNet. Average ("Avg") and standard errors ("SE") of topological distances and runtimes are listed (n = 20).

We applied FastNet on the 1070-gene yeast dataset of Salichos & Rokas (2013) and estimated a

phylogeny by coupling FastNet analysis with slope analysis to identify the number of reticulations in the output phylogeny. As shown in Figure B.2, the FastNet-inferred phylogeny with one reticulation was preferred to the FastNet-inferred phylogeny with two reticulations and the inferred tree of Salichos & Rokas (2013), which served as FastNet's guide phylogeny. The preferred phylogeny (see Figure 4.2) contained the inferred phylogeny of Salichos & Rokas (2013) where the tree edges were all subsumed by the preferred phylogeny. Furthermore, we identified gene flow between the clade containing *Kluyveromyces*, *Saccharomyces kluyveri*, and *Eremothecium gossypii* with the clade containing *Candida*. This reticulation is similar but more ancient compared to the one identified in the study of Yu & Nakhleh (2015).



(a) One-reticulation-node FastNet network

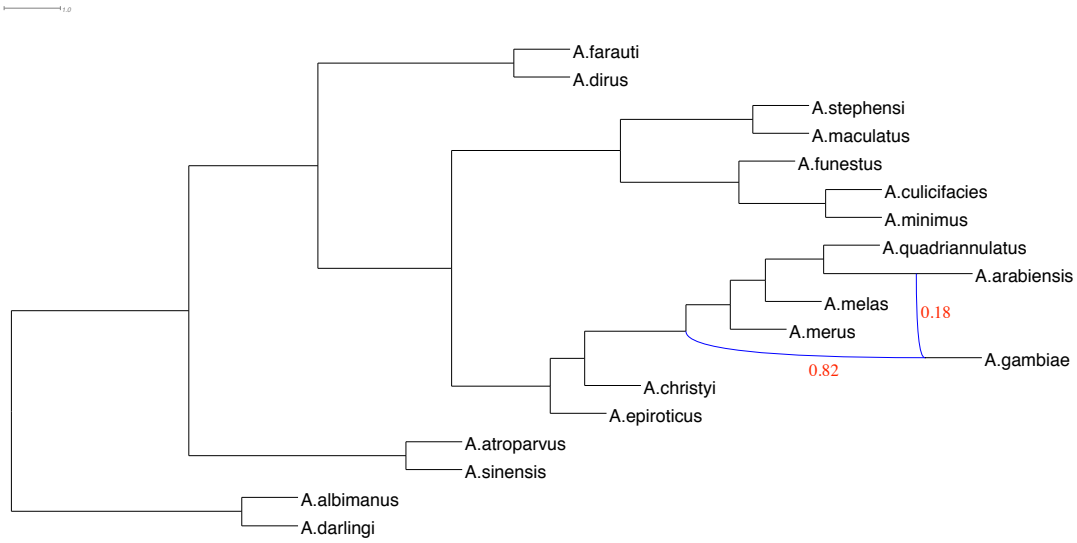


(b) Two-reticulation-node FastNet network

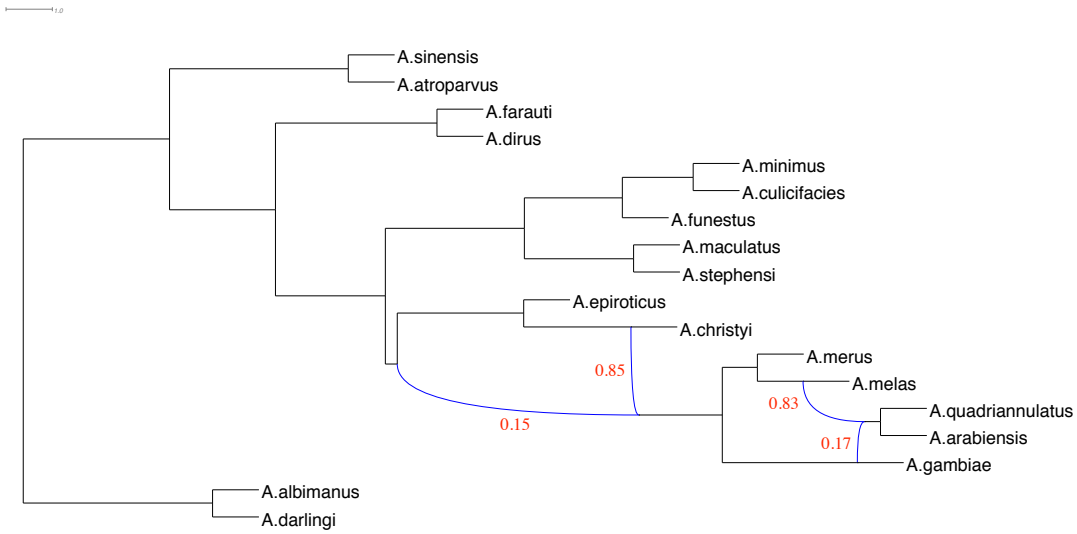
Figure 4.2: **The species phylogeny inferred by FastNet on the 1070-gene yeast dataset of Salichos & Rokas (2013) using (a) one reticulation and (b) two reticulations.** Reticulation edges are shown using blue curved lines. Inheritance probabilities are shown using red. Branches were scaled according to their branch lengths, which are measured in coalescent units. Dendroscope (Huson & Scornavacca, 2012) was used to plot the phylogenies. Using slope analysis as a model selection approach, the FastNet inferred network with one reticulation was preferred (see Figure B.2).

We further applied FastNet on the mosquito dataset of Neafsey *et al.* (2015) and estimated a phylogeny by coupling FastNet analysis with standard model selection approaches to identify the

number of reticulations in the output phylogeny. As shown in Figure B.3, the FastNet-inferred phylogeny with two reticulations was preferred to the FastNet-inferred phylogeny with one reticulation and the inferred tree of Neafsey *et al.* (2015), which served as FastNet's guide phylogeny. The preferred phylogeny (see Figure 4.3) contained the inferred phylogeny of Neafsey *et al.* (2015) where the tree edges were all subsumed by the preferred phylogeny. Furthermore, we identified an ancestral reticulation between the clade containing *A. quadriannulatus*, *A. arabiensis*, *A. gambiae*, *A. melas*, and *A. merus* with the other clade that has *A. christyi* and *A. epiroticus*. We identified another reticulation within the members of the *A. gambiae* complex which has been previously described in the study of Fontaine *et al.* (2015).



(a) One-reticulation-node FastNet network



(b) Two-reticulation-node FastNet network

Figure 4.3: **The species phylogeny inferred by FastNet on the mosquito dataset of Neafsey *et al.* (2015) using (a) one reticulation and (b) two reticulations.** Using slope analysis as a model selection approach, the FastNet inferred network with two reticulations was preferred (see Figure B.3). Figure layout and description are otherwise similar to Figure 4.2.

4.6 Discussion

In this chapter, we introduced FastNet, a new computational method for inferring phylogenetic networks from large-scale genomic sequence datasets. FastNet utilizes a divide-and-conquer algorithm to constrain two different aspects of scale: the number of taxa and evolutionary divergence.

We evaluated the performance boost of FastNet in comparison to state-of-the-art phylogenetic inference methods. We found that FastNet was comparable to or improved upon existing methods in terms of computational efficiency and topological accuracy. FastNet was up to an order of magnitude faster compared to other state-of-the-art phylogenetic network inference methods. Furthermore, FastNet's topological accuracy was typically better than all other methods in our study.

We explored the impact of multiple factors upon FastNet's topological accuracy and boosting effect. FastNet retained its accuracy and runtime performance advantage as we increased the number of taxa from 15 to 30, where the boosting effect tended to increase as we increased the dataset size. We observed a similar boosting effect outcome when we increased the number of reticulations from one to four, where a relatively greater performance impact on topological accuracy and runtime was observed. The performance boost of FastNet relative to the base methods was observed regardless of whether inferred or true gene trees were used as input, with the performance boosting effect greater when true gene trees were used. We further explored the performance of FastNet using model conditions including deep and non-deep gene flow. A relatively greater performance impact on topological accuracy and runtime was seen in the presence of deep gene flow compared to non-deep gene flow model conditions. However, for evolutionary scenarios involving either deep or non-deep gene flow, FastNet's accuracy was relatively robust to the dataset sizes explored in our study (in terms of the number of taxa).

We note that the base methods (i.e. MLE-length and MPL) were run in default mode. More intensive settings for each base method's optimization procedures may allow a tradeoff between topological accuracy and computational runtime. We stress that our goal was not to make specific recommendations about the nuances of running the base methods. Rather, FastNet's divide-and-conquer framework can be viewed as orthogonal to the specific algorithmic approaches utilized by the base method to be boosted. In this sense, improvements to the latter accrue to the former in a straightforward and modular manner.

We consider the procedures used in FastNet's inference of a guide phylogeny, subproblem de-

composition, and merge technique to be reasonable approaches, but more sophisticated alternatives can (and should) be proposed. Rather, we decided to focus our effort on the part of the phylogenetic network inference problem that we hypothesized to both have a first-order impact on inference accuracy and that substantially differentiates the problem from species tree inference: namely, gene flow detection and inferring phylogenies on subproblems. Despite these methodological limitations, we were able to obtain consistent improvements in topological accuracy and computational runtime when FastNet was used to boost the performance of a base method. Taken together, these results suggest that FastNet is robust to the choice of guide phylogeny, subproblem decomposition, merge technique, and gene tree error.

CHAPTER 5

COAL-MAP: MAPPING THE GENOMIC ARCHITECTURE OF QUANTITATIVE TRAITS WITH COMPLEX EVOLUTIONARY ORIGINS

There are many scenarios, spanning a diverse array of organisms across the Tree of Life, where complex evolution occurred and played a role in shaping regions of the genome encoding important adaptive traits. However, what remains poorly understood are the patterns and causes of phenotypic and genetic variation within and between populations. One way to address this question is through association mapping (AM), which pinpoints statistical associations between genotypes and phenotypes to uncover the underlying genetic factors contributing to variation in a trait of interest. One of the issues that need to be addressed when conducting an AM study is sample relatedness (Price *et al.*, 2010a; Devlin & Roeder, 1999), which induces spurious associations between the genotypic and trait data when the evolutionary relatedness between samples is not accounted for or modeled incorrectly. Current state-of-the-art methods address this issue by modeling the global sample relatedness measured across all genomic loci. Present-day populations arose through complex evolutionary histories that involved mutation, gene duplication and loss, recombination, ancestral polymorphism, natural selection, and gene flow. The complex interplay of the aforementioned evolutionary processes played a primary role in genome evolution and introduced different loci that exhibit local genealogical variation where gene trees differ from each other and the species phylogeny. Therefore, the assumption that there is a fixed sample relatedness across the genome could lead to spurious inferences. In this chapter, we mitigate this problem by developing Coal-Map (Hejase & Liu, 2016a), which models local and global sample relatedness using a generalized linear mixed model. We evaluate the performance of Coal-Map across a wide range of evolutionary scenarios and show that its performance is comparable or typically better than EIGENSTRAT. We further apply Coal-Map on an empirical dataset making use of hundreds of mouse genomes for which adaptive interspecific introgression has recently been described.

5.1 Linear mixed model

We consider the general problem of AM that involves identifying the underlying genetic factors contributing to variation in a phenotype of interest and accounts for local variation of sample relatedness across genomic sequences as well as global sample relatedness. We now describe our new AM method, which we refer to as Coal-Map. The input to Coal-Map is a multiple sequence alignment X containing n aligned sequences and k sites and a phenotypic vector y containing n quantitative traits. Each sequence in X has a corresponding phenotype such that the i^{th} sequence in X has a phenotypic value y_i where $1 \leq i \leq n$. An additional input to Coal-Map is a breakpoint vector b containing m breakpoints in ascending order. A local partition X_l is denoted as a region in the multiple sequence alignment containing sites falling in the closed interval b_l and b_{l+1}. The output of Coal-Map is a statistical score p_j, which measures the statistical association between x_j and y, for each site such that $1 \leq j \leq k$. The association score p_j is calculated using the following linear mixed model (following the notation of Zhou & Stephens (2012)):

$$\begin{aligned} y &= W_j\alpha + x_j\beta + \epsilon \\ \epsilon &\sim MVN_n(0, \tau^{-1}I_n) \end{aligned} \tag{5.1}$$

where y (n x 1) is the phenotypic vector, W (n x c) includes the fixed effects used to account for global sample relatedness and additional covariates used to model local sample relatedness, α (c x 1) encodes a coefficient for each covariate in W, x_j (n x 1) is the test SNP, β is the effect size of x_j, ϵ is a random effect that follows an n-dimensional multivariate normal distribution and is used to model unexplained variation in y, τ is the variance of residual errors, and I_n is an n by n identity matrix. The parameters $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\tau}$ are estimated using maximum likelihood where p_j for x_j is computed using a likelihood-ratio test between the fitted model against a null model with no SNP effect. The LMM takes the following form $\mathcal{L}(\lambda, \tau, \alpha, \beta) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|H| - \frac{1}{2}\tau(y - W\alpha - x_i\beta)^T H^{-1}(y - W\alpha - x_i\beta)$ where $H = \lambda K_{\text{global}} + I_n$ (reproduced from equation (3) in Zhou & Stephens (2012)).

5.2 Breakpoint inference

We infer local-phylogeny-switching breakpoints, where a pair of neighboring breakpoints represents a locus to be used in Coal-Map. Local coalescent-history mapping is an integral part of the design of Coal-Map. We utilize coalescent-based models to capture local genealogical variation alongside global sample relatedness. Different genealogies are observed for different genomic loci, depending on the specific coalescent history of each locus. Therefore, better coalescent-based modeling of the evolutionary origins of local genealogical variation may permit more accurate breakpoint inference. The input is X and the output is b, a breakpoint vector containing m breakpoints in ascending order. A local partition X_l is denoted as a genomic region falling in the closed interval b_l and b_{l+1}. Many methods have been proposed to reconstruct local genealogical histories such as PhyloNet-HMM (Liu *et al.*, 2014), RecHMM (Westesson & Holmes, 2009), and the Four-Gamete Test (Hudson & Kaplan, 1985). For the simulation study, we used the Four-Gamete Test, which identifies segregating sites that did not arise without either recombination or a repeat mutation, to infer local-phylogeny-switching breakpoints. The Four-Gamete Test is an appropriate choice to detect breakpoints due to the simplifying assumptions of our simulation study (infinite sites model, free recombination between loci, and complete linkage within each locus). For the empirical study, PhyloNet-HMM, which is a probabilistic inference method that uses the coalescent model, phylogenetic networks, and hidden Markov models to ascribe local genealogical variation to one of several evolutionary processes such as interspecific introgression, ILS, recombination, back mutation, and any combination thereof, was used to infer the loci.

5.3 Modeling sample relatedness

We applied PCA on X and used the top five principal components, which are ranked in descending order based on their eigenvalues, to model global sample relatedness. Let $w_1 ... w_5$ represent the top five global principal components. The local sample relatedness of X_l is represented using the top five principal components computed after applying PCA on X_l. Let $w_{l_1} ... w_{l_5}$ represent the top five local principal components of a genomic local region X_l. For each test locus x_j, one

could use a model that only accounts for global sample relatedness; use $w_1 ... w_5$ as covariates in W. An alternative way is to use a model that accounts for global and local sample relatedness; use $w_1 ... w_5$ as global covariates and $w_{l_1} ... w_{l_5}$ as local covariates in W. The model that only uses global sample relatedness is a nested version of the model that uses both local and global sample relatedness. A heuristic approach was used to select between the two aforementioned models. For each test locus x_j, a likelihood ratio test was applied for each model against a model with no SNP effect; the model with the smaller p-value was selected. The motivation behind using local sample relatedness is for the local principal components (covariates) to only contribute to the linear mixed model when the current local partition contains causal genomic sites.

5.4 Simulation study

We evaluated the performance of Coal-Map using neutral simulations of multi-locus sequence data where local genealogical variation was due to non-tree-like evolutionary scenarios. The program ms (Hudson, 2002), which generates samples under neutral models, was used to simulate coalescent histories and embedded gene trees under a coalescent model with an admixture process similar to the instantaneous unidirectional admixture described in the study of Durand *et al.* (2011). Under our coalescent model, admixture occurs at t_2 = 2.0 where a lineage from H comes from B according to probability γ and A according to probability 1 - γ (see Figure 5.1 and Appendix C.1.1). The choice of γ allows us to explore the impact of non-tree-like evolution (i.e. different rates of gene flow) in our simulation study, where we set γ to either 0.01, 0.1, 0.25, or 0.5. Each replicate dataset sampled 10 loci with a multiple sequence alignment containing 1000 individuals and a sequence length of 2500 bp (250 bp/locus). The local partition breakpoint vector b required as input to Coal-Map was inferred using the Four-Gamete Test.

We further evaluated the performance of Coal-Map using non-neutral simulations of multi-locus sequence data where local genealogical variation was due to non-tree-like evolutionary scenarios. The program msms (Ewing & Hermisson, 2010) was used to generate a forward-time simulation that explicitly modeled positive selection for the causal loci in the "neutral with non-tree-like

model phylogeny" model conditions (see Appendix C.1.2). The msms-based simulation utilized a sequence mutation model that allowed recurrent mutations between two alleles. Our forward-time coalescent simulation used a selection coefficient $s = 0.56$ which was based upon previously reported estimates from natural mouse populations that were involved in adaptive introgression linkage to emulate the genomic patterns of positive selection (Song *et al.*, 2011).

We used a trait model that is an extension to the one used in the study of Long & Langley (1999). Twenty causal SNPs were randomly selected from one, two, and ten loci which we refer to as 10% causal loci, 20% causal loci, and 100% causal loci, respectively. The following equation, which includes a genotypic and an environmental component, was used to simulate a quantitative trait for every sample:

$$y_i = \pi \sum_{j \in \delta} \frac{Q_{i,j}}{|\delta|} + (1 - \pi) N(0, 0.01) \tag{5.2}$$

where π represents the variation attributed to the genotypic component, which is the set of causal SNPs ($\pi = 0.5$ was used in this performance study). δ is the set of causal SNPs. $Q_{i,j}$ is the state of sample i at causal site j. Given that an infinite sites model is used, $Q_{i,j}$ can take a value of 0 or 1 representing homozygous non-mutant or mutant alleles, respectively. The environmental component is represented using a standard normal distribution with mean of 0 and standard deviation of 0.01. We randomly selected twenty causal sites, which have a minor allele frequency ranging between 0.1 and 0.3, to represent δ.

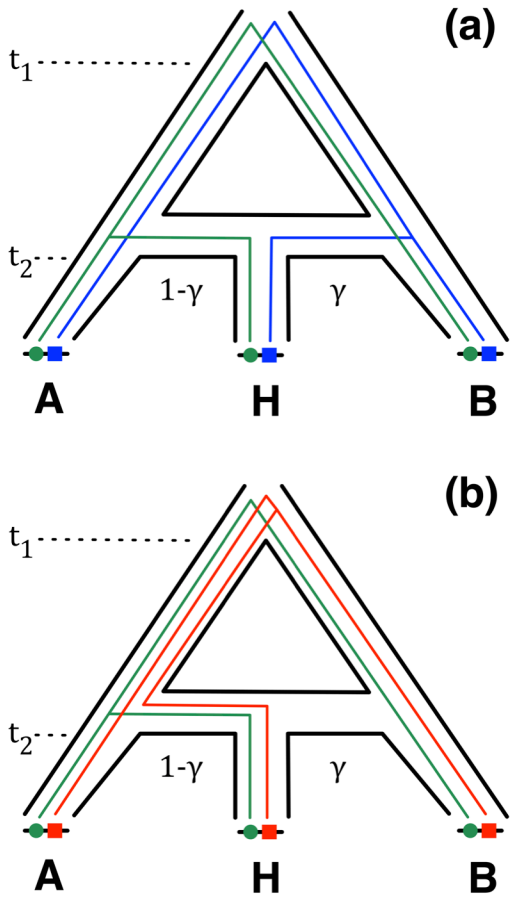


Figure 5.1: **Local genealogical variation due to gene flow and ILS.** (a) A phylogenetic network containing 3 populations. At t_2, a lineage ancestral to the gene sampled from population H coalesce with either lineages ancestral to populations A or B. The probabilities 1 - γ and γ determine the ancestral population that an H lineage comes from. At t_1, the lineages ancestral to populations A and B coalesce. Due to the presence of an admixed population H, two discordant gene trees exist (green and blue). (b) In the presence of ILS, we have a red gene tree that is discordant with the green gene tree. For the red gene tree, because the lineage ancestral to the gene sampled from population H fails to coalesce in the population ancestral to A, this lineage coalesces with the lineage ancestral to population B before coalescing with the lineage ancestral to population A.

5.5 Empirical study

Coal-Map was further applied on empirical mouse genomes of natural *Mus musculus* and *Mus spretus* populations. The empirical dataset has 744 haploid mouse genomes that are either wild

or wild-derived samples. The total length of each haploid genome is 414,376 SNPs genotyped across all samples. PhyloNet-HMM (Liu *et al.*, 2014), which is a probabilistic inference method that uses the coalescent model, phylogenetic networks, and hidden Markov models to ascribe local genealogical variation to one of several evolutionary processes: interspecific introgression, ILS, recombination, back mutation, and any combination thereof, was used to infer the introgressed regions. We used a synthetic phenotypic trait for each sample based on the above trait model. The effect size and minor allele frequency were specified as input parameters for the trait model. We focused on mouse chromosomes 7, 15, and 17, which are the only chromosomes that consist of at least two introgressed regions with a total length of at least 100 bp. For each chromosome, 10% causal loci and 20% causal loci trait simulations were repeated to generate twenty replicate datasets. The empirical genomic sequence data, synthetic trait data, and local partition breakpoints were provided to Coal-Map as inputs.

5.6 Results

The simulation study included model conditions across a range of genetic architectures that included neutral or non-neutral with non-tree-like model phylogenies with a range of hybridization frequencies (γ = 0.01, 0.1, 0.25, and 0.5), and trait models including 10% causal loci, 20% causal loci, and 100% causal loci. Figure 5.2 compares the performance of Coal-Map and EIGENSTRAT using a ROC curve on model conditions incorporating positive selection with hybridization frequency γ of 0.5 and trait simulations that include 10% causal loci, 20% causal loci, and 100% causal loci. Coal-Map offered equal or better power and comparable type I error to EIGENSTRAT with a performance advantage in terms of AUROC of 0.058, 0.038, and 0.029 for 10% causal loci, 20% causal loci, and 100% causal loci model conditions, respectively. This performance advantage was significant for the 10% causal loci, 20% causal loci, and 100% causal loci model conditions using the approach of DeLong *et al.*, with p-values of 2 x 10^{-14}, 5 x 10^{-5}, and 0.006, respectively. As measured by AUROC, the performance advantage of Coal-Map over EIGENSTRAT was largest on the 10% causal loci model condition and smaller as more loci contributed causal SNPs.

This is also consistent with each method's TPR at a typical FPR. At an FPR of 0.05, Coal-Map's TPR improved upon EIGENSTRAT's by 0.152 and 0.054 on the 10% causal loci and 20% causal loci model conditions, respectively; on the 100% causal loci model condition, the TPR difference between the two methods was less than 0.039.

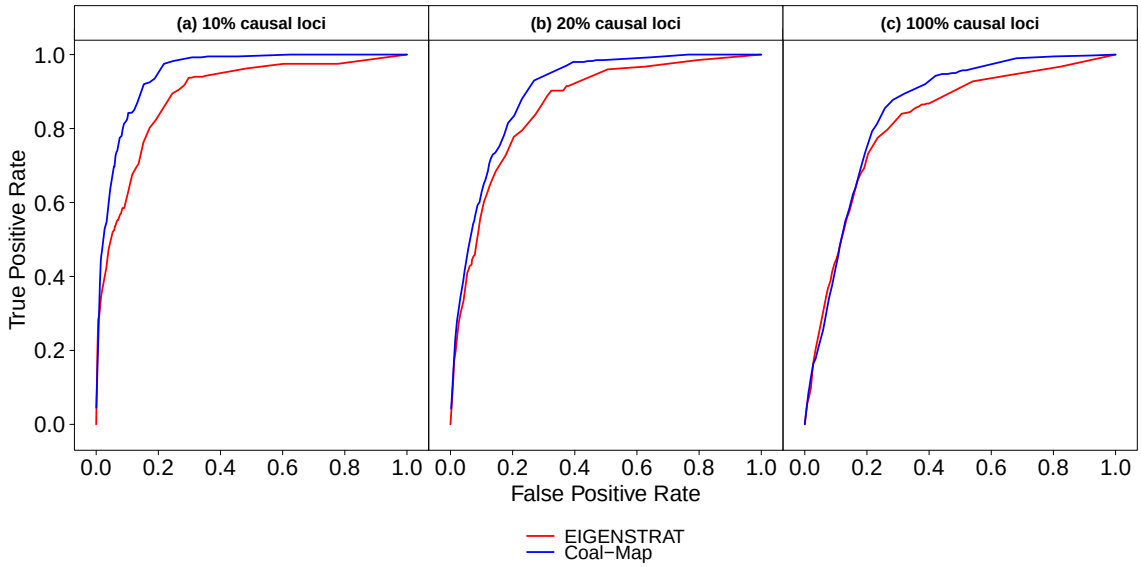


Figure 5.2: **For simulations involving adaptive gene flow (hybridization frequency γ = 0.5 and selection coefficient s = 0.56), Coal-Map has an equal or better power and comparable type I error to EIGENSTRAT.** The ROC curve shows the relationship between FPR and TPR. The blue and red ROC curves report the performance of Coal-Map and EIGENSTRAT, respectively.

Coal-Map's performance advantage over EIGENSTRAT was similarly observed on model conditions that involved a range of hybridization frequencies from γ = 0.01 to 0.5 and intra-locus linkage that did not incorporate positive selection. Figure 5.3 compares the performance of Coal-Map and EIGENSTRAT using a ROC curve on model conditions with the highest level of gene flow (γ = 0.5). Across the different trait architectures, Coal-Map offered equal or better power and comparable type I error to EIGENSTRAT with a performance advantage in terms of AUROC of 0.068, 0.038, and 0.01 on the 10% causal loci, 20% causal loci, and 100% causal loci model conditions, respectively. The performance advantage was significant using the statistical test of DeLong *et al.* for the 10% causal loci and 20% causal loci model conditions with p-value $< 10^{-5}$ but not for the 100% causal loci model condition (p-value = 0.16). As measured by AUROC,

the performance advantage of Coal-Map over EIGENSTRAT was largest on the 10% causal loci, with TPR improving upon EIGENSTRAT by 0.111 at an FPR of 0.05, and smaller as more loci contributed causal SNPs. We further compared the reported p-values at causal SNPs in Figure 5.4 using a cumulative histogram for Coal-Map and EIGENSTRAT where Coal-Map reported smaller p-values compared to EIGENSTRAT.

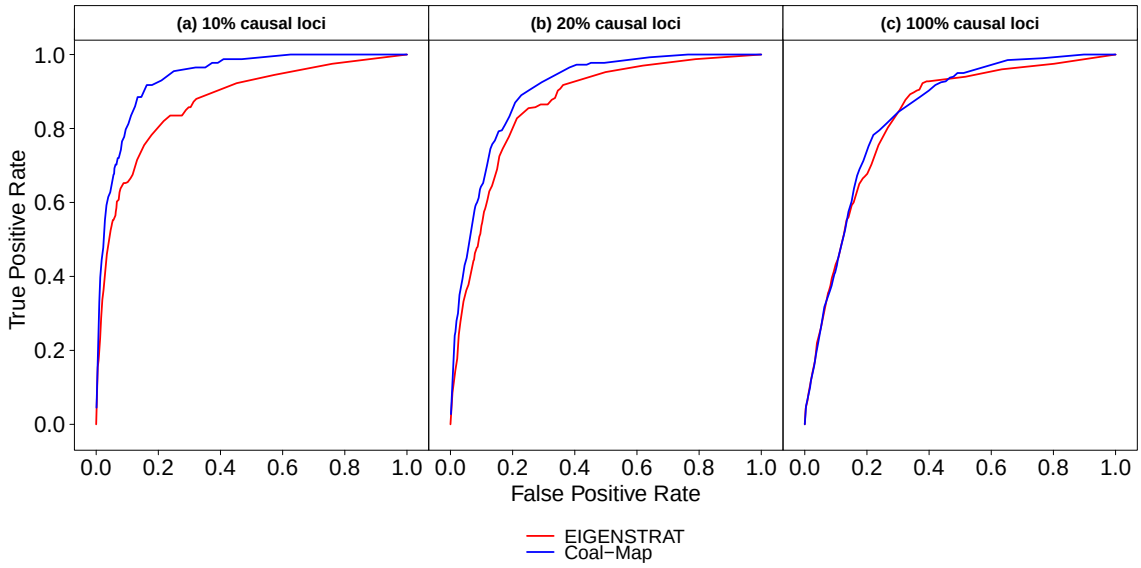


Figure 5.3: **For simulations involving neutral gene flow (hybridization frequency γ = 0.5), Coal-Map has an equal or better power and comparable type I error to EIGENSTRAT.** Figure layout and description are otherwise similar to Figure 5.2.

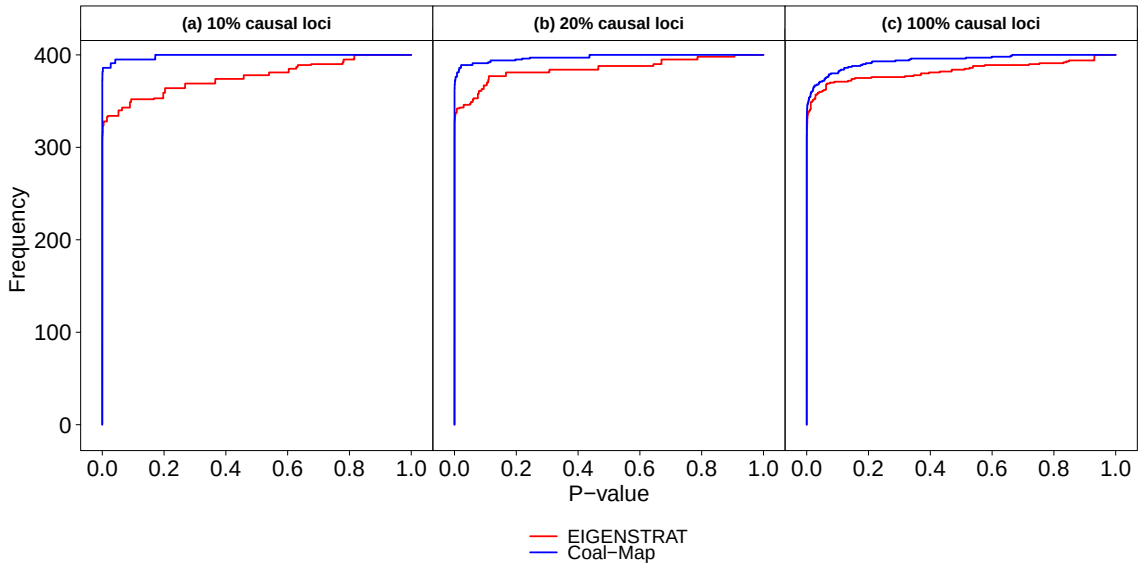


Figure 5.4: **For simulations involving neutral gene flow (hybridization frequency γ = 0.5), the cumulative histogram of p-values at causal sites for Coal-Map and EIGENSTRAT are reported.** The x-axis reports the test statistic (i.e. p-value) while the y-axis reports the cumulative frequency over twenty replicates for each model condition. Results are shown for three model conditions: (a) 10% causal loci, (b) 20% causal loci, and (c) 100% causal loci.

We observed that the performance advantage of Coal-Map over EIGENSTRAT did not diminish and in fact remained roughly the same as we examined smaller levels of gene flow ($\gamma = 0.01$, 0.1, and 0.25) on 10% causal loci and 20% causal loci model conditions (see Figures C.1 and C.2). Table 5.1 shows that the performance advantage of Coal-Map over EIGENSTRAT remained significant using the statistical test of DeLong *et al.* across the different hybridization frequencies. Thus, on the model condition with negligible gene flow (i.e. $\gamma = 0.01$), virtually all local genealogical variation was due to ILS. Using the statistical test of DeLong *et al.*, the difference in AUROC between Coal-Map and EIGENSTRAT was significant across all trait architectures. This is consistent with the difference in the TPR between Coal-Map and EIGENSTRAT at different FPR values. At an FPR value of 0.05, the TPR of Coal-Map was greater than EIGENSTRAT by 0.111 and 0.084 for the 10% causal loci and 20% causal loci model conditions, respectively. As for the 100% causal loci model condition, the difference in TPR between Coal-Map and EIGENSTRAT was less than 0.002.

10% causal loci			
Hybridization frequency γ	**Coal-Map**	**EIGENSTRAT**	**Corrected q value**
0.5	0.938	0.870	$< 10^{-5}$
0.25	0.935	0.882	$< 10^{-5}$
0.1	0.928	0.890	$< 10^{-5}$
0.01	0.917	0.845	$< 10^{-5}$

20% causal loci			
Hybridization frequency γ	**Coal-Map**	**EIGENSTRAT**	**Corrected q value**
0.5	0.898	0.860	$< 10^{-5}$
0.25	0.911	0.860	$< 10^{-5}$
0.1	0.881	0.843	$< 10^{-5}$
0.01	0.879	0.834	$< 10^{-5}$

100% causal loci			
Hybridization frequency γ	**Coal-Map**	**EIGENSTRAT**	**Corrected q value**
0.5	0.836	0.826	0.16
0.25	0.842	0.808	0.001
0.1	0.854	0.842	0.093
0.01	0.847	0.817	0.002

Table 5.1: **The performance of Coal-Map and EIGENSTRAT based on AUROC is compared across model conditions involving neutral evolution with ILS and a wide range of gene flow.** On 10% causal loci and 20% causal loci model conditions, Coal-Map has AUROC that is significantly better than EIGENSTRAT, using the statistical test of DeLong *et al.* with Benjamini-Hochberg correction (Benjamini & Hochberg, 1995), across different hybridization frequencies ranging from a relatively large level of gene flow ($\gamma = 0.5$) to negligible amounts of gene flow ($\gamma = 0.01$). On 100% causal loci model conditions, Coal-Map had a diminished performance advantage in terms of AUROC, and the improvement was either weakly significant or not significant (under the same test).

We examined the sensitivity of Coal-Map to the number of covariates (five and twenty). The performance improvement of Coal-Map, using five covariates to represent global and local sample relatedness, was significantly greater than EIGENSTRAT in terms of AUROC for the 10% causal loci (p-value $< 10^{-5}$) and 20% causal loci (p-value $< 10^{-5}$) model conditions but not for the 100% causal loci model condition (p-value = 0.26). At an FPR of 0.05, Coal-Map's TPR improved upon EIGENSTRAT's by 0.230 and 0.079 on the 10% causal loci and 20% causal loci model

conditions, respectively; on the 100% causal loci model condition, the TPR difference between the two methods was less than 0.001. We observed consistent results when the sensitivity of Coal-Map was explored using twenty covariates to represent global and local sample relatedness. Overall, we found that Coal-Map's performance was robust to the number of covariates used to represent sample relatedness. We further compared the performance of EIGENSTRAT to an approach that only accounts for local sample relatedness with out explicitly modeling global sample relatedness. We observed that modeling local sample relatedness alone resulted in a marked decrease in performance with the resulting power and false positive rates being worse than EIGENSTRAT (see Figure 5.5). We explored a trait model that only accounts for the genotypic component while lacking a random effect due to environment. We observed (see Figure 5.6) that Coal-Map performs better than EIGENSTRAT using 10% causal loci and 20% causal loci model conditions with an AUROC improvement of 0.039 and 0.011, respectively, with this performance improvement being significant (p-value $< 10^{-5}$) for both model conditions. Overall, we found that Coal-Map's performance advantage over EIGENSTRAT was greater on model conditions lacking a random effect due to environment compared to model conditions that included both genotypic and environmental effects. We explored other model selection approaches (i.e. forward selection). The forward selection approach is conservative and biased towards selecting the model with fewer parameters. In our simulation study, the forward selection approach was biased towards selecting the model that only accounts for global sample relatedness. Figure 5.7 shows the performance of Coal-Map and EIGENSTRAT using ROC curves. Using the statistical test of DeLong *et al.*, Coal-Map's performance improvement was significantly greater than EIGENSTRAT in terms of AUROC for the 10% causal loci (p-value $< 10^{-5}$) and 20% causal loci (p-value = 0.0002) model conditions but not for the 100% causal loci model condition (p-value = 0.83).

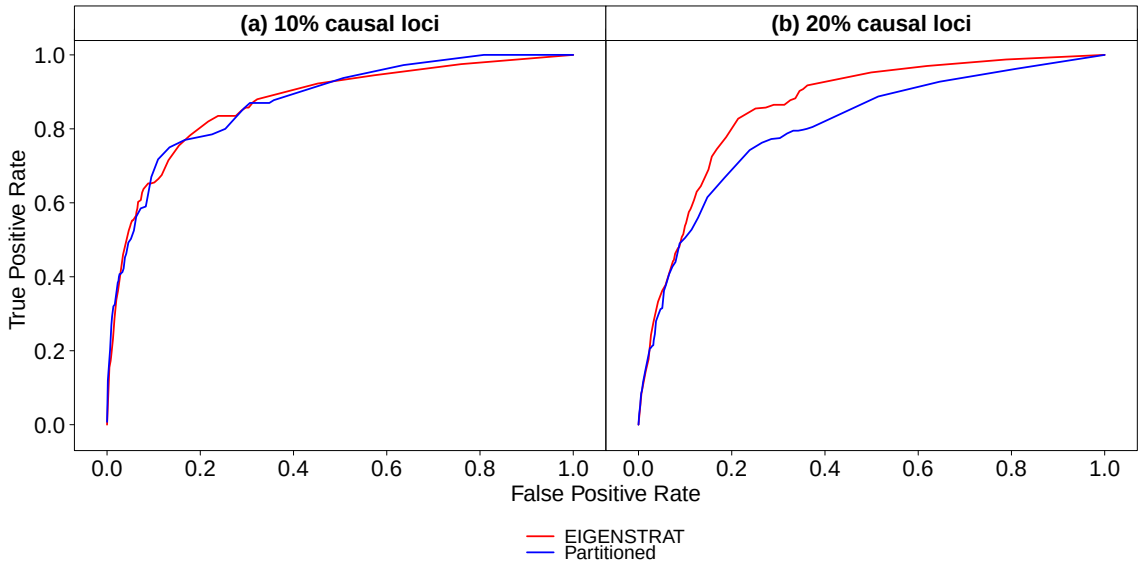


Figure 5.5: **For simulations involving neutral gene flow (hybridization frequency γ = 0.5), EIGENSTRAT has an equal or better power and comparable type I error compared to a partitioned approach that only accounts for local sample relatedness.** Results are shown for two model conditions: (a) 10% causal loci and (b) 20% causal loci. Figure layout and description are otherwise similar to Figure 5.2.

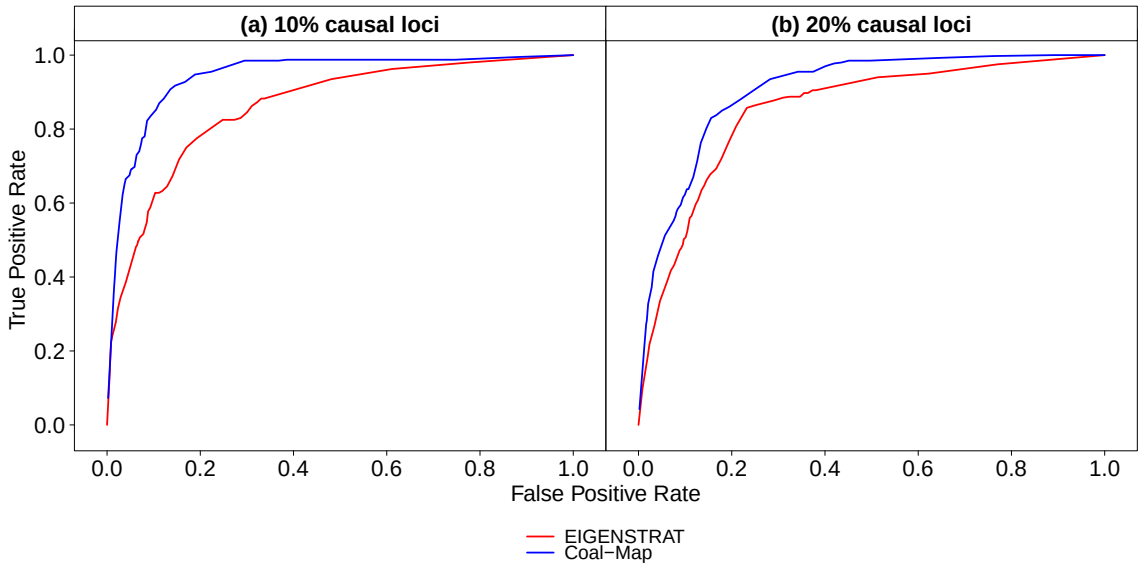


Figure 5.6: **For simulations involving neutral gene flow (hybridization frequency γ = 0.5) with the trait having no environmental effect (proportion of trait variation contributed by the genotypic effect π = 1), Coal-Map has an equal or better power and comparable type I error to EIGENSTRAT.** Results are shown for two model conditions: (a) 10% causal loci and (b) 20% causal loci. Figure layout and description are otherwise similar to Figure 5.2.

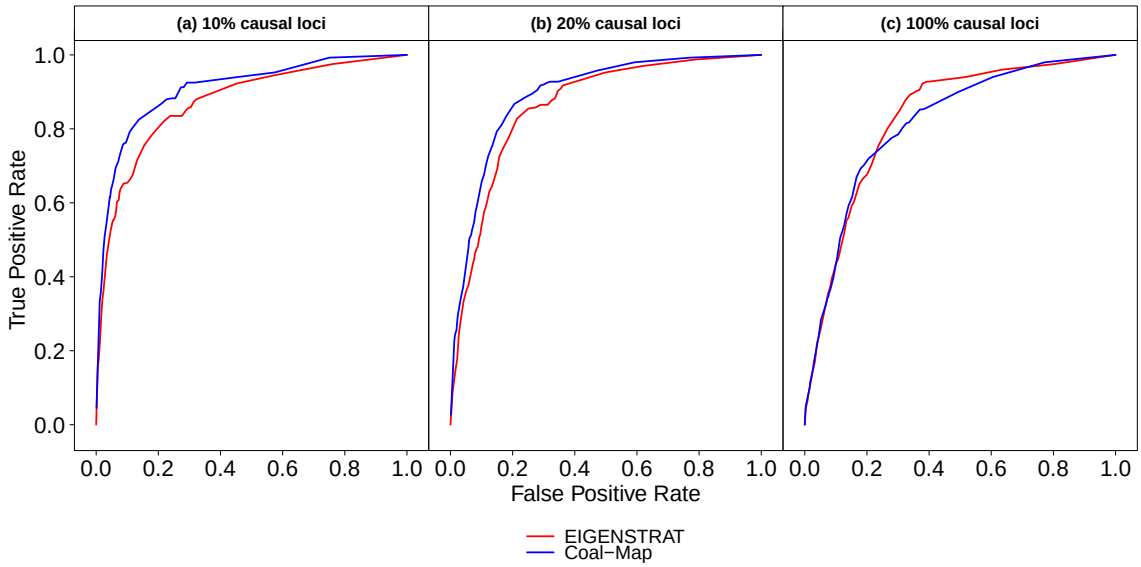
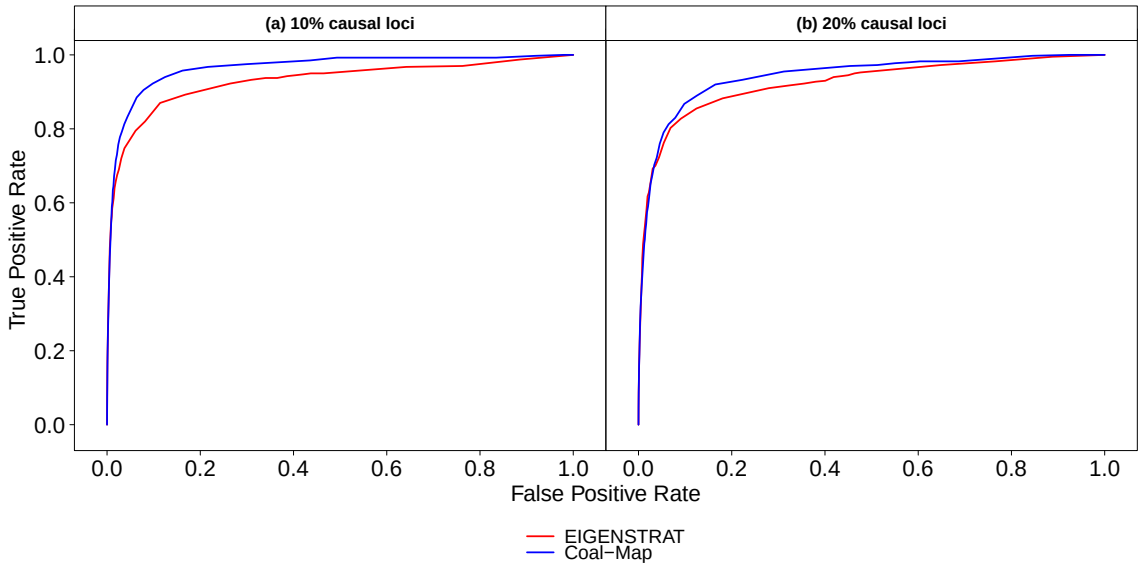


Figure 5.7: **For simulations involving neutral gene flow (hybridization frequency $\gamma = 0.5$), Coal-Map, using forward selection as a model selection approach, has an equal or better power and comparable type I error to EIGENSTRAT.** Figure layout and description are otherwise similar to Figure 5.2.

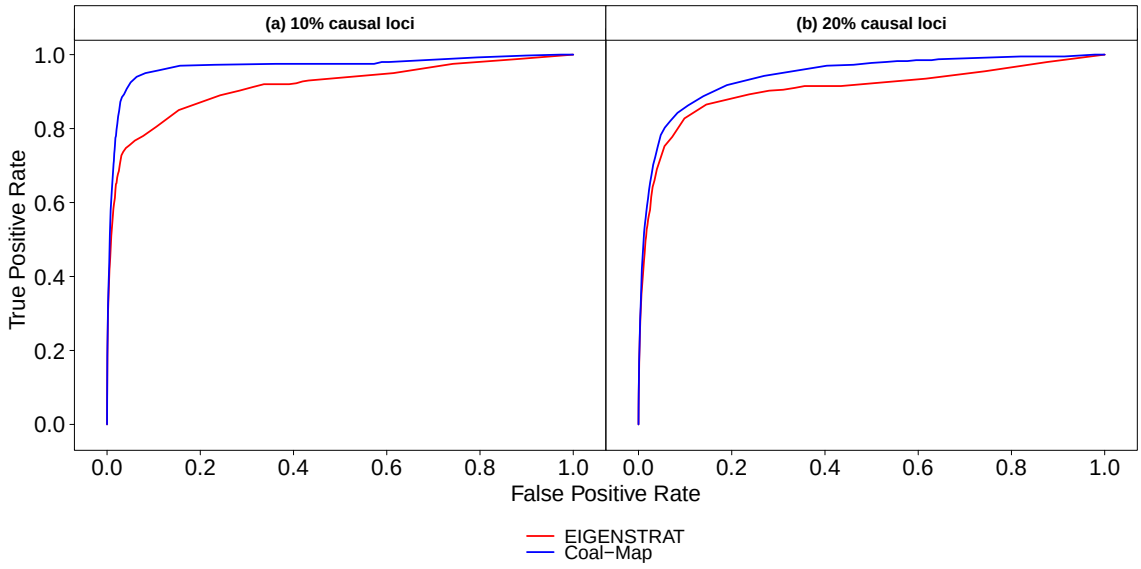
We further compared the performance of Coal-Map and EIGENSTRAT on chromosomes 7, 15, and 17 of empirical mouse genomes. For chromosomes 7 and 17, Coal-Map offered better power than EIGENSTRAT using 10% causal loci and 20% causal loci model conditions, respectively (Figure 5.8). The AUROC improvement of Coal-Map over EIGENSTRAT was significant (see Table 5.2) for both trait architectures, with a TPR improvement at an FPR value of 0.05 for chromosomes 7 and 17 of 0.076 and 0.166 for 10% causal loci, and 0.030 and 0.057 for 20% causal loci, respectively. We further explored the reported p-values at causal SNPs. Overall, Coal-Map reported smaller p-values at causal SNPs compared to EIGENSTRAT (Figure 5.9). We note that chromosomes 7 and 17 exhibited the greatest amount of introgression in our study. In contrast, chromosome 15 had the fewest number of introgressed sites in our study, with the AUROC improvement of Coal-Map over EIGENSTRAT being weakly significant on the 10% causal loci model condition and not significant on the 20% causal loci model condition (see Figures C.3 and C.4).

Chromosome	10% causal loci			20% causal loci		
	Coal-Map	EIGENSTRAT	Corrected q value	Coal-Map	EIGENSTRAT	Corrected q value
7	0.964	0.928	$< 10^{-5}$	0.942	0.923	0.003
15	0.940	0.922	0.014	0.917	0.919	0.587
17	0.968	0.914	$< 10^{-5}$	0.942	0.904	1.6×10^{-5}

Table 5.2: **The performance of Coal-Map and EIGENSTRAT based on AUROC is compared using empirical mouse chromosomes and simulated traits.** On the two mouse chromosomes with the greatest number of introgressed sites in our study – chromosomes 7 and 17 – Coal-Map's performance was significantly better than EIGENSTRAT for both 10% causal loci and 20% causal loci traits using the statistical test of DeLong *et al.* with Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). We observed a reduced performance improvement on chromosome 15, which had relatively fewer introgressed sites: the improvement was weakly significant for 10% causal loci traits and not significant for 20% causal loci traits (using the same test).

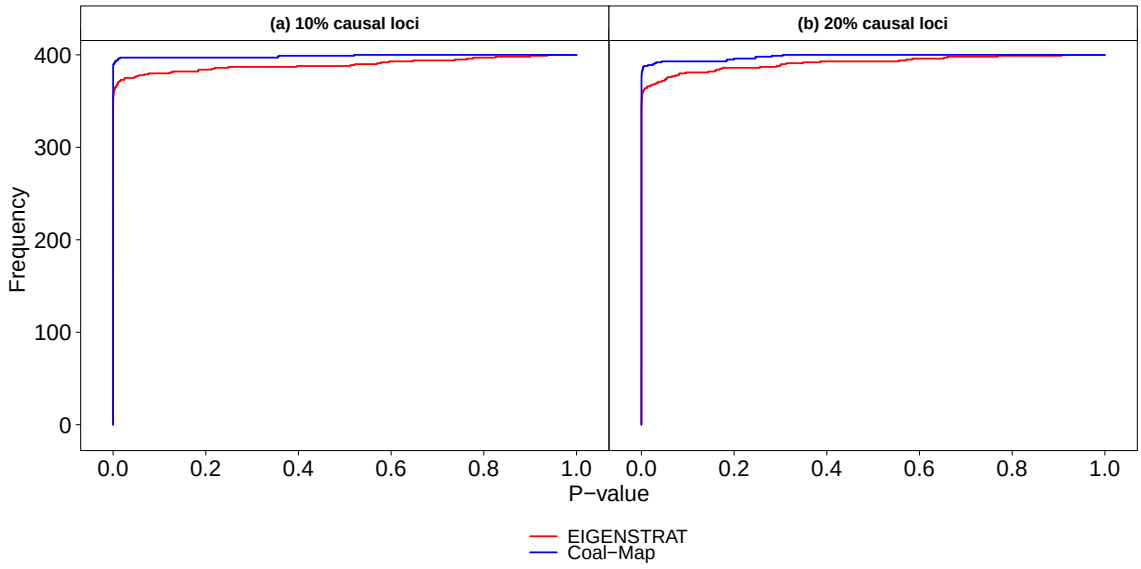


(a) Chromosome 7

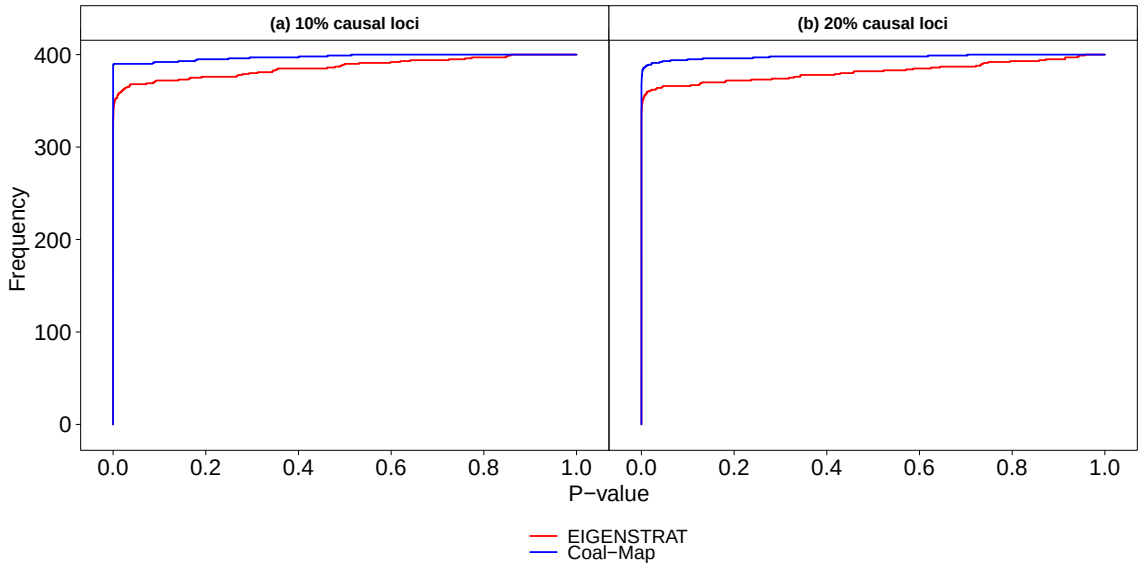


(b) Chromosome 17

Figure 5.8: **Using empirical genomic data from mouse chromosomes 7 and 17, Coal-Map has an equal or better power and comparable type I error to EIGENSTRAT.** Results are shown for two model conditions: (a) 10% causal loci and (b) 20% causal loci. Figure layout and description are otherwise similar to Figure 5.2.



(a) Chromosome 7



(b) Chromosome 17

Figure 5.9: **Using empirical genomic data from mouse chromosomes 7 and 17, the cumulative histogram of p-values at causal sites for Coal-Map and EIGENSTRAT are reported.** Results are shown for two model conditions: (a) 10% causal loci and (b) 20% causal loci. Figure layout and description are otherwise similar to Figure 5.4.

5.7 Discussion

In this chapter, we introduced Coal-Map, a new AM method, that explicitly models local and global sample relatedness and found that Coal-Map's performance is comparable or better

than EIGENSTRAT in terms of statistical power and FPR. Additionally, Coal-Map's performance advantage was greatest on model conditions that most closely resembled empirically observed scenarios of adaptive introgression.

Coal-Map was evaluated using simulated and empirical data on model conditions reflecting a wide range of evolutionary scenarios. We showed that Coal-Map had similar or better performance compared to EIGENSTRAT, a state-of-the-art method in terms of its popularity, power, and type I error control. The number of causal loci is a factor that impacts the performance of Coal-Map. As reported, Coal-Map performs best using 10% or 20% causal loci and as the number of causal loci increases to cover the entire genome (i.e. 100% causal loci), the performance of Coal-Map degrades and becomes comparable to EIGENSTRAT. We attribute this decrease in performance of Coal-Map on the 100% causal loci model condition due to the increase in local genealogical variation across the different causal loci, which removes the local sample relatedness effect on the phenotype; therefore, global sample relatedness will dominate. Based on previous studies of adaptive introgression in mice (Liu *et al.*, 2015; Consortium, 2012), we hypothesize that 10% causal loci and 20% causal loci model conditions are most relevant to an empirical study.

We compared the performance of Coal-Map and EIGENSTRAT across a wide range of gene flow scenarios. The simulations included model conditions with minimal gene flow where local genealogical variation was mainly due to ILS. Additionally, the simulations included non-tree like evolutionary scenarios that incorporated positive selection. As seen in the results, Coal-Map outperformed EIGENSTRAT across the different model conditions. We observed that a model that only accounts for local sample relatedness resulted in reduced power and higher FPR compared to Coal-Map, which accounts for both local and global sample relatedness. Our finding is consistent with the findings of Shriner (2011). The intuitive explanation is that local sample relatedness in the current partition (i.e. the partition enclosing the test SNP) should be modeled when the current partition contains causal SNPs, but otherwise not. The choice of the number of covariates used to represent sample relatedness in Coal-Map was based upon a previous algorithmic design study examining the use of fixed effects models for AM (Price *et al.*, 2006). To further explore

the ramifications of this choice, we conducted an algorithmic design experiment to explore the impact of the number of covariates used in Coal-Map's model upon its performance. We found that Coal-Map's performance was robust to this design choice. Furthermore, Coal-Map's performance advantage over EIGENSTRAT was retained across different levels of environmental contribution to traits. A larger performance improvement was seen on model conditions with only a genotypic contribution to traits, which we ascribe to the lack of sample relatedness inherent in the additive environmental noise. The empirical study examined in this work involved even a wider range of evolutionary processes, including recombination, where the results were consistent with the simulation study. Overall, we observed that Coal-Map's performance was comparable or better than EIGENSTRAT.

CHAPTER 6

COAL-MINER: A STATISTICAL METHOD FOR GWA STUDIES OF QUANTITATIVE TRAITS WITH COMPLEX EVOLUTIONARY ORIGINS

In this chapter, we adopt an evolutionary approach to model and account for sample relatedness. Sample relatedness can be a confounding factor in AM studies if modeled incorrectly or not modeled at all (Devlin & Roeder, 1999; Price *et al.*, 2010a; Marchini *et al.*, 2004; Voight & Pritchard, 2005). Inferring the genealogical history of each genomic locus provides a way to account for sample relatedness, which is crucial for AM. We introduce Coal-Miner (Hejase *et al.*, 2017a); a method that captures local genealogical variation using a generalized linear mixed model. In the previous chapter, we introduced Coal-Map, which evaluates the relationship between genotype and phenotype while accounting for and modeling local and global sample relatedness. We note here some improvements which we address in this follow-up method. Coal-Map focuses in its entirety on one cause of local genealogical variation, which is introgression/gene flow. In Coal-Miner, we conduct simulation and empirical studies on a broader set of applications, beyond just mapping introgressed traits, such as gene flow due to genetic admixture, ILS, recombination, and natural selection as the primary causes of local genealogical variation. In this work, we introduce the use of search techniques to make an explicit map and detect candidate loci (i.e. loci that contain putatively associated sites), which provides assistance in downstream stages of the Coal-Miner pipeline (i.e. traditional GWA single-marker test). This detection technique was inspired from the work of Speed & Balding (2014) on a different computational problem. Searching for candidate loci assists in improving the modeling of the global and local sample relatedness. The main motivation behind searching for candidate loci is that we only want to select loci to contribute to the linear mixed model when they also contribute to the phenotype. Global and local sample relatedness could be modeled using a combination of random and fixed effects instead of relying only on fixed effects (i.e. Coal-Map utilizes a fixed effects model). Recent methods that model sample relatedness using a combination of fixed and random effects have shown through simulation studies that mixed models

perform better in terms of modeling sample relatedness than fixed effects models (Price *et al.*, 2010a; Zhang *et al.*, 2010). We extend Coal-Map to model the local and global sample relatedness as a function of fixed and random effects. We show through a simulation study that examines models with different evolutionary scenarios that Coal-Miner's performance surpasses state-of-the-art AM methods such as Coal-Map, GEMMA, and EIGENSTRAT in terms of statistical power and FPR. We further apply Coal-Miner on empirical datasets that include bacteria in the family *Burkholderiaceae*, the flowering plant *Arabidopsis thaliana*, and the butterfly *Heliconius erato* to identify known and novel genes associated with different traits of interest. The Coal-Miner pipeline involves three stages. In stage one of Coal-Miner, we infer local-phylogeny-switching breakpoints, where a pair of neighboring breakpoints describes a locus to be used in downstream stages. In stage two of Coal-Miner, we detect candidate loci, which are loci that likely contain putatively associated sites. Finally, in stage three of Coal-Miner, we perform a traditional GWA single-marker test using a generalized linear mixed model to pinpoint associated sites.

6.1 Method

In this work, we consider the general problem of AM which involves identifying the underlying genetic factors contributing to variation in a phenotype of interest. We now describe our new AM method, which we refer to as Coal-Miner. The input to Coal-Miner is a multiple sequence alignment X and a phenotypic vector y. Let X be a multiple sequence alignment containing n aligned sequences and k sites. y is the phenotypic vector containing n quantitative traits. Each sequence in X has a corresponding phenotype such that the i^{th} sequence in X has a phenotypic value y_i where $1 \leq i \leq n$. The output of Coal-Miner is a statistical score p_j for each site such that $j = 1...k$. p_j measures the statistical association between x_j and y. Coal-Miner utilizes a LMM to evaluate the relationship between y and x:

$$\begin{aligned} y &= (1-r)W_j\alpha + x_j\beta + u + \epsilon \\ u &\sim MVN_n(0, r\lambda\tau^{-1}K_{\text{global}}) \\ \epsilon &\sim MVN_n(0, \tau^{-1}I_n) \end{aligned} \tag{6.1}$$

where y is the phenotype vector, W_j includes the fixed effects which are used to model local sample relatedness, α encodes the coefficients of the covariates located in W_j, x_j is the test locus, β is the effect size of x_j, r is the proportion of candidate loci, u is a random effect that follows an n-dimensional multivariate normal distribution, K_{global} is a kinship matrix which is represented as a pairwise genotypic similarity between individuals and is used to model global sample relatedness, λ is the ratio between two variance components (genetic and environmental effects), τ is the variance of residual errors, ϵ is a random effect that follows an n-dimensional multivariate normal distribution and is used to model any unexplained variation in y, and I_n is an n by n identity matrix. The parameters $\hat{\alpha}$, $\hat{\beta}$, $\hat{\tau}$, and $\hat{\lambda}$ are estimated using maximum likelihood where p_j for x_j is computed using likelihood-ratio test between the fitted model against a null model with no SNP effect. The following three stages describes the details of the Coal-Miner pipeline (see pseudocode in Algorithm 6 for more details):

Algorithm 6 Coal-Miner Design

variable n ▷ Number of samples
variable k ▷ Number of sites
variable X ▷ An n by k multi-locus sequence data matrix
variable y ▷ An n by 1 phenotype vector
variable l^* ▷ Number of candidate loci

procedure LocalPhylogenySwitchingBreakpointsInference(X, Ψ)
    X_i = Partition(X, Ψ) ▷ Partition the sites in X into loci X_i using method Ψ, where $\cup X_i = X$ for $1 \leq i \leq l$
    b = GetBreakpoints(X_i) ▷ Get the local partition breakpoints in ascending order
    return b ▷ Return a local-phylogeny-switching breakpoint vector

procedure IdentifyCandidateLoci(X, y, l^*, b)
    $\{X_i\}$ = GetPartitionedLoci(X, b)
    l = GetNumberOfLoci(b)
    for $i = 1$ **to** l **do**
        w_i = PCA(X_i) ▷ Get the top five principal components from PCA applied to X_i
        $W = w_i$ ▷ Assign the five covariates in w_i to W
        K_{global} = ComputeGlobalKinship(X_i)
        $H = \lambda K_{\text{global}} + I_n$
        $\mathcal{L}(\lambda, \tau, \alpha)_i = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi)$ ▷ LMM log-likelihood optimization criterion
        $-\frac{1}{2}\log|H| - \frac{1}{2}\tau(y - W\alpha)^T H^{-1}(y - W\alpha)$
    $\{X_j^*\}$ = GetCandidateLoci($\{\mathcal{L}(\lambda, \tau, \alpha)_i\}$, $\{X_i\}$, l^*) ▷ Get the set of candidate loci consisting of the top l^* loci based upon the fitted LMM likelihood
    return $\{X_j^*\}$ ▷ Returns a set of candidate loci

procedure SnpBasedAssociationTesting(X, y, l^*)
    b = LocalPhylogenySwitchingBreakpointsInference(X, Ψ)
    $\{X_j^*\}$ = IdentifyCandidateLoci(X, y, l^*, b)
    for $i = 1$ **to** k **do** ▷ Test each SNP $x_i \in X$ for association under Coal-Miner's LMM
        if $x_i \in \{X_j^*\}$ **then**
            w_j = PCA(X_j^*) ▷ Gets the top principal component from PCA applied to X_j^*
            $W = w_j$ ▷ Assign the covariate in w_j to W
        K_{global} = ComputeGlobalKinship(X)
        $H = \lambda K_{\text{global}} + I_n$
        $\mathcal{L}(\lambda, \tau, \alpha, \beta) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi)$ ▷ LMM log-likelihood optimization criterion
        $-\frac{1}{2}\log|H| - \frac{1}{2}\tau(y - W\alpha - x_i\beta)^T H^{-1}(y - W\alpha - x_i\beta)$
        p_i = LRT($\mathcal{L}_{fitted}$, $\mathcal{L}_{null}$) ▷ Association score for x_i using a likelihood ratio test
        AppendResultToVectorP(p_i, p)
    return p ▷ Returns a vector of p values

6.1.1 Stage one: Infer local-genealogy-switching breakpoints under extended coalescent model

The input to stage one is a multiple sequence alignment X and a breakpoint inference method Ψ. We use Ψ to infer breakpoints on X, where the sites in X are partitioned into loci X_i such that $\cup X_i = X$ for $1 \leq i \leq l$, and l is the total number of loci inferred by Ψ. The general approach to address this computational problem is to infer local coalescent histories under an appropriate multi-species extension of the coalescent model, and then to assign breakpoints based upon gene tree discordance. Each pair of neighboring breakpoints delineates a locus for use in downstream stages of the Coal-Miner pipeline. The specific choice of method Ψ depends upon the relevant evolutionary processes involved in multi-locus sequence evolution, particularly regarding the source(s) of local genealogical discordance. We explored two choices for Ψ. Based on the simplifying assumptions of our simulation study (infinite sites model, free recombination between loci, and complete linkage within each locus), we used the Four-Gamete Test (Hudson & Kaplan, 1985) to infer local-phylogeny-switching breakpoints. For the empirical study analyses, we did not make use of the infinite sites model and its assumptions about sequence evolution. Furthermore, multiple evolutionary processes were known to be involved in multi-locus sequence evolution, including genetic drift/ILS, recombination/gene conversion, gene flow/horizontal gene transfer (HGT), and natural selection. We used RecHMM (Westesson & Holmes, 2009) to infer recombination-free genomic regions. Rec-HMM performs fixed-species-phylogeny inference of local genealogies under a statistical model that combines a finite-sites substitution model and a hidden Markov model which is meant to capture intra-sequence dependence (such as arises from recombination). The output of stage one is a local-phylogeny-switching breakpoint vector b encoding a set of breakpoints in ascending order, where a pair of neighboring breakpoints describes a locus to be used in downstream stages (i.e. stages two and three).

6.1.2 Stage two: Detect candidate loci

The input to stage two are the partitioned multiple sequence alignments $\{X_i\}$. Our general approach to this problem consists of a search among possible sets of loci $\{X_i\}$ using optimization under a "null" version of Coal-Miner's LMM, where we do not consider a test SNP (i.e. $\beta = 0$ in Coal-Miner's LMM) and the phenotypic contributions from putatively associated SNPs in each locus are captured by covariates in W. We apply PCA on each partitioned locus X_i for $1 \leq i \leq l$ and use the top five principal components to model local sample relatedness. Local sample relatedness should only contribute to the linear mixed model when the current scanned locus contains causal sites. We therefore used a search technique to identify candidate loci by utilizing a prediction model where we omit any SNP effect ($\beta = 0$). The LMM log-likelihood optimization criterion used was $\mathcal{L}(\lambda, \tau, \alpha) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|H| - \frac{1}{2}\tau(y - W\alpha)^T H^{-1}(y - W\alpha)$, where W encodes the five covariates obtained above. This approach only includes an environmental residual effect along with a fixed effect W corresponding to a distinct local partition X_i. We computed a LMM log-likelihood optimization score $\mathcal{L}(\lambda, \tau, \alpha)$ for each locus in the set $\{X_i\}$ and then obtained the candidate loci that consist of the top l^* loci based upon the fitted LMM likelihood. The output of stage two is a set of candidate loci $\{X_j^*\}$. We obtained estimates of λ in the range of $[10^{-5}, 1]$ using the optimization heuristic implemented in the GEMMA software library (Zhou & Stephens, 2012), which combines Brent's method (Brent, 1973) and the Newton-Raphson method.

6.1.3 Stage three: Test each marker for significant association with phenotypic character under linear mixed model

We perform a GWA single-marker test to pinpoint associated sites. Coal-Miner utilizes a LMM to evaluate the relationship between y and x. The input to stage three of Coal-Miner are the set of candidate loci $\{X_j^*\}$. For each test SNP x_i, we test it for association with y under the following LMM: $\mathcal{L}(\lambda, \tau, \alpha, \beta) = \frac{n}{2}\log(\tau) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|H| - \frac{1}{2}\tau(y - W\alpha - x_i\beta)^T H^{-1}(y - W\alpha - x_i\beta)$ where $H = \lambda K_{\text{global}} + I_n$. The global sample relatedness is represented using K_{global}, which is a kinship matrix obtained from X that describes the genotypic similarity between individuals. For

each test SNP x_i, if it falls within a candidate locus X_j^*, we compute the top principal component from PCA applied to X_j^*, and assign it as a covariate to W. The local sample relatedness for each test SNP x_i is represented using a single covariate if it falls within a candidate locus but otherwise not. Therefore, we use a model that only accounts for global sample relatedness (use K_{global} and omit W) if x_i does not belong to a candidate locus, or a model that accounts for global and local sample relatedness (use K_{global} and one covariate in W) if x_i belongs to a candidate locus. We obtained an association score for each test SNP x_i using a likelihood ratio test between a fitted LMM model against a null model with no SNP effect.

6.2 Simulation study

We evaluated the performance of Coal-Miner using simulated genetic sequence data where local genealogical variation was due to gene flow and/or incomplete lineage sorting. We used a simulation setup similar to the one described in chapter 5. A set of biallelic loci that follows an infinite sites model was simulated using ms (Hudson, 2002), which is a program for generating samples under neutral models. We utilized a coalescent model with an admixture process similar to the instantaneous unidirectional admixture model described in the study of Durand *et al.* (2011). In our coalescent model, admixture occurs at $t_1 = 2.0$ where a lineage from H comes from B according to probability γ and A according to probability 1 - γ (see Figure D.7 for the description of the model phylogeny used in this study). The choice of branch lengths ($t_0 = 3.0$, $t_1 = 2.0$) used in our simulations are inline with literature evidence related to empirical mice data (Liu *et al.*, 2015). Each replicate consisted of 10 loci. We simulated a multiple sequence alignment containing 1000 sequences with a sequence length of 2500 bp (250 bp/locus). The local partition breakpoint vector b required as input to Coal-Miner was inferred using the Four-Gamete Test (Hudson & Kaplan, 1985). A trait model similar to the one described in chapter 5 was used. Twenty causal sites were randomly selected from one, two, and three loci which we refer to as 10% causal loci, 20% causal loci and 30% causal loci, respectively. Our simulation study also included non-neutral simulations that incorporated positive selection. We used msms (Ewing & Hermisson, 2010)

to conduct forward-time coalescent simulations of genotypic sequence evolution (in place of an otherwise equivalent neutral backward-time coalescent simulation using ms), where causal loci were evolved under deme-dependent positive selection with a finite sites mutation model and all other loci evolved neutrally (as discussed above in the neutral simulation procedure). We used a selection coefficient of $s = 0.56$, which is in line with estimates from prior studies of positive selection in natural *Mus* populations (Song *et al.*, 2011). To recap, the model conditions differed in terms of the proportion of causal loci (either 10%, 20%, or 30%), model phylogeny (either tree-like or non-tree-like), and the presence or absence of positive selection. For each model condition, we repeated the simulation procedure to obtain 20 replicate datasets.

We further included simulations where we varied the admixture times (see Appendix D.1.1) and split/divergence times (see Appendix D.1.2), included an isolation-with-migration model of gene flow (see Appendix D.1.3), and included recombination (see Appendix D.1.4).

6.3 *Arabidopsis* dataset

We reanalyzed the dataset introduced by The 1001 Genomes Consortium (Alonso-Blanco *et al.*, 2016). We downloaded 1,135 *Arabidopsis* strains curated from previous studies (Long *et al.*, 2013; Horton *et al.*, 2012; Cao *et al.*, 2011; Schmitz *et al.*, 2013; Gan *et al.*, 2011; Hagmann *et al.*, 2015) and two traits of interest (the measured flowering time under 10 °C and 16 °C) from www.1001genomes.org. After filtering and quality controls, the genomes contained 10,707,430 biallelic SNPs. These genotypes are native to diverse geographic regions (see Figure D.10 for the inferred phylogeny with the country of origin mapped to the tips of the tree). RecHMM (Westesson & Holmes, 2009) was used to identify and infer recombination-free genomic regions for each chromosome. We sampled four distant taxa from the following geographic regions: Spain, Sweden, USA, and Russia, and used their multiple sequence alignment as input to RecHMM. RecHMM identified 2134, 924, 1574, 695, and 1246 genomic regions with an average length of 14 kb, 21 kb, 15 kb, 26 kb, and 21 kb for chromosomes 1, 2, 3, 4, and 5, respectively. Using CoalMiner on the flowering at 10 °C model condition, we inferred 153, 59, 90, 123, and 99 candidate

loci for chromosomes 1, 2, 3, 4, and 5, respectively. Using Coal-Miner on the flowering at 16 °C model condition, we inferred 129, 132, 132, 153, and 103 candidate loci for chromosomes 1, 2, 3, 4, and 5, respectively. We used a heuristic based on the point of inflection from the distribution of likelihood scores to determine the number of candidate loci (see Figure D.9). Coal-Miner with a minor allele frequency of 0.03 was then applied on the inferred loci and two model conditions (flowering time at 10 °C and 16 °C) to pinpoint regions associated with flowering.

6.4 *Burkholderiaceae* dataset

Bacteria belonging to the *Burkholderiaceae* are of interest given their importance in human and plant disease, but also given their role as plant and fungal endosymbionts and their metabolic capacity to degrade xenobiotics. Fully sequenced (closed) genomes belonging to *Burkholderiaceae* were selected and downloaded from the PATRIC web portal (www.patricbrc.org). Supplementary Table S8 in Hejase *et al.* shows the species names along with other information (IDs, groups, and niche). We chose to maximize phylogenetic and ecological diversity in this sampling, so we included available genomes belonging to free-living, pathogenic, and endosymbiotic species spanning across the genera *Burkholderia*, *Ralstonia*, *Pandoraea*, *Cupriavidus*, *Mycoavidus*, and *Polynucleobacter*. Genomes ranged in size from 2048 (1.56 MB) and 9172 (9.70 MB) coding DNA sequences (CDS). The software package Proteinortho (Lechner *et al.*, 2011) was used to select for single copy orthologs across selected genomes based upon sequence similarity and using default parameters. The collected data contained 57 genomes that are either free-living (52 samples) or endosymbiont (5 samples) and 549 orthologs. We applied Coal-Miner on the inferred orthologs to identify genomic sites or regions behind variation in the following phenotype of interest: animal pathogen vs. non-animal pathogen. We examined the implicated regions inferred by Coal-Miner and identified candidate genes that are associated with the phenotype along with their gene ontology (Ashburner *et al.*, 2000) and KEGG (Kanehisa & Goto, 2000) pathway assignments.

6.5 *Heliconius erato* butterfly dataset

Previous studies have shown that adaptive interspecific introgression has played a key role in the evolution of mimetic butterfly wing patterns (Consortium, 2012; Supple *et al.*, 2013). We re-analyzed data from a previous study (Supple *et al.*, 2013) that involves 45 *Heliconius erato* butterflies collected from four hybrid zones (Peru, Ecuador, French Guiana, and Panama) and belonging to two different red phenotypes: postman and rayed (number of postman = 28 and number of rayed = 17). This dataset constituted from a 400kb region, also known as D interval, which is known to modulate red phenotypic variation in *Heliconius erato*. We used Coal-Miner to compare the two major red phenotypes (postman and rayed) using 56,862 biallelic SNPs. The goal from this empirical study was to identify the variants that modulate red variation. RecHMM was used to infer seven genomic loci in the D interval. We further identified one candidate locus in stage two of the Coal-Miner pipeline. In stage three of the Coal-Miner pipeline, we identified regions associated with red variation using a GWA single-marker test.

6.6 Simulation study results

The simulation study included model conditions that differed in terms of model phylogeny (either tree-like or non-tree-like) and the presence or absence of positive selection. A range of genetic architectures were simulated, where one, two, or three loci contained causal sites. Figure 6.1 compares the performance of Coal-Miner to Coal-Map, GEMMA, and EIGENSTRAT using ROC curves on model conditions with non-tree-like model phylogeny (γ = 0.5) in the absence of positive selection. Coal-Miner offered better power than Coal-Map, GEMMA, and EIGENSTRAT across the different trait model conditions. The performance improvement was significant in terms of AUROC for all model conditions (DeLong *et al.* (1988), α = 0.05), with p-values < 0.00001 for 10% causal loci, 20% causal loci, and 30% causal loci model conditions, respectively. As measured by AUROC, Coal-Miner's performance advantage over the other methods was largest on the 10% causal loci model condition with an AUROC improvement of 0.023, 0.096, and 0.091 for Coal-Map, GEMMA, and EIGENSTRAT, respectively, and smaller as more loci contributed causal

SNPs. This can also be seen based on each method's power at typical false positive rates. At an FPR of 0.1, Coal-Miner's TPR improved upon Coal-Map's, GEMMA's, and EIGENSTRAT's by 0.128, 0.256, and 0.28 on the 10% causal loci model condition, and 0.178, 0.156, and 0.284 on the 20% causal loci model condition. On the 30% causal loci model condition, the TPR difference between Coal-Miner versus Coal-Map, GEMMA, and EIGENSTRAT was 0.169, 0.155, and 0.239, respectively.

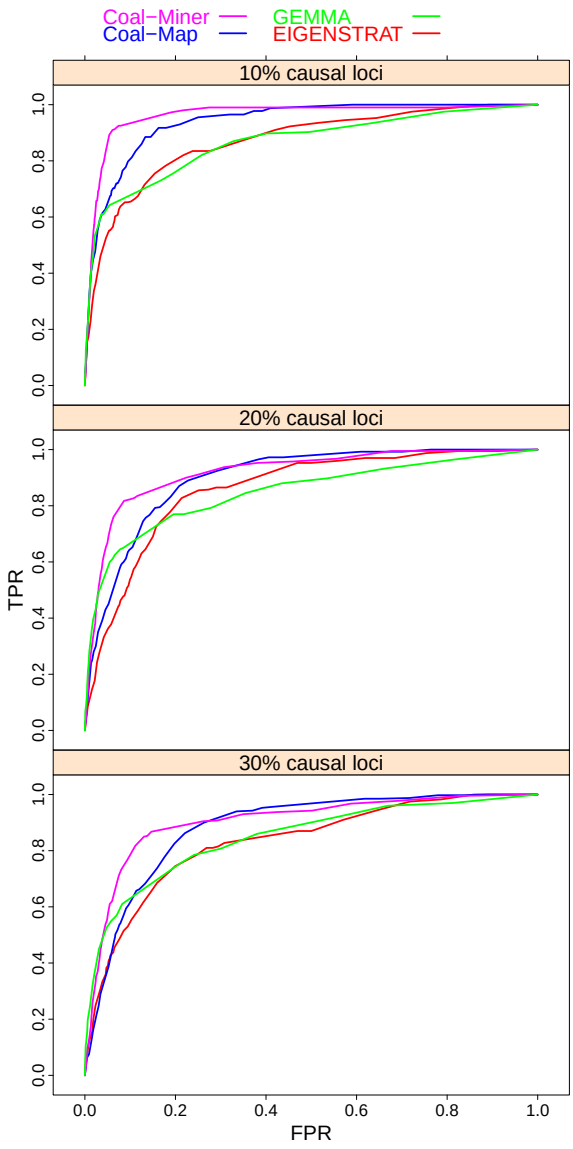


Figure 6.1: **For simulations involving neutral with non-tree-like model phylogeny (hybridization frequency $\gamma = 0.5$), Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA.** The ROC curve shows the relationship between FPR versus TPR. The AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA for 10% causal loci model condition are 0.962, 0.939, 0.871, and 0.866, respectively. For the 20% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.924, 0.899, 0.859, and 0.849, respectively. For the 30% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.905, 0.882, 0.832, and 0.847, respectively.

Coal-Miner's performance advantage over Coal-Map, GEMMA, and EIGENSTRAT was similarly observed on neutral with tree-like model phylogeny (hybridization frequency $\gamma = 0$). The

synthetic traits incorporated genetic contributions from one, two, or three loci. In Figure 6.2, the performance of Coal-Miner, Coal-Map, GEMMA, and EIGENSTRAT on model conditions with no gene flow is shown using ROC curves. Across the different trait architectures (i.e. causal SNPs drawn from one, two, or three loci in the 10% causal loci, 20% causal loci, and 30% causal loci model conditions, respectively), Coal-Miner offered comparable or better power than Coal-Map, GEMMA, and EIGENSTRAT for a given FPR. The performance improvement, using the statistical test of DeLong *et al.*, of Coal-Miner over Coal-Map, GEMMA, and EIGENSTRAT was significant in terms of AUROC for the 10% causal loci, 20% causal loci, and 30% causal loci model conditions with p-values of 0.0053, 0.00001, and 0.00003, respectively. Coal-Miner's performance advantage over Coal-Map, GEMMA, and EIGENSTRAT was largest on the 10% causal loci model condition with an AUROC improvement of 0.021, 0.073, and 0.11, respectively. As more loci contributed causal SNPs, Coal-Miner retained its performance advantage over Coal-Map, GEMMA, and EIGENSTRAT with an AUROC improvement of 0.057, 0.061, and 0.091 for the 20% causal loci model condition, and 0.051, 0.06, and 0.105 for the 30% causal loci model condition, respectively.

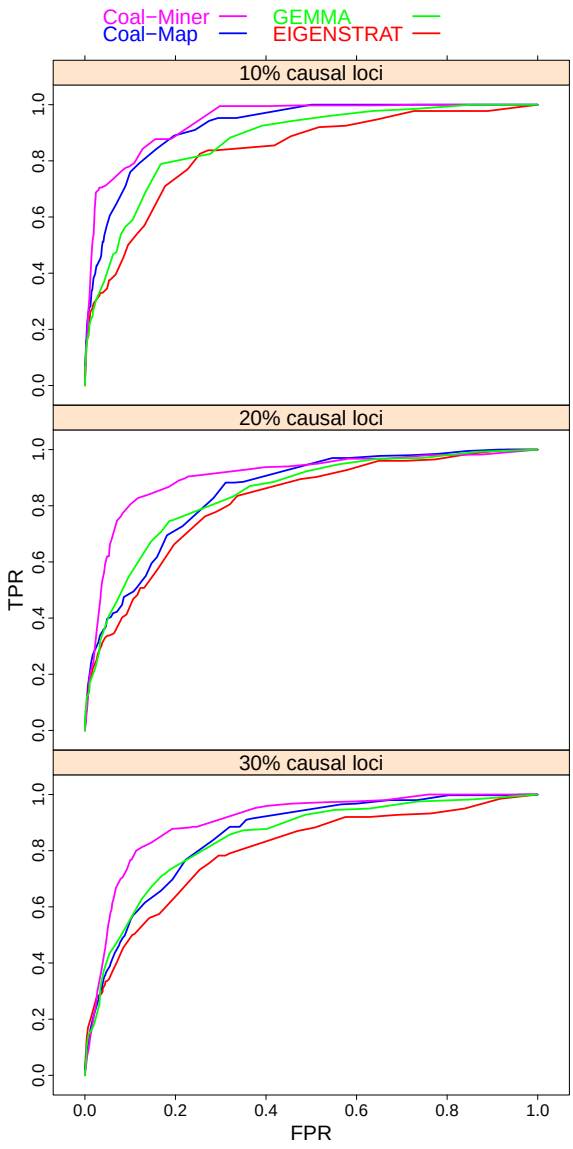


Figure 6.2: **For simulations involving neutral with tree-like model phylogeny (hybridization frequency γ = 0), Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA.** The AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA for 10% causal loci model condition are 0.953, 0.916, 0.876, and 0.938, respectively. For the 20% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.936, 0.889, 0.856, and 0.904, respectively. For the 30% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.908, 0.858, 0.834, and 0.887, respectively. Figure layout and description are otherwise similar to Figure 6.1.

As we examined 10% causal loci, 20% causal loci, and 30% causal loci model conditions on tree-like or non-tree-like model phylogenies in the presence of positive selection, we observed that

Coal-Miner's performance advantage over the other methods did not diminish and in fact remained roughly the same. As shown in Figures D.1 and D.2, the performance improvement in terms of AUROC and TPR at an FPR of 0.1 of Coal-Miner over the other methods remained significant across all trait architectures. On the model condition with no gene flow (i.e. $\gamma = 0$) in the presence of positive selection, virtually all local genealogical variation was due to ILS or positive selection. At an FPR of 0.1, Coal-Miner's TPR improved upon Coal-Map, GEMMA, and EIGENSTRAT by 0.205, 0.369, and 0.406 on 10% causal loci model condition, 0.182, 0.191, and 0.288 on 20% causal loci model condition, and 0.174, 0.16, and 0.217 on 30% causal loci model condition, respectively.

Multi-locus sequence evolution in our simulation study is impacted by genetic drift and ILS, admixture, positive selection, and combinations of these processes. Our simulation study also included additional model conditions that involved alternative models of multi-locus sequence evolution. Each model condition was an extension of the above neutral model condition with 10% causal loci. One set of model conditions varied split time t_1 in the model tree shown in Figure D.7 panel (a). Another set of model conditions varied admixture time t_1 in the model phylogeny network shown in Figure D.7 panel (b), where $\gamma = 0.5$. The impact of recombination was explored in a model condition which made use of the coalescent-with-recombination model (Hudson, 1983). The simulations generated 2.5 kb alignments under a finite-sites model of recombination with per-generation crossover probability between adjacent sites of $10^{-9.85}$, which is 1-2 orders of magnitude smaller than estimates for mouse, rat and human (Jensen-Seaman *et al.*, 2004). We further explored the impact of gene flow using a model condition which substituted the isolation-with-migration model (Notohara, 1990) in place of the IUA model. Table D.1 shows an AUROC comparison of Coal-Miner and the other AM methods on the additional model conditions (see Figures D.3, D.4, D.5, and D.6 for ROC curves).

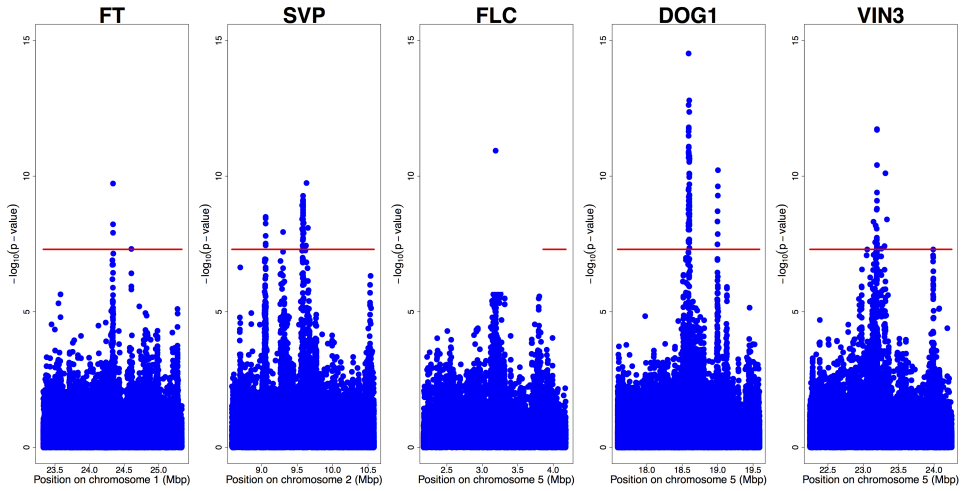
For model conditions that varied divergence time, involved recombination, or incorporated an isolation-with-migration (IM) model of gene flow, Coal-Miner returned significantly improved AUROC compared to the next best method based upon the statistical test of DeLong *et al.*, and the other AM methods were ranked similarly to the experiments which varied the proportion of causal

loci. A similar ranking was obtained when performance was measured using TPR at an FPR of 0.1. Coal-Miner returned a comparable AUROC (within 0.027) as the divergence time t_1 increased from 1.0 to 2.9. The other methods performed similarly, except that the AUROC difference was larger (within 0.031). In the IM-based model condition, all methods returned AUROC that was comparable relative to experiments using the IUA model that were otherwise equivalent. For IUA-based model conditions that varied the admixture time t_1, Coal-Map and Coal-Miner had comparable AUROC which was better than GEMMA and EIGENSTRAT. When comparing TPR at an FPR of 0.1, Coal-Miner returned a significant performance improvement relative to Coal-Map and the other AM methods. Among the AM methods in our study, Coal-Miner's AUROC was least impacted by the choice of admixture time and differed by at most 0.029 as the time t_1 increased from 1.0 to 2.9. The AUROC of the other AM methods became smaller as the admixture time became more ancient, and the AUROC difference was relatively greater than Coal-Miner (as much as 0.086). Overall, Coal-Miner retained its performance advantage relative to the state-of-the-art, with one exception: Coal-Miner and Coal-Map had comparable AUROC on model conditions involving neutral evolution on non-tree-like model phylogenies and 10% causal loci, although Coal-Miner's TPR at an FPR of 0.1 was significantly better than Coal-Map's. These model conditions involved the smallest proportion of causal loci in our study. We note that Coal-Map's performance tended to degrade more rapidly than Coal-Miner as the proportion of causal loci increased, and the relative performance of the two methods may have changed for model conditions with higher proportions of causal loci that are otherwise equivalent.

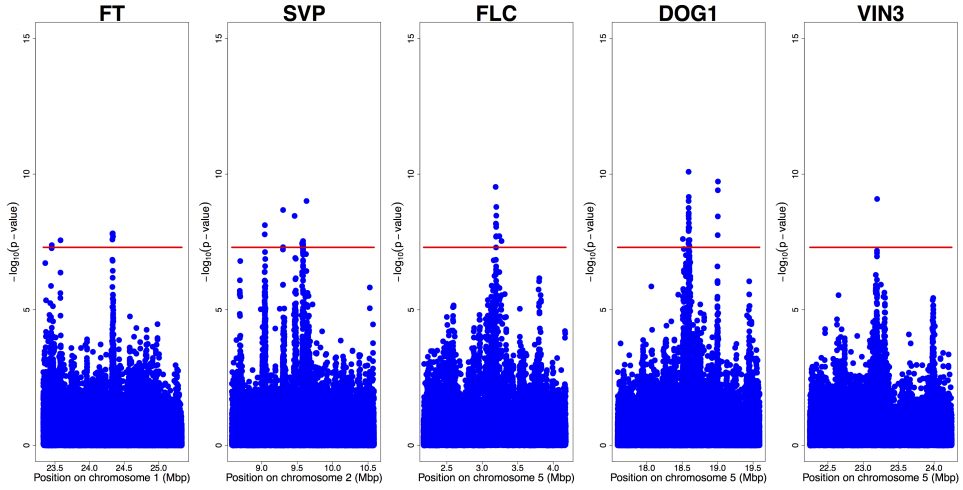
6.7 Empirical study results

We applied Coal-Miner on an *Arabidopsis thaliana* genomic dataset consisting of 1,135 high quality re-sequenced natural lines adapted to different environments with varying local climates (i.e. temperature) (Alonso-Blanco *et al.*, 2016). Coal-Miner identified several significant peaks across the scanned chromosomes associated with flowering time under high and low temperatures. Figure 6.3 displays the Manhattan plots using two model conditions: (a) flowering time at 10 °C

and (b) flowering time at 16 °C. We identified five genes (FT, SVP, FLC, DOG1, VIN3) that are known to regulate flowering and contribute to flowering time variation in *Arabidopsis* (Amasino & Michaels, 2010). Plants rely on both by endogenous and environmental (i.e. temperature and photoperiod) cues to initiate flowering. These five genes encode major components of the vernalization (exposure to the prolonged cold) and autonomous pathways known to regulate the initiation of flowering in *Arabidopsis*. We identified the same five genes to also be associated with flowering time at 16 °C. During cold exposure, VERNALIZATION INSENSITIVE3 (VIN3) functions to repress Flowering Locus C (FLC) to delay the initiation of flowering. A previous GWAS study (Alonso-Blanco *et al.*, 2016) also identified these five genes at 10 °C, but no other candidates, and identified only DELAY OF GERMINATION1 (DOG1) at 16 °C. DOG1 is known to be involved in determining seasonal timing of seed germination and influences flowering time in *Arabidopsis* (Huo *et al.*, 2016). Allelic and copy number variants (CNV) for many of these genes, including FLC, are known to serve important roles in generating novel variation in flowering time and permit plants to adapt to new climates (Méndez-Vigo *et al.*, 2011). We evaluated whether GEMMA detected significantly associated markers in five genomic regions centered on positive control genes identified by Coal-Miner, which are known to regulate flowering time in *Arabidopsis*. We used a Bonferroni-corrected threshold for significance. For three of the five genomic regions (FT, DOG1, and VIN3) in the 10 °C dataset, GEMMA returned association scores that were significant. On the 16 °C dataset, GEMMA returned association scores that were significant on two (FT and DOG1) genomic regions. The corresponding Manhattan plot for the GEMMA analysis is shown in Figure D.8.



(a) 10 °C



(b) 16 °C

Figure 6.3: **Results showing the Manhattan plots after applying Coal-Miner on the *Arabidopsis* dataset using two model conditions: (a) flowering time at 10 °C and (b) flowering time at 16 °C.** The x axis represents the chromosomal position, and the y axis shows the -$\log_{10}$ p-value for all SNPs. The genome-wide significant threshold (p-value = 5 x 10^{-8}) is indicated by the red line. The identified genes that are known to regulate flowering were added based on their respective position. Minor allele frequency of 0.03 was used in the analysis.

We applied Coal-Miner on an empirical dataset of complete genomes of bacteria belonging to the *Burkholderiaceae* and spanning a diversity of ecological states including animal and plant pathogens. Table D.2 shows the genes inferred by Coal-Miner to be associated with human pathogenicity, along with their inferred KEGG pathway and gene ontology assignments. In total, we identified 12 genes associated with human pathogenicity in *Burkholderia*. Four of these genes

have been implicated in pathogenicity by others, and in some cases validated through gene knockout and experimental evolution experiments. For example, the cell division protien FtsK that Coal-Miner associated with human pathogenicity was found to be one of three genes under positive selection in *Burkholderia multivorans* during a 20-year cystic fibrosis infection (Silva *et al.*, 2016). Modifications of another gene identified by Coal-Miner, DNA gyrase subunit A, are well known to be implicated with virulence and antibiotic resistance to quinolone and ciprofloxacin in pathogenic *Burkholderia* (Beceiro *et al.*, 2013; Sousa *et al.*, 2017). For example, Lieberman *et al.* (2011) found that the DNA gyrase subunit A gene was under positive selection during a *Burkholderia dolosa* outbreak among multiple patients with cystic fibrosis. Another gene identified by Coal-Miner, Excinuclease ABC subunit A, has been shown to bind to previously published vaccine targets (Munikumar *et al.*, 2013). Coal-Miner also associated the protein dihydrofolate synthase with animal pathogenity. Point mutations leading to nonsynonymous base changes in the dihydrofolate reductase gene have previously been demonstrated to be associated with trimethoprim resistance in cystic fibrosis patients infected by *Burkholderia cenocepacia* (Drevinek & Mahenthiralingam, 2010; Lefebre & Valvano, 2002).

Figure 6.4 displays the Manhattan plot generated after applying Coal-Miner on the *Heliconius erato* dataset across the D interval. We identified two significant peaks ranging from 502 kb to 592 kb and 658 kb to 682 kb, respectively. The second peak is located at the $3'$of the optix transcription factor, a gene previously shown to be behind the red phenotype variation in *Heliconius*. The first peak is located in a noncoding region more distant from the $3'$of the optix transcription factor.

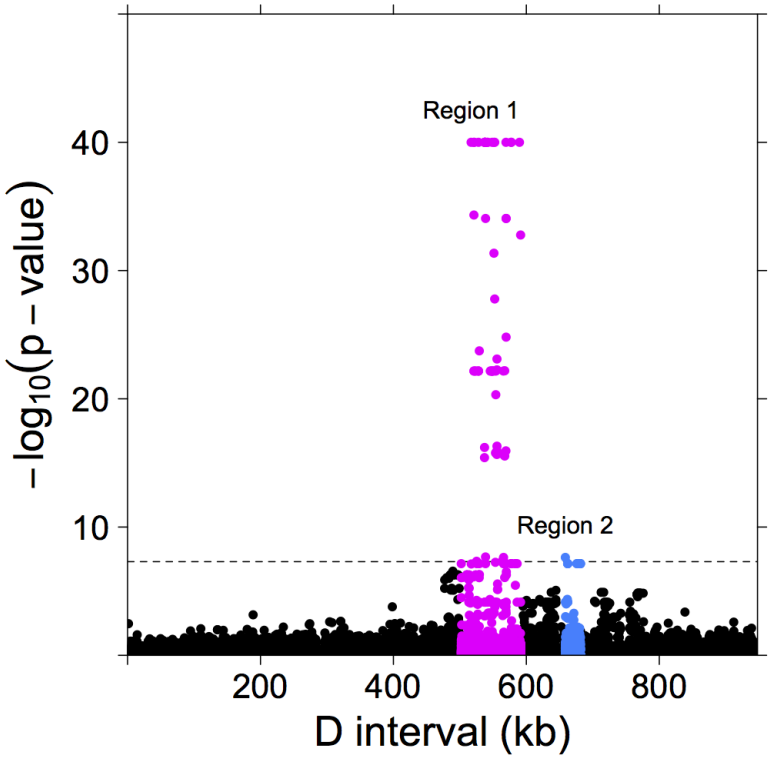


Figure 6.4: **Manhattan plot showing the empirical study results involving *Heliconius erato* butterflies across the D interval.** The x axis represents the genomic position across the D interval, and the y axis shows the -$\log_{10}$ p-value for all SNPs. The genome-wide significant threshold (p-value = 5 x 10^{-8}) is indicated by the dotted black line. The dots indicate genotype by phenotype association calculated for biallelic SNPs using Coal-Miner for four hybrid zones: Peru, Ecuador, French Guiana, and Panama (number of postman = 28; number of rayed = 17). The magenta and blue regions represent the two significant peaks identified by Coal-Miner.

6.8 Discussion

In the simulation study, we explored model conditions where we varied the proportion of causal loci (i.e. 10%, 20%, or 30%) with neutral or non-neutral evolution on tree-like or non-tree-like model phylogenies. As measured using AUROC and TPR at an FPR of 0.1, we observed that Coal-Miner had comparable or better performance to all the other state-of-the-art methods (i.e. Coal-Map, GEMMA, and EIGENSTRAT). Furthermore, as we varied the proportion of causal loci, Coal-Miner retained its performance advantage over the other methods, which suggests that Coal-Miner's performance advantage is robust to the specific proportion of causal loci that have

genetic effects contributions to the quantitative trait. We note that, as even more causal loci are added beyond the proportions explored in our study, the effects contributed by any individual locus becomes more diffuse, and global sample structure will become a more reasonable approximation of different causal loci with different local sample structures. In general, we found traits with "diffuse" genomic architecture (i.e. traits with a relatively high proportion of causal loci) to be challenging for all methods. Coal-Miner tended to cope better with the challenge relative to the other methods in our study, which we attribute to the design of the second stage in the Coal-Miner pipeline (i.e. candidate locus detection). Consistent performance trends were observed when comparing neutral versus non-neutral simulations, which suggests that, for the model conditions that we explored in our study, Coal-Miner's performance is robust to the presence or absence of positive selection. A similar outcome was observed when comparing IUA model-based experiments involving two different types of model phylogenies: tree-like and non-tree-like. Taken together, the model conditions included multiple sources of local genealogical variation, including genetic drift/ILS, gene flow, positive selection, recombination, and combinations thereof. We found that regardless of the evolutionary process that contributed to local genealogical variation, Coal-Miner retained its performance advantage over the other methods. This suggest that Coal-Miner's model and algorithm may be generalized to other evolutionary scenarios, so long as the breakpoint inference method used in the Coal-Miner pipeline suitably accounts for evolutionary processes with first-order contributions to genome evolution.

The empirical datasets in our study were more challenging than the simulated datasets because the former likely involved more complex evolutionary scenarios compared to the latter. Additional evolutionary processes which may have played an important role include other types of natural selection, recombination and demographic events. For the *Arabidopsis* and *Heliconius erato* datasets, Coal-Miner retrieved known associations across all the positive control regions. Furthermore, Coal-Miner analysis of the *Arabidopsis* dataset identified putatively novel markers (i.e. markers which were not tagged using other AM methods). Additional comparative and functional analyses are needed to interpret these findings. For the *Burkholderiaceae* dataset, Coal-Miner detected novel

significant associations, where we validated some of these new findings via a thorough literature review.

In the empirical study, we utilized a data-driven slope heuristic based on the point of inflection from the distribution of likelihood scores in the second stage of Coal-Miner to detect candidate loci. On average, we detected around 10% as candidate loci. We hypothesize that the 10% causal loci model condition in the simulations is the most relevant to our empirical study. We note here that these data-driven slope heuristics have been used in the phylogenetics community for model selection (Solís-Lemus & Ané, 2016). However, other model selection strategies such as Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978) can be used instead to infer the candidate loci.

In this chapter, we introduced Coal-Miner that models candidate loci (local sample relatedness) and global sample relatedness using a generalized LMM. We show through simulation studies that cover a range of evolutionary scenarios that the performance of Coal-Miner is comparable or typically better than the state-of-the-art. We further apply Coal-Miner on three empirical datasets where we identify known and novel genes that encodes for variation in target traits of interest. We conclude with some future research directions. First, other model selection techniques such as cross-validation could be explored for detecting the number of candidate loci in the second stage of Coal-Miner. Cross-validation has been shown to perform well in practice as long as there is sufficient amount of data (Yu *et al.*, 2014). Second, a combined approach that simultaneously performs breakpoint inference, detection of candidate loci, and AM could yield further performance improvements. Finally, the path from genotype to phenotype involves many intermediate layers such as gene expression, metabolic networks, and protein-protein interactions. Integrating those different types of biological data into our pipeline may offer additional performance improvements beyond those observed in this study.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

This dissertation work addresses two main issues. First, we develop an algorithm that is accurate and fast in terms of inferring phylogenetic networks using large-scale genomic sequence datasets. Second, we use better phylogenies to improve the functional interpretation of genomes. Addressing these issues will enable us to better understand and uncover the underlying pinnings of a biological system, which aids in a wide range of applications such as studying health and disease, and understanding fundamental biological processes.

We evaluated the scalability of state-of-the-art methods for inferring phylogenetic networks from multi-locus sequences under genetic drift/ILS, gene flow, and point mutations, where much of the difficulty of this inference problem is due to the complex interplay of all three evolutionary processes, and further quantified the performance of the methods in terms of computational runtime, main memory usage, and topological accuracy on datasets that varied along two separate dimensions of scale: the number of taxa and sequence divergence. We found that these methods face tremendous scalability challenges, in terms of accuracy and speed, on datasets that are well within the scope of today's phylogenomic studies. To address these scalability challenges, we introduced FastNet, which is a new computational method for inferring phylogenetic networks from large-scale genomic sequence datasets. FastNet utilizes a divide-and-conquer algorithm to constrain two different aspects of scale: the number of taxa and evolutionary divergence. We evaluated the performance of FastNet in comparison to state-of-the-art phylogenetic inference methods and found that FastNet improves upon existing methods in terms of computational efficiency and topological accuracy. FastNet was an order of magnitude faster than the most accurate state-of-the-art phylogenetic network inference method. Furthermore, FastNet's topological accuracy was comparable to or typically better than all the other state-of-the-art methods.

We introduced Coal-Map, a new AM method which explicitly models both local sample relatedness, such as arises in a genomic region containing tracts of common introgressive origin, and global sample relatedness. Coal-Map is a methodological pipeline that incorporates recent theoretical innovations that bridge population-level evolution under the coalescent with traditional phylogenetic models of biomolecular sequence evolution. We validated the performance of Coal-Map using synthetic and empirical data. The datasets in our study featured local genealogical variation due to gene flow as well as ILS, sequence mutation, positive selection, and in the case of the empirical mouse genomes: recombination. We compared the performance of Coal-Map to EIGENSTRAT, a leading AM method and consistently observed the same outcome across all of the datasets in our study: Coal-Map's performance in terms of power and FPR was comparable or better than EIGENSTRAT in all cases.

We further introduced Coal-Miner which offered more contributions relative to Coal-Map. First, Coal-Miner utilizes a linear mixed model with multiple effects to explicitly capture the genomic architecture of a phenotype, where both genotypic and phenotypic characters are the product of a complex evolutionary history which can cause sample relatedness to vary locally across genomic loci. Second, the pipeline-based design of Coal-Miner incorporates an intermediate stage to infer candidate loci for use in the new linear mixed model, where a candidate locus is a locus that is inferred to contain one or more putatively associated SNPs. We showed that across a range of genomic architectures and evolutionary scenarios explored in our study, Coal-Miner had comparable or typically improved statistical power and type I error control compared to state-of-the-art AM methods (including Coal-Map). These scenarios included different evolutionary processes such as genetic drift and ILS, positive selection, gene flow, and recombination - all of which can generate local genealogical variation that differs from the true species phylogeny.

7.2 Future work

Several aspects of our performance study can also be revisited in the future to better understand the performance of FastNet and related methods. Dynamic programming based upon Δ assign-

ments will likely be necessary to retain computational efficiency as the number of reticulation nodes increases. Furthermore, more sophisticated techniques for gene tree inference, inferring a guide phylogeny, subproblem decomposition, and merging phylogenetic networks inferred on subproblems can be substituted for the approaches used in our study. Third, the use of a guide phylogeny naturally invites iteration: the output phylogeny from one iteration of the FastNet algorithm would be used as the guide phylogeny for a subsequent iteration of the algorithm. This requires modifying step one of FastNet to utilize a guide network in lieu of a guide tree.

We conclude with our thoughts on future work related to AM. As an alternative to the pipeline-based design of Coal-Miner, simultaneous inference of local coalescent histories and AM model parameters will avoid error propagation across different stages of a pipeline-based algorithm. Furthermore, viewed through the lens of evolution, genotype and phenotype are arguably two sides of the same coin. The same could be said of "intermediate-scale" characters (i.e. interactomic characters). A combination of the extended coalescent models and linear mixed models could be used to capture evolutionary relatedness of and functional dependence between heterogeneous biological characters across multiple scales of complexity and at higher evolutionary divergences.

APPENDICES

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

A.1 Supplementary Tables

Average (SE) topological distance between inferred phylogenetic networks			
	MLE-length	**MP**	**SNaQ**
MLE-length	.11 (.02)	.42 (.06)	.44 (.04)
MP		.36 (.03)	.52 (.05)
SNaQ			.23 (.02)

Table A.1: **Topological distances between inferred phylogenies in the empirical study.** Phylogenies were inferred using a representative method from each category of multi-locus methods: MLE-length (a full likelihood probabilistic method), MP (a parsimony-based method), and SNaQ (a pseudo-likelihood-based probabilistic method). The normalized tripartition distance between solutions that included gene flow (i.e. phylogenetic networks with one reticulation) is shown as an average (standard error) across replicates (n = 20). When constrained to infer a phylogenetic tree rather than a phylogenetic network, all methods inferred an identical species tree across all replicates. Each replicate dataset consists of randomly selecting a sample from the following mouse species and subspecies: *Mus musculus domesticus*, *Mus musculus musculus*, *Mus musculus castaneus*, *Mus spretus*, *Mus spicilegus*, and *Mus macedonicus*.

Sample name	Type (Origin)
B9	Wild caught (Hamm, North Rhine-Westphalia, Germany)
B10	Wild caught (Hamm, North Rhine-Westphalia, Germany)
B11	Wild caught (Hamm, North Rhine-Westphalia, Germany)
C1	Wild caught (Hamm, North Rhine-Westphalia, Germany)
C2	Wild caught (Hamm, North Rhine-Westphalia, Germany)
C3	Wild caught (Hamm, North Rhine-Westphalia, Germany)
MWN1287	Wild caught (Roca del Valles, Catalunya, Spain)
PERC/EiJ	Wild-derived laboratory strain (Nana Village, Rimac Valley, Peru)
WSB/EiJ	Wild-derived laboratory strain (Centerville, Maryland, US)
ZALENDE/EiJ	Wild-derived laboratory strain (Zalende, Switzerland)
MWN1279	Wild caught (Arel, Mallorca island, Spain)
RDS12763	Wild caught (Tubingen, Germany)
KCT222	Wild caught (Remderoda, Germany)
MWN1194	Wild caught (Korinthos, Velo, Peleponissos, Greece)
MWN1198	Wild caught (Laganas, Zakinthos Island, Greece)
MWN1026	Wild caught (San Girogio, Curone Valley, Piamonte, Italy)
MWN1030	Wild caught (Menconico, Staffora Valley, Lombardia, Italy)
MWN1106	Wild caught (Cassino, Lazio, Italy)
MWN1214	Wild caught (Milazzo, Olivarella, Sicily, taly)
22MO	Wild-derived laboratory strain (Monastir, Tunisia)
WMP	Wild-derived laboratory strain (Monastir, Tunisia)
DMZ	Wild-derived laboratory strain (Azemmour, Moroco)
BZO	Wild-derived laboratory strain (Oran, Algeria)
DCA	Wild-derived laboratory strain (Akrotiri, Cyprus)
DCP	Wild-derived laboratory strain (Paphos, Cyprus)
CZECHII/EiJ	Wild-derived laboratory strain (Bratislava, Slovak Republic)
PWK/PhJ	Wild-derived laboratory strain (Lhotka, Bohemia, Czech Repuplic)
SKIVE/EiJ	Wild-derived laboratory strain (Skive, Denmark)
BAG102	Wild caught (Gabortelep, Bekes, Hungary)
BAG3	Wild caught (Bukovce, Slovak Republic)
BAG56	Wild caught (Pomykow, Lublin, Poland)
BAG68	Wild caught (Wola Duza, Lublin, Poland)
BAG74	Wild caught (Krasne, Podkarpackie, Poland)
BAG94	Wild caught (Szepes, Debrecen, Hajdu-Bihar, Hungary)
BAG99	Wild caught (Szomolyom, Hajdu-Bihar, Hungary)
RDS10105	Wild caught (Monchhof, Austria)
RDS13554	Wild caught (Hubinger-Leitham, Austria)
Yu2097m	Wild caught (Urumqi, Xinjiang, China)
Yu2099f	Wild caught (Urumqi, Xinjiang, China)
Yu2115m	Wild caught (Yutian, Xinjiang, China)
Yu2120f	Wild caught (Hebukesaier, Xinjiang, China)
CIM1	Wild-derived laboratory strain (Masinagudi, India)
POHN	Wild-derived laboratory strain
CAST/EiJ	Wild-derived laboratory strain (Thonburi, Thailand)
SPRET/EiJ	Wild caught (Puerto Real, Cadiz Province, Spain)
SEG1	Inbred lab
ZRU1	Inbred lab
YCA1	Inbred lab
XBS1	Inbred lab

Table A.2: **Empirical mice genomic data along with their type and origin (City, Province, Country).** Origin was only reported for the wild-derived and wild caught laboratory strains.

A.2 Supplementary Figures

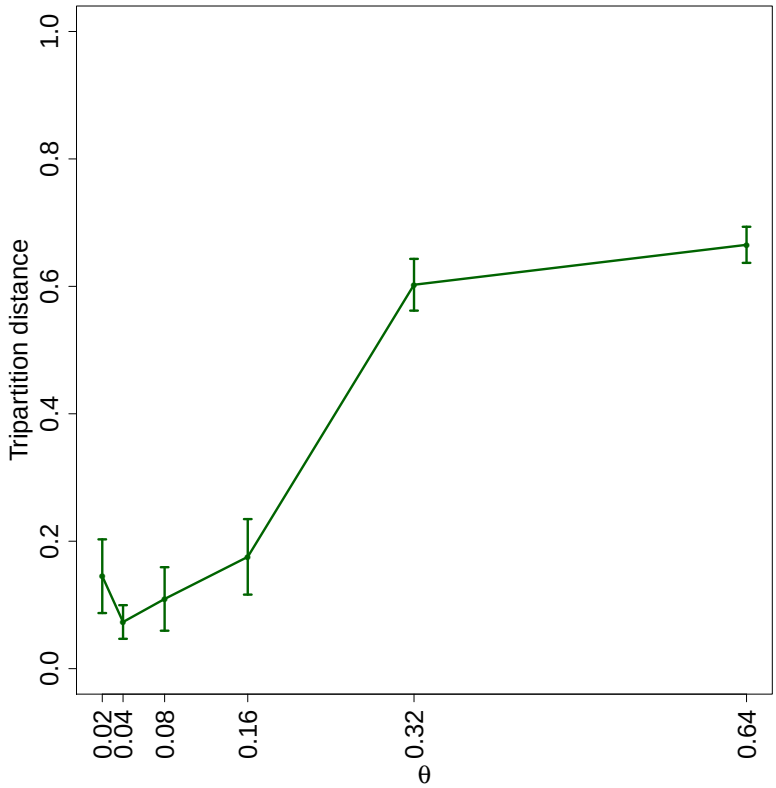


Figure A.1: **The impact of sequence divergence on the topological error of MLE-length**. We assessed the performance of MLE-length to characterize the accuracy of multi-locus inference methods since MLE-length was generally more accurate than MLE, SNaQ, MPL, and MP. Results are shown on seven-taxon datasets across six model conditions where θ ranged from 0.02 to 0.64. The average and standard error of the tripartition distance between the inferred and model networks are reported for twenty replicates.

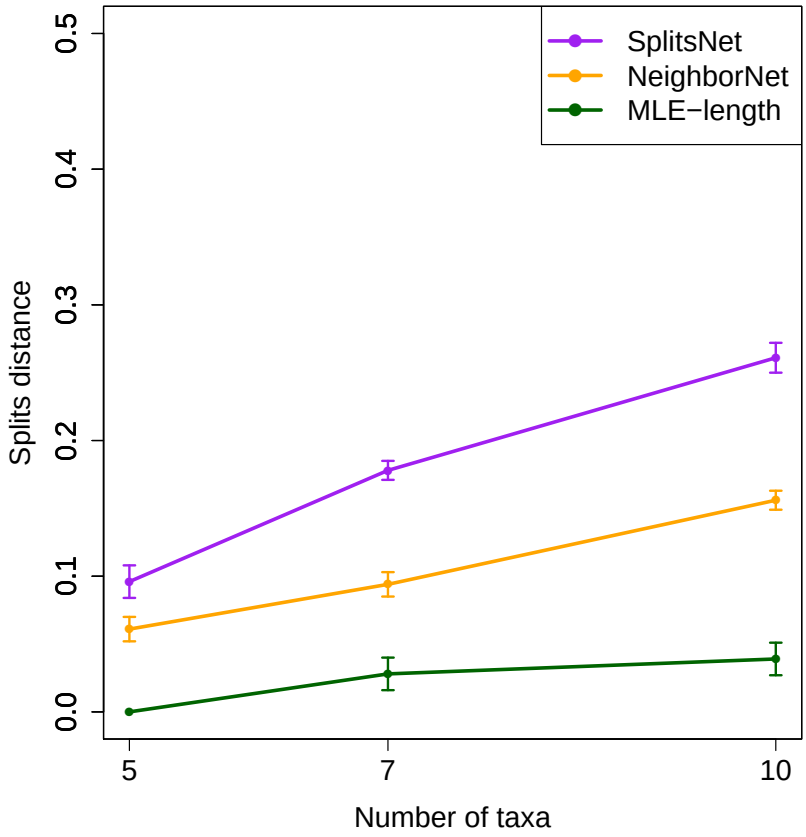


Figure A.2: **Performance comparison of concatenation-based (SplitsNet and NeighborNet) and summary-based (MLE-length) inference methods across different dataset sizes.** We assessed the performance of MLE-length to characterize the accuracy of multi-locus inference methods since MLE-length was generally more accurate than MLE, SNaQ, MPL, and MP. Results are shown for three model conditions where the number of taxa ranged from five to ten with θ of 0.08. True gene trees were used as input to MLE-length. The average and standard error of the splits distance between the inferred and model networks are reported for twenty replicates.

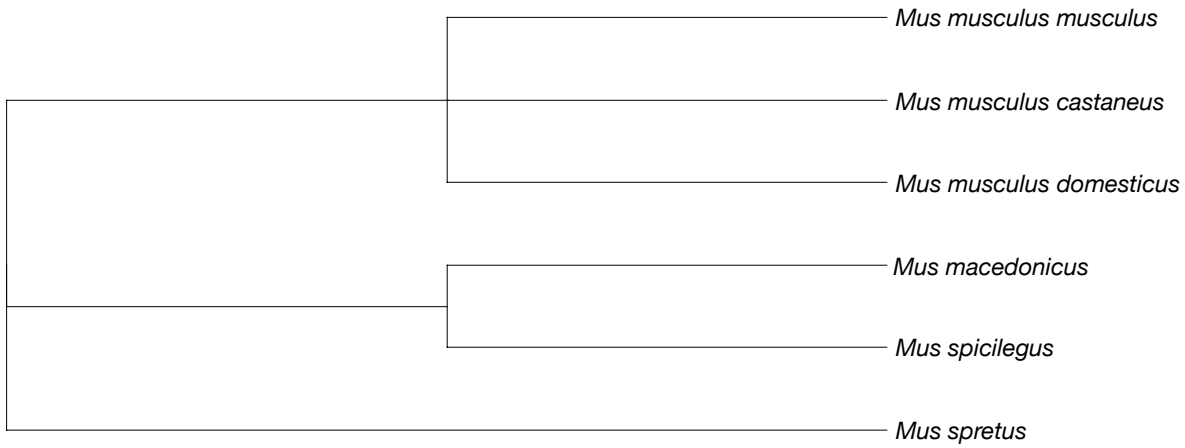


Figure A.3: **The *Mus* consensus phylogeny proposed by Guénet & Bonhomme (2003).** Previous studies (Liu *et al.*, 2015; Staubach *et al.*, 2012) identified gene flow between the *Mus musculus* subspecies and between *Mus musculus domesticus* and *Mus spretus*.

APPENDIX B

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

B.1 Supplementary Tables

Number of taxa	Topological distance			Runtime in hours		
	Avg	SE	q value	Avg	SE	q value
15	0.066	0.001	1.5×10^{-2}	34.975	4.129	1.3×10^{-7}
20	0.070	0.014	1.1×10^{-2}	71.138	7.668	8.7×10^{-8}

Table B.1: **Average distances and runtimes for the performance boost of FastNet (with MLE-length as a base method) over the base method itself using model conditions containing 15 or 20 taxa.** The topological distance between the inferred and model phylogenies was measured using the tripartition distance. The model conditions involved model phylogenies that contained one reticulation node with deep gene flow. True gene trees were used as input to FastNet and MLE-length. Average ("Avg") and standard errors ("SE") for the performance improvement of topological distances and runtimes are listed (n = 20). A one-sided t-test comparing the performance advantage of FastNet over the boosted method (MLE-length) for the evaluation criteria (i.e. topological distance and runtime) was conducted. Corrected q-values are reported where multiple test correction was performed using Benjamini & Hochberg.

Number of taxa	Topological distance			Runtime in hours		
	Avg	SE	q value	Avg	SE	q value
15	0.015	0.017	3.8×10^{-1}	2.334	0.227	5.1×10^{-4}
20	0.166	0.035	3.2×10^{-3}	8.021	1.473	3.2×10^{-3}

Table B.2: **Average distances and runtimes for the performance boost of FastNet (with MPL as a base method) over the base method itself using model conditions containing 15 or 20 taxa.** Table layout and description are otherwise similar to Table B.1.

B.2 Supplementary Figures

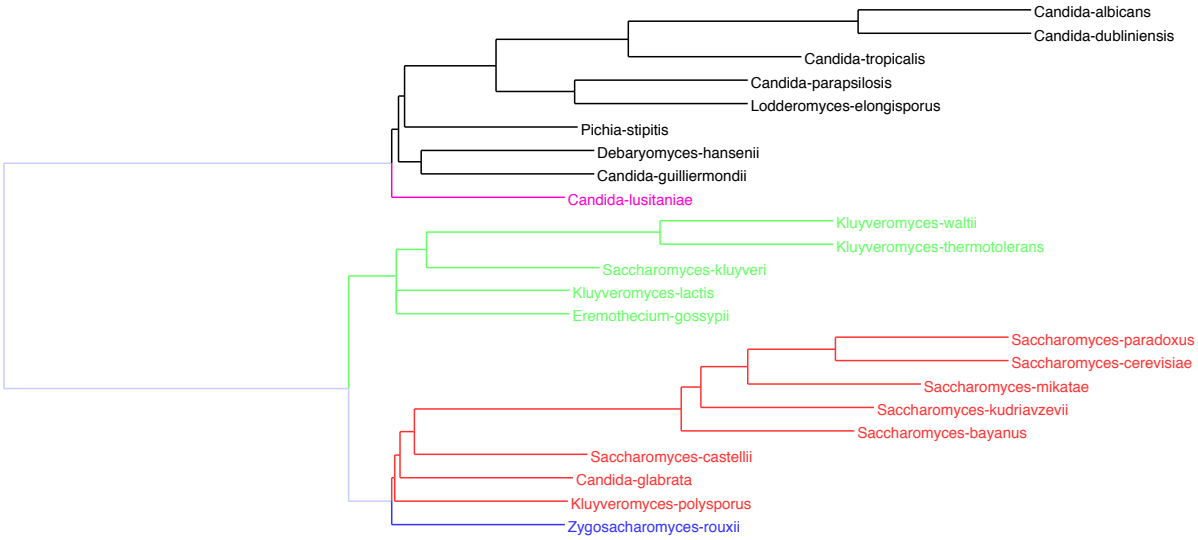


Figure B.1: **Subproblem decomposition of FastNet on the 1070-gene yeast dataset.** Each leaf tip is colored according to one of five subproblems (i.e. black, magenta, green, red, or blue). We note here that Candida-lusitaniae (colored in magenta) and Zygosacharomyces-rouxii (colored in blue) belong to subproblems containing one taxon.

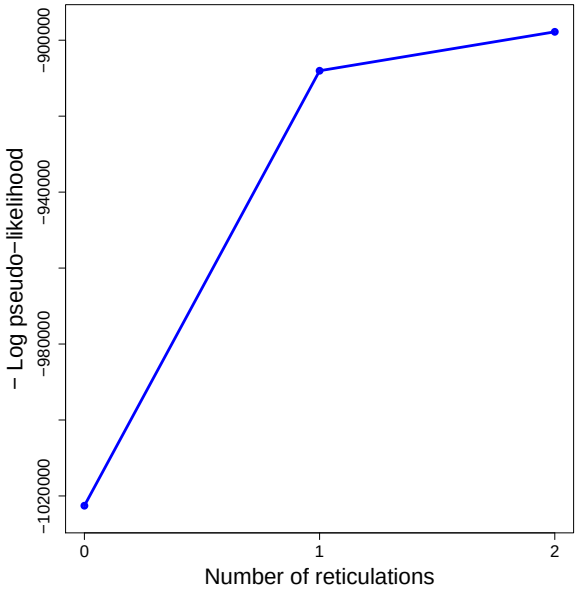


Figure B.2: **A slope analysis of the inferred phylogenies using pseudo-likelihood scores on the 1070-gene yeast dataset.** The x axis shows the number of reticulations used to infer a FastNet network while the y axis shows the - log pseudo-likelihood score for each FastNet inferred network.

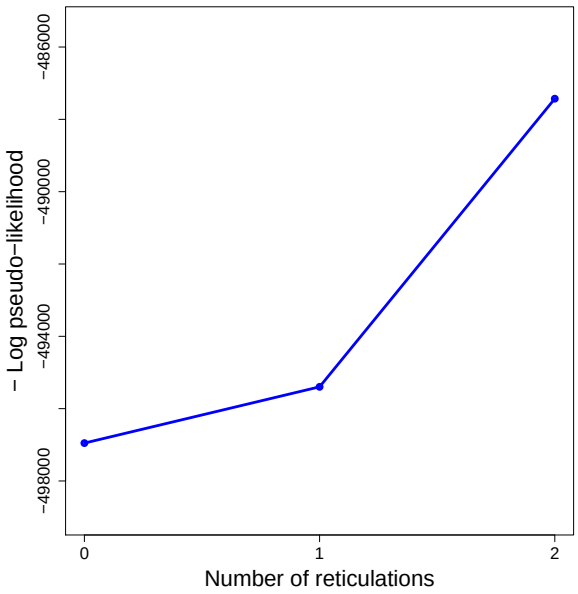


Figure B.3: **A slope analysis of the inferred phylogenies using pseudo-likelihood scores on the mosquito dataset.** Figure layout and description are otherwise similar to Figure B.2

APPENDIX C

SUPPLEMENTARY MATERIALS FOR CHAPTER 5

C.1 Simulation study

C.1.1 Neutral with non-tree-like model phylogeny

The following ms command was used to generate a multiple sequence alignment for the neutral model conditions with non-tree-like model phylogenies that include drift/ILS and gene flow:

```
ms 1000 10 -t 4.0 -s 250 -T -I 4 250 250 Ca Cb -ej 3.0 2 1
-ej 2.0 3 1 -ej 2.0 4 2
```

where the number of taxa is 1000, the number of gene trees is 10, the -t switch represents the mutation parameter $4N_0\mu$ where N_0 is the diploid population size (N_0 = 2.5 x 10^5) and μ is the neutral mutation rate for a locus (μ = 4 x 10^{-6}), the number of segregating sites is 250, and the -T parameter outputs the gene trees, which represent the evolutionary history of the sampled taxa. The -I parameter is followed by the number of subpopulations (k = 4) and a list of integers (n_A = 250, n_B = 250, n_{Ca} = Ca, n_{Cb} = Cb) that represent the number of taxa sampled for each subpopulation. Ca and Cb vary across loci and are dependent on γ (0.01 or 0.1 or 0.25 or 0.5). The -ej switch (-ej t i j) moves all lineages from subpopulation i to subpopulation j at time t.

C.1.2 Non-neutral with non-tree-like model phylogeny

The following msms command was used to generate a multiple sequence alignment for the non-neutral model conditions with non-tree-like model phylogenies that include drift/ILS, gene flow, and positive selection:

```
java -jar msms.jar 1000 <Number of causal loci> -t 4.0 -s 250 -T
-I 4 250 250 Ca Cb 0 -ej 3.0 2 1 -ej 2.0 3 1 -ej 2.0 4 2
-SI 2.0 4 0 0 0 0 -Sc 0 4 11200 6272 0 -Sc 0 3 11200 6272 0
-Smu 4.0 -N 10000
```

where the -SI switch (-SI t <number of populations> A B Ca Cb) sets the start of selection to time t forward in time from this point, the -Sc switch (-Sc t i α_{AA} α_{Aa} α_{aa}) sets the selection strength in population i pastward from time t to 2Ns, the -Smu switch sets the forward mutation rate for the selected allele, and the -N switch is the effective population size.

C.2 EIGENSTRAT

The following command was used to generate the principal components:

```
smartpca.perl -i example.geno -a example.snp -b example.ind -k 10 -q YES -o example.pca
-p example.plot -e example.eval -l example.log -m 5 -t 2 -s 6
```

where the -i parameter specifies the genotype file, the -a parameter specifies the SNP file, the -b parameter specifies the individual file, the -k parameter specifies the number of principal components to output, the -q parameter specifies whether the phenotype is quantitative, the -o parameter specifies the output file of principal components, the -p parameter specifies the prefix of output plot files of top two principal components, the -e parameter specifies the output file of all eigenvalues, the -l parameter specifies the output log file, the -m parameter specifies the maximum number of outlier removal iterations, the -t parameter specifies the number of principal components along which to remove outliers, and the -s parameter specifies the number of standard deviations which an individual must exceed to be removed as an outlier.

The following command was used to apply the association analysis:

smarteigenstrat.perl -i example.geno -a example.snp -b example.ind -q YES -p example.pca -k 10 -o example.chisq -l example.log

where the -p parameter specifies the input file of principal components, the -k parameter specifies the number of principal components along which to correct for population structure, the -o parameter specifies the χ^2 association statistics, and the -l parameter specifies the standard output file.

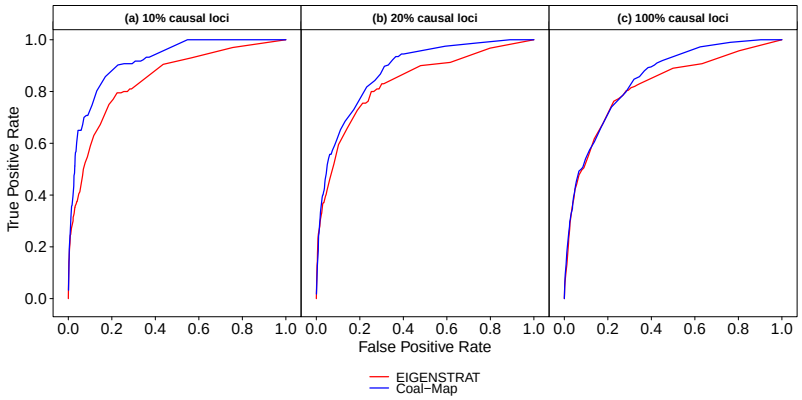
C.3 Inferring local-phylogeny-switching breakpoints

The local partition breakpoint vector b for the simulation study required as input to Coal-Map was inferred using the Four-Gamete Test (Hudson & Kaplan, 1985), which identifies segregating sites that did not arise without either recombination or a repeat mutation. The Four-Gamete Test is an appropriate choice to detect breakpoints due to the simplifying assumptions of our simulation study (infinite sites model, free recombination between markers, and complete linkage within each marker). For the empirical study, we used RecHMM (Westesson & Holmes, 2009), an HMM-based method for computing local-phylogeny-switching breakpoints. The following command was used to run RecHMM:

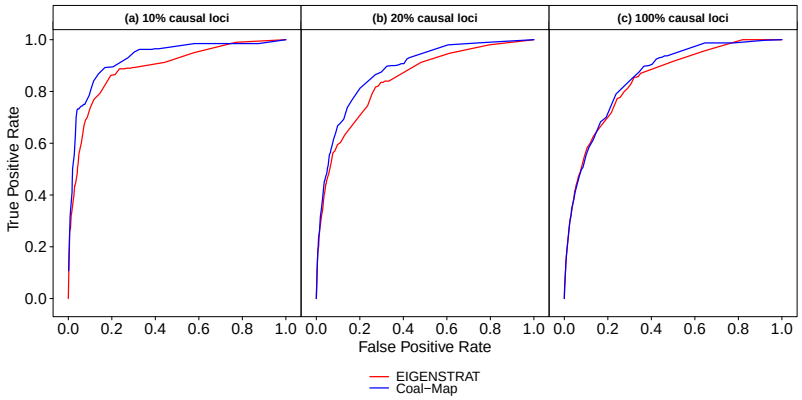
./runTraining.py <FASTA input alignment> -lb -prefix <empty existing working directory> -k <number of hidden states> -lt

Using 2 states for the -k option corresponds to two parental trees for the model network.

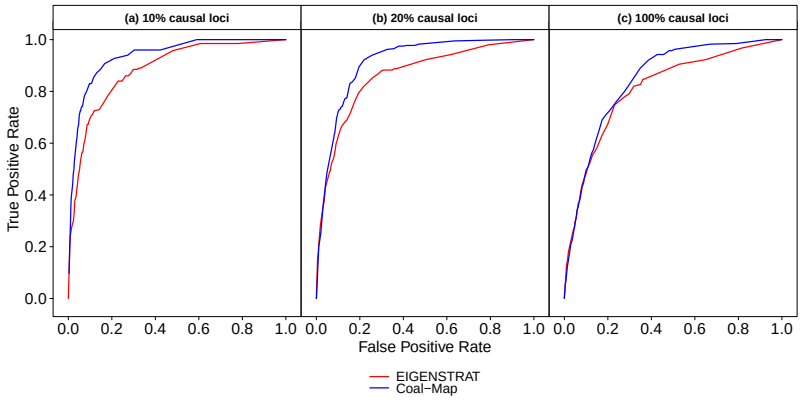
C.4 Supplementary Figures



(a) $\gamma = 0.01$

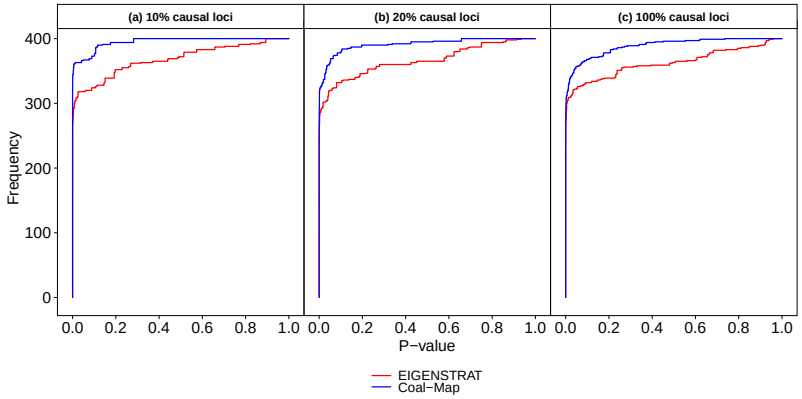


(b) $\gamma = 0.1$

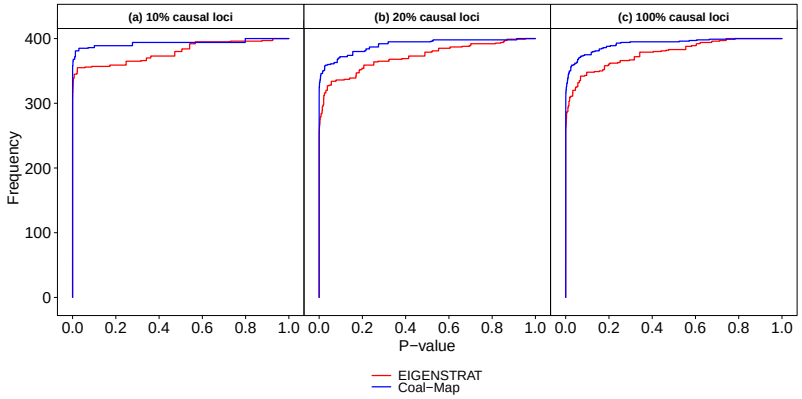


(c) $\gamma = 0.25$

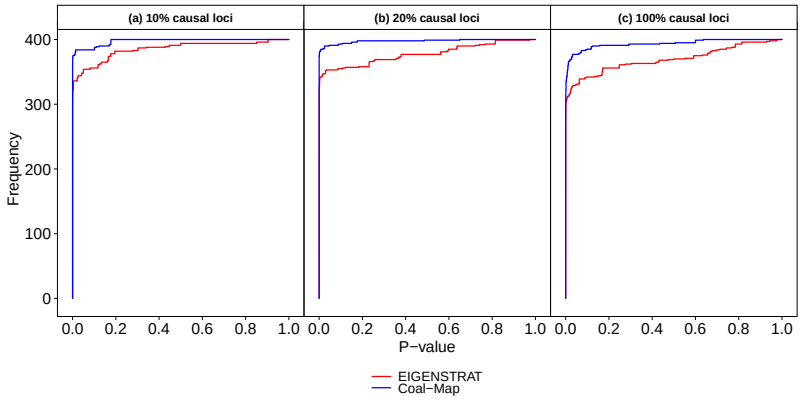
Figure C.1: **For simulations involving neutral gene flow with non-tree-like model phylogeny across a range of hybridization frequencies (γ = 0.01, 0.1, or 0.25), Coal-Map has an equal or better power and comparable type I error to EIGENSTRAT.** Figure layout and description are otherwise similar to Figure 5.2.



(a) $\gamma = 0.01$



(b) $\gamma = 0.1$



(c) $\gamma = 0.25$

Figure C.2: **For simulations involving neutral gene flow with non-tree-like model phylogeny across a range of hybridization frequencies (γ = 0.01, 0.1, or 0.25), the cumulative histogram of p-values at causal sites for Coal-Map and EIGENSTRAT are reported.** Figure layout and description are otherwise similar to Figure 5.4.

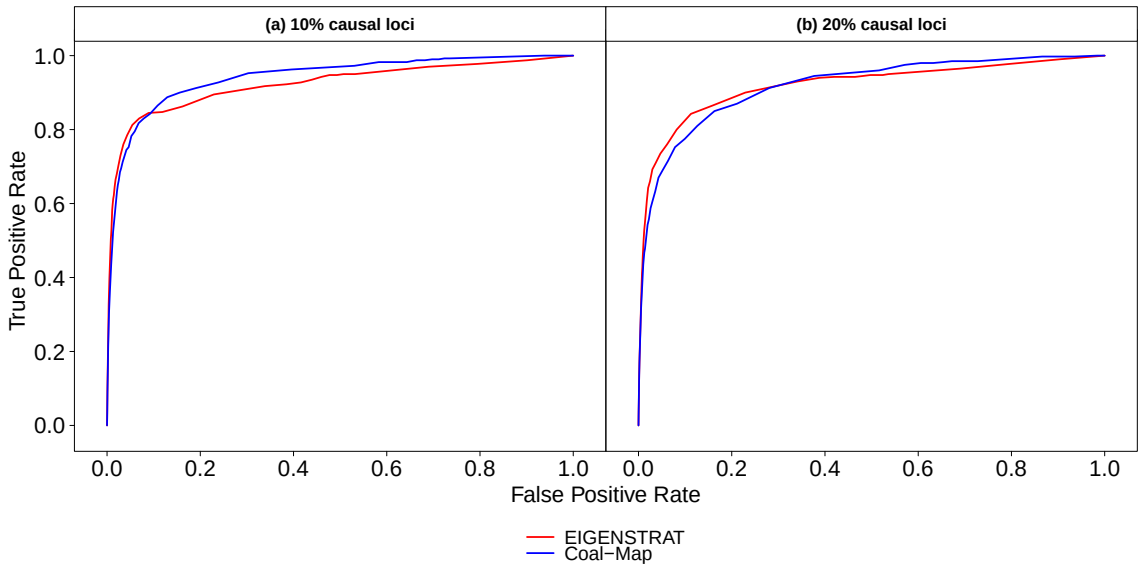


Figure C.3: **Using empirical genomic data from mouse chromosome 15, Coal-Map has an equal or better power and comparable type I error to EIGENSTRAT.** Results are shown for two model conditions: (a) 10% causal loci and (b) 20% causal loci. Figure layout and description are otherwise similar to Figure 5.2.

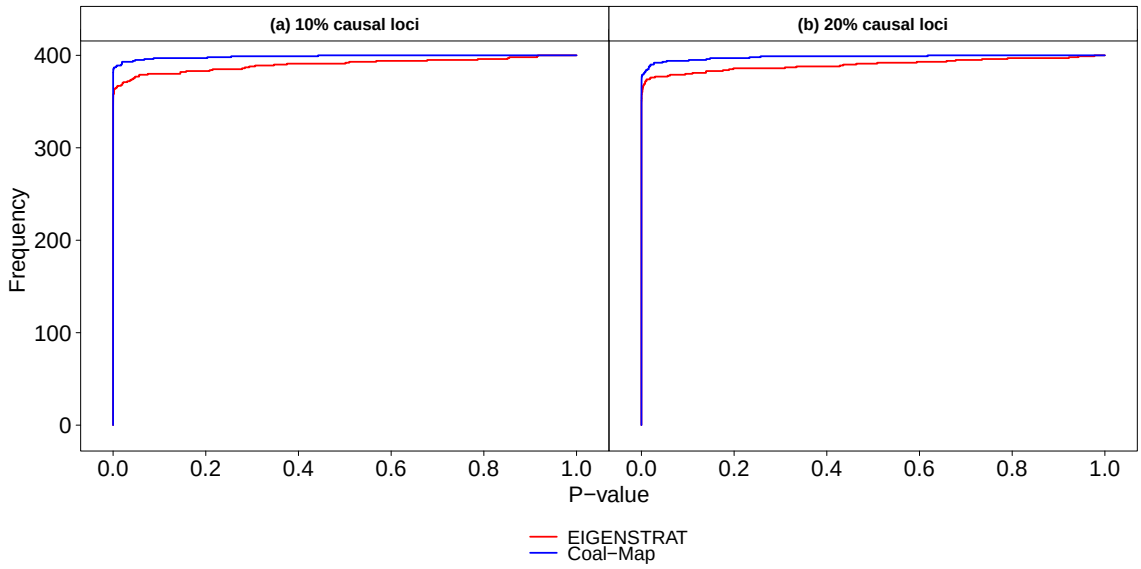


Figure C.4: **Using empirical genomic data from mouse chromosome 15, the cumulative histogram of p-values at causal sites for Coal-Map and EIGENSTRAT are reported.** Results are shown for two model conditions: (a) 10% causal loci and (b) 20% causal loci. Figure layout and description are otherwise similar to Figure 5.4.

APPENDIX D

SUPPLEMENTARY MATERIALS FOR CHAPTER 6

D.1 Simulation study

D.1.1 Neutral with non-tree-like model phylogeny

We explored the impact of different admixture times by simulating two datasets with admixture occurring at $t_1 = 1.0$ and $t_1 = 2.9$. We used the following ms commands to generate the aforementioned simulations:

```
ms 1000 10 -t 4.0 -s 250 -T -I 4 250 250 Ca Cb -ej 3.0 2 1
-ej 1.0 3 1 -ej 1.0 4 2
```

```
ms 1000 10 -t 4.0 -s 250 -T -I 4 250 250 Ca Cb -ej 3.0 2 1
-ej 2.9 3 1 -ej 2.9 4 2
```

D.1.2 Neutral with tree-like model phylogeny

We further explored the impact of different split times by simulating two more datasets with divergence occurring at $t_1 = 1.0$ and $t_1 = 2.9$. We used the following ms commands to generate the aforementioned simulations:

```
ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej 1.0 3 2 -ej 3.0 2 1
ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej 2.9 3 2 -ej 3.0 2 1
```

D.1.3 Isolation with migration

ms (Hudson, 2002) was used to simulate a multiple sequence alignment for the neutral model conditions with non-tree-like model phylogenies incorporating an isolation-with-migration (IM) model of gene flow:

```
ms 1000 10 -t 4.0 -s 250 -T -I 3 250 250 500 -ej 2.0 3 2 -ej 3.0 2 1
-em 1 1 3 1
```

where the -em switch (-em t i j x) sets $4N_0m_{ij}$ (m_{ij} = 10^{-6}) to x at time t and m_{ij} is the fraction of subpopulation i in each generation which consist of migrants from subpopulation j. The migration rate used in this simulation is inline with previous studies (Hejase & Liu, 2016b).

D.1.4 Recombination

We simulated a multiple sequence alignment under the coalescent model with uniform recombination rate across a locus. We used a total sequence length of 2.5 kb, and a p parameter of 0.35, which is $4N_0$r, where r is the probability of cross-over per generation between the ends of the locus. The per-generation crossover probability of $10^{-9.85}$ between adjacent sites was used. Therefore, the probability of cross-over between the ends of the locus is: $10^{-9.85}$ x (2500-1) = 3.5 x 10^{-7} and p = 4 x 2.5 x 10^5 x 3.5 x 10^{-7} = 0.35. On average, we obtained 10 recombinant regions per replicate. The following ms command was used to generate a multiple sequence alignment for the neutral model conditions with tree-like model phylogenies incorporating recombination:

```
ms 1000 1 -t 4.0 -s 2500 -T -I 3 250 250 500 -ej 2.0 3 2 -ej 3.0 2 1
-r 0.35 2500
```

D.2 GEMMA

We used GEMMA (Zhou & Stephens, 2012) which utilizes a linear mixed model to account for sample relatedness. GEMMA represents the phenotype y as a function of fixed ($W\alpha + X\beta$) and random ($u + \epsilon$) effects:

$$\begin{aligned} y &= W\alpha + x\beta + u + \epsilon \\ u &\sim MVN_n(0, \lambda\tau^{-1}K) \\ \epsilon &\sim MVN_n(0, \tau^{-1}I_n) \end{aligned} \tag{D.1}$$

where y is the phenotype vector, W includes the fixed effects, α encodes the coefficients of the covariates located in W, x is the test locus, β is the effect size of x, u is a random effect that follows an n-dimensional multivariate normal distribution, K is a kinship matrix which is represented as a pairwise genotypic similarity between individuals, λ is the ratio between two variance components (genetic and environmental effects), τ is the variance of residual errors, ϵ is a random effect that follows an n-dimensional multivariate normal distribution and is used to model any unexplained variation in y, and I_n is an n by n identity matrix. The parameters $\hat{\alpha}$, $\hat{\beta}$, $\hat{\tau}$, and $\hat{\lambda}$ are estimated using maximum likelihood where the association test statistics for x are generated using likelihood-ratio test between the fitted model against a null model with no SNP effect.

The following command was used to generate a kinship matrix:

> gemma -g <specify input genotype file name> -p <specify input phenotype file name> -a <specify input SNPs annotation file name> -gk 1 <kinship/relatedness matrix type> -o <specify output file prefix>

The following command was used to run the association test:

> gemma -g <specify input genotype file name> -p <specify input phenotype file name> -a

<specify input SNPs annotation file name> -n 1 <specify phenotype column in the phenotype file> -maf 0 <specify minor allele frequency threshold> -r2 1 <specify r-squared threshold> -k <specify input kinship/relatedness matrix file name> -lmm 2 <specify frequentist analysis choice> -o <specify output file prefix>

D.3 Supplementary Tables

Model condition	AUROC				q-value
	Coal-Miner	Coal-Map	GEMMA	EIGENSTRAT	
Non-tree-like model phylogeny with admixture time $t_1 = 1.0$	0.959	**0.963**	0.922	0.905	0.9959
Non-tree-like model phylogeny with admixture time $t_1 = 2.9$	**0.933**	0.922	0.836	0.843	< 0.00001
Tree-like model phylogeny with split time $t_1 = 1.0$	**0.959**	0.899	0.884	0.852	< 0.00001
Tree-like model phylogeny with split time $t_1 = 2.9$	**0.932**	0.895	0.853	0.849	< 0.00001
Coalescent-with-recombination	**0.841**	0.768	0.77	0.738	< 0.00001
Isolation-with-migration	**0.953**	0.931	0.881	0.868	< 0.00001

Table D.1: **Additional evolutionary scenarios exploring other evolutionary processes that can generate local genealogical variation.** The additional model conditions were variants of the model condition with neutral evolution on a tree-like or non-tree-like model phylogenies and 10% causal loci. Each model condition incorporated an alternative evolutionary scenario (see Appendix D.1.1, D.1.2, D.1.3, and D.1.4 for more details). The performance of each AM method was evaluated based on AUROC where we report each method's AUROC as an average across twenty replicate datasets for each model condition. The AUROC of the most accurate method is shown in bold. We report Coal-Miner's performance advantage based upon the AUROC of the most accurate of the other AM methods, based upon the test of DeLong *et al.*. We corrected for multiple tests using the approach of Benjamini & Hochberg, and corrected q-values are shown.

Proteins	Pathway Assignments	Gene Ontology Assignments
Dihydrofolate synthase	KEGG:00790 Folate biosynthesis	GO:0008841 dihydrofolate synthase activity, GO:0004326 tetrahydrofolylpolyglutamate synthase activity
Aspartokinase	KEGG:00260 Glycine, serine and threonine metabolism, KEGG:00270 Cysteine and methionine metabolism, KEGG:00300 Lysine biosynthesis	GO:0004072 aspartate kinase activity
NADH-ubiquinone oxidoreductase chain G	KEGG:00190 Oxidative phosphorylation, KEGG:00910 Nitrogen metabolism	GO:0008137 NADH dehydrogenase (ubiquinone) activity
Excinuclease ABC subunit A	-	GO:0005524 ATP binding, GO:0016887 ATPase activity
Carboxyl-terminal protease	-	-
Homoserine O-acetyltransferase	KEGG:00270 Cysteine and methionine metabolism, KEGG:00920 Sulfur metabolism	GO:0004414 homoserine O-acetyltransferase activity
Glutamate-ammonia-ligase adenylyltransferase	-	GO:0008882 [glutamate-ammonia-ligase] adenylyltransferase activity
Undecaprenyl-diphosphatase	KEGG:00550 Peptidoglycan biosynthesis	GO:0050380 undecaprenyl-diphosphatase activity
Cell division protein FtsK	-	-
DNA gyrase subunit A	-	GO:0003918 DNA topoisomerase (ATP-hydrolyzing) activity
Diaminohydroxyphosphoribosylaminopyrimidine deaminase	KEGG:00740 Riboflavin metabolism	GO:0008703 5-amino-6-(5-phosphoribosylamino)uracil reductase activity, GO:0008835 diaminohydroxyphosphoribosylaminopyrimidine deaminase activity
Ribonucleotide reductase of class Ia (aerobic), alpha subunit	KEGG:00230 Purine metabolism, KEGG:00240 Pyrimidine metabolism, KEGG:00480 Glutathione metabolism	GO:0004748 ribonucleoside-diphosphate reductase activity
DNA gyrase subunit B	-	GO:0003918 DNA topoisomerase (ATP-hydrolyzing) activity
Ketol-acid reductoisomerase	KEGG:00290 Valine, leucine and isoleucine biosynthesis, KEGG:00770 Pantothenate and CoA biosynthesis	GO:0004455 ketol-acid reductoisomerase activity
Phosphoribosylformylglycinamidine synthase, synthetase subunit	KEGG:00230 Purine metabolism	GO:0004642 phosphoribosylformylglycinamidine synthase activity
DNA polymerase I	KEGG:00230 Purine metabolism, KEGG:00240 Pyrimidine metabolism	GO:0003887 DNA-directed DNA polymerase activity

Table D.2: **Empirical study results involving bacteria belonging to the *Burkholderiaceae*.** Results are shown for proteins inferred to be associated with human pathogenicity along with their KEGG pathway and gene ontology assignments.

D.4 Supplementary Figures

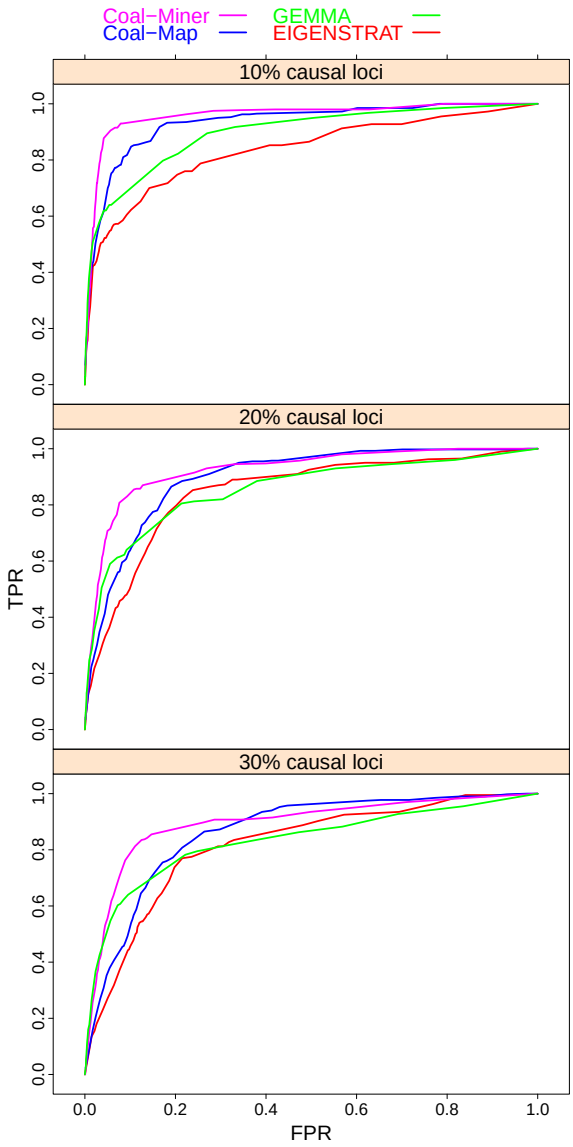


Figure D.1: **For simulations involving non-neutral with non-tree-like model phylogeny (hybridization frequency γ = 0.5 and selection coefficient s = 0.56), Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA.** The AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA for the 10% causal loci model condition are 0.959, 0.933, 0.836, and 0.896, respectively. For the 20% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.926, 0.897, 0.847, and 0.856, respectively. For the 30% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.894, 0.863, 0.816, and 0.832, respectively. Figure layout and description are otherwise similar to Figure 6.1.

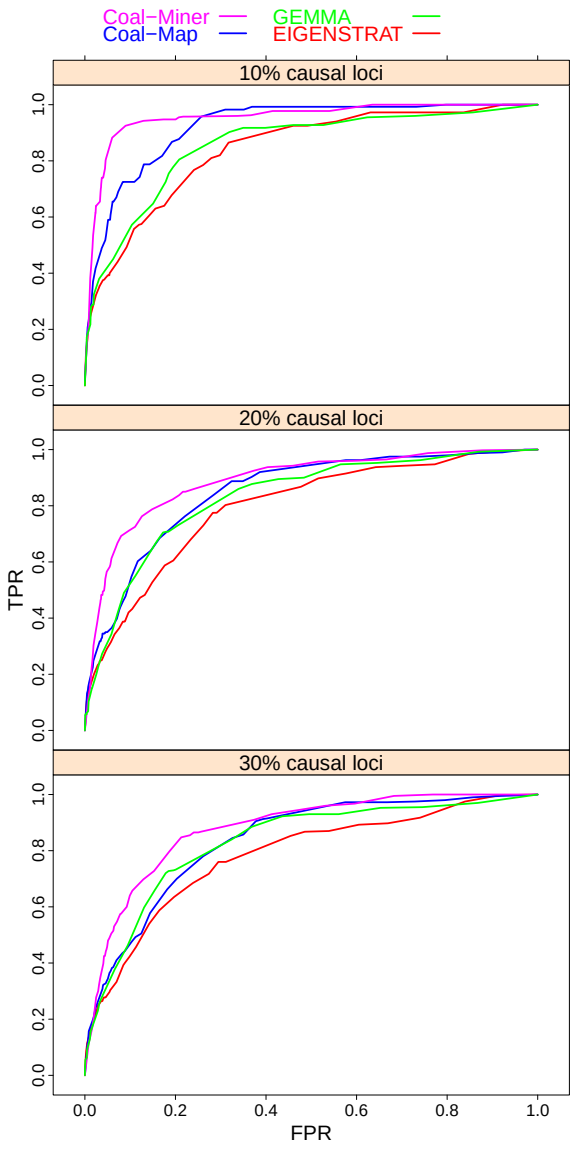
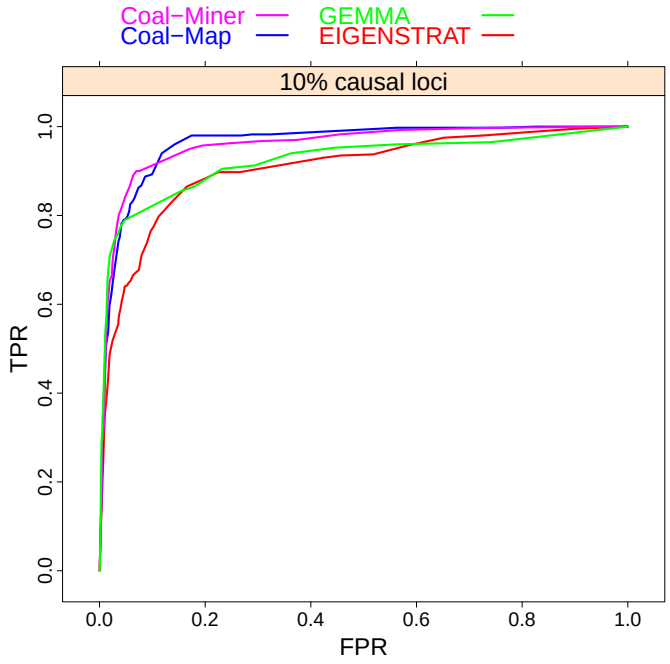
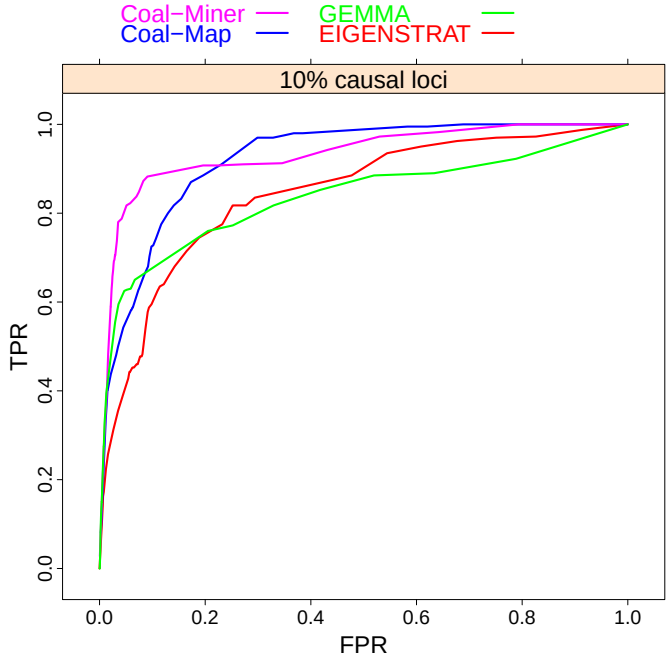


Figure D.2: **For simulations involving non-neutral with tree-like model phylogeny (hybridization frequency γ = 0 and selection coefficient s = 0.56), Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA.** The AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA for the 10% causal loci model condition are 0.954, 0.922, 0.841, and 0.856, respectively. For the 20% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.890, 0.850, 0.796, and 0.832, respectively. For the 30% causal loci model condition, the AUROC of Coal-Miner, Coal-Map, EIGENSTRAT, and GEMMA are 0.879, 0.836, 0.783, and 0.830, respectively. Figure layout and description are otherwise similar to Figure 6.1.

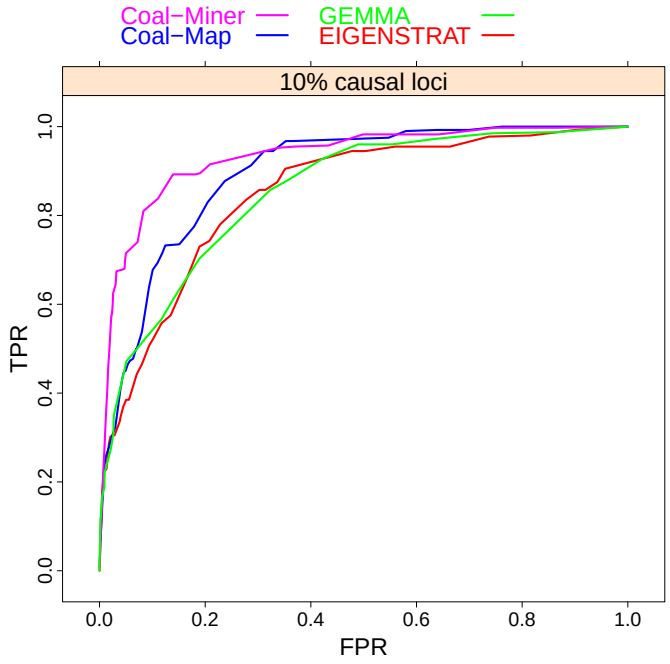


(a) $t_1 = 1$

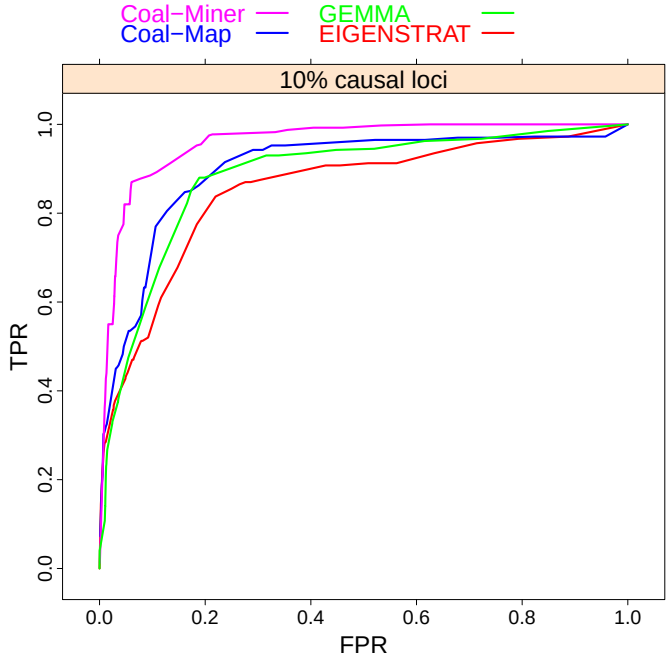


(b) $t_1 = 2.9$

Figure D.3: **Simulations involving neutral with non-tree-like model phylogeny (hybridization frequency $\gamma = 0.5$) along with admixture times of (a) $t_1 = 1$ and (b) $t_1 = 2.9$.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Figure 6.1.



(a) $t_1 = 1$



(b) $t_1 = 2.9$

Figure D.4: **Simulations involving neutral with tree-like model phylogeny (hybridization frequency $\gamma = 0$) along with divergence times of (a) $t_1 = 1$ and (b) $t_1 = 2.9$.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Figure 6.1.

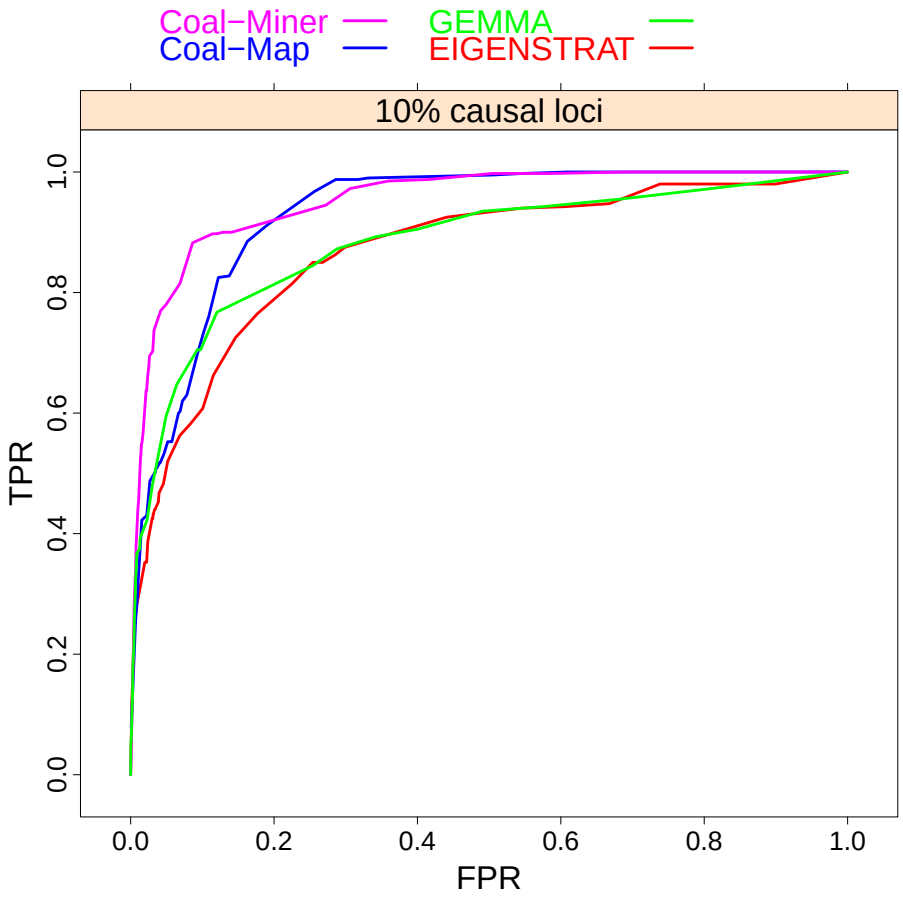


Figure D.5: **Simulations involving neutral with non-tree-like model phylogeny incorporating an isolation-with-migration (IM) model of gene flow.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Figure 6.1.

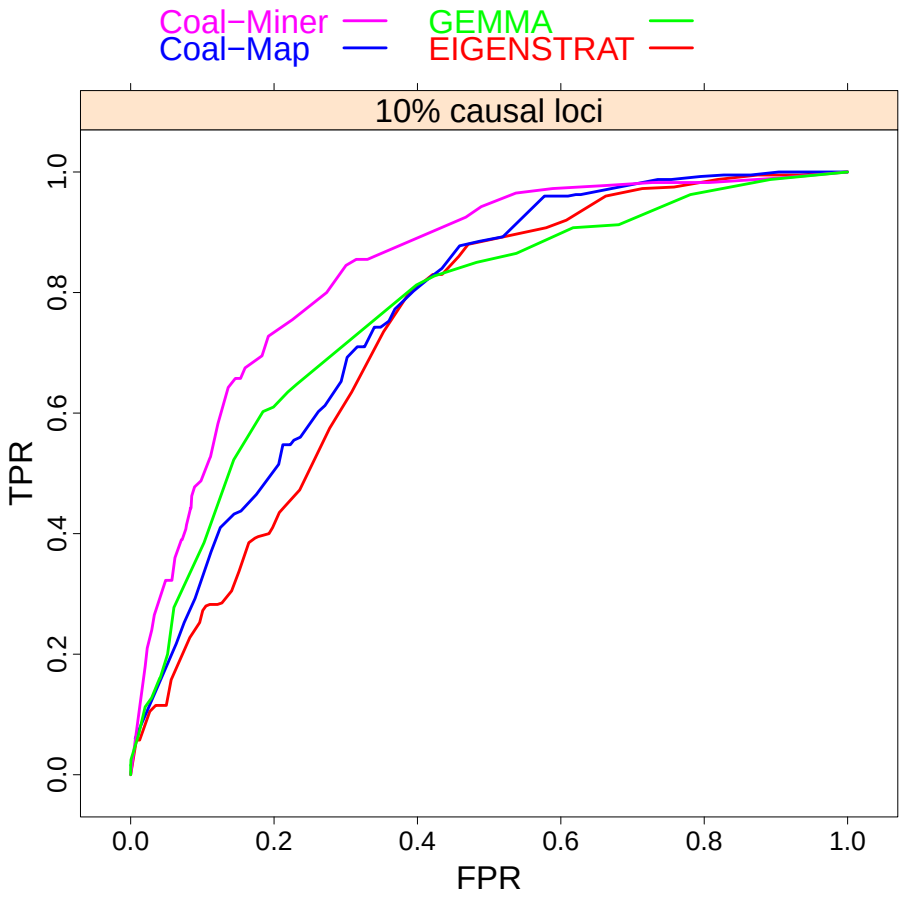


Figure D.6: **Simulations involving neutral with tree-like model phylogeny incorporating recombination.** Coal-Miner has an equal or better power and comparable type I error to Coal-Map, EIGENSTRAT, and GEMMA. Figure layout and description are otherwise similar to Figure 6.1.

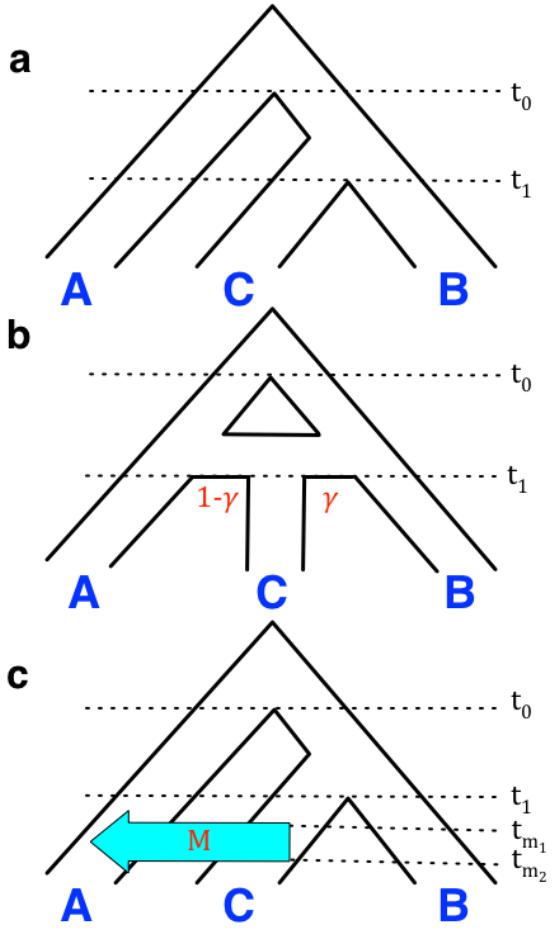
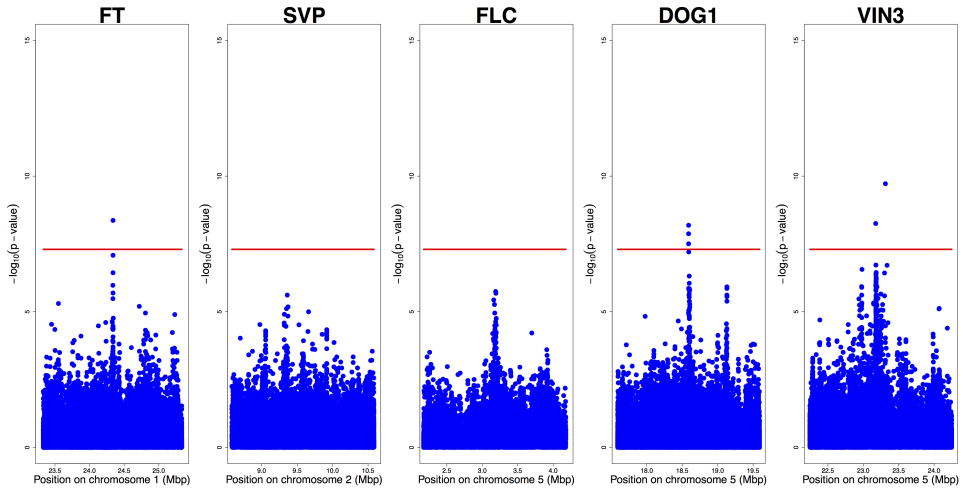
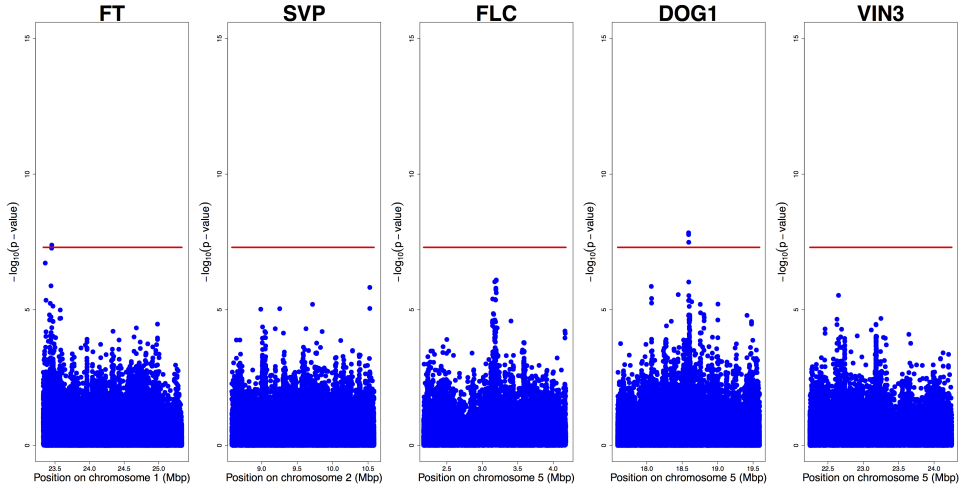


Figure D.7: **Model phylogenies used in the simulation study.** (a) Tree-like phylogeny, (b) Non-tree-like phylogeny with instantaneous unidirectional admixture (IUA), and (c) Non-tree-like phylogeny incorporating an isolation-with-migration (IM) model of gene flow.



(a) 10 °C



(b) 16 °C

Figure D.8: **Results showing the Manhattan plots after applying GEMMA on the *Arabidopsis* dataset using two model conditions: (a) flowering time at 10 °C and (b) flowering time at 16 °C.** The genes known to regulate flowering were added based on their respective position. Figure layout and description are otherwise similar to Figure 6.3.

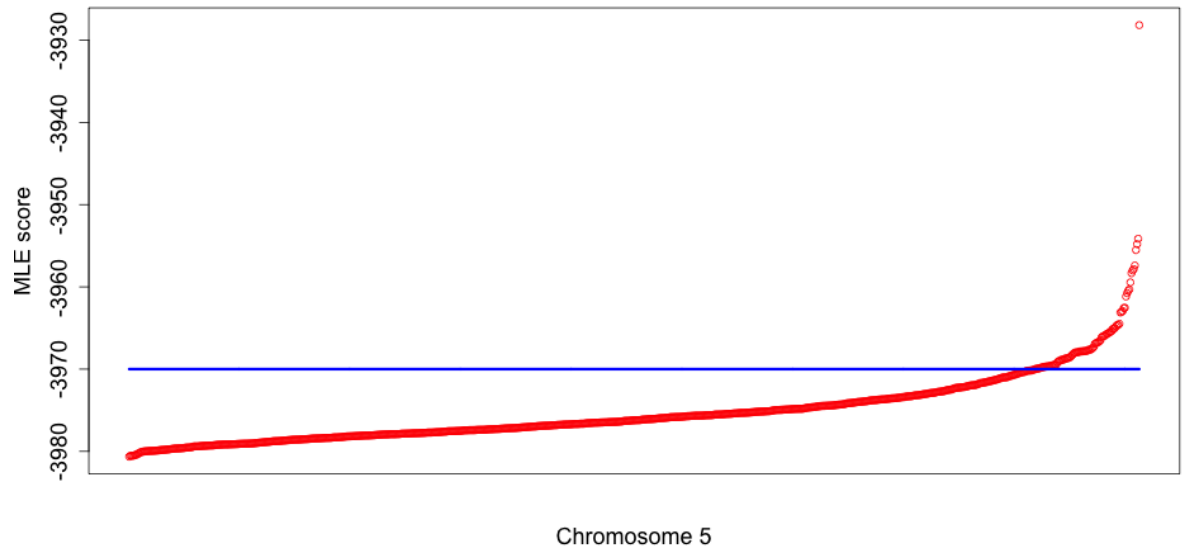


Figure D.9: **Distribution of likelihood scores in the second stage of Coal-Miner for loci in chromosome 5 (*Arabidopsis* dataset).** The x axis represents chromosome 5, and the y axis represents the likelihood scores. Results are shown for the flowering time at 10 °C model condition. Each circle represents a genomic locus. The blue line represents the threshold, which is the point of inflection in the distribution, that was used to detect candidate loci (i.e. loci that contain putatively associated sites). Any circles located above the threshold are considered candidate loci.

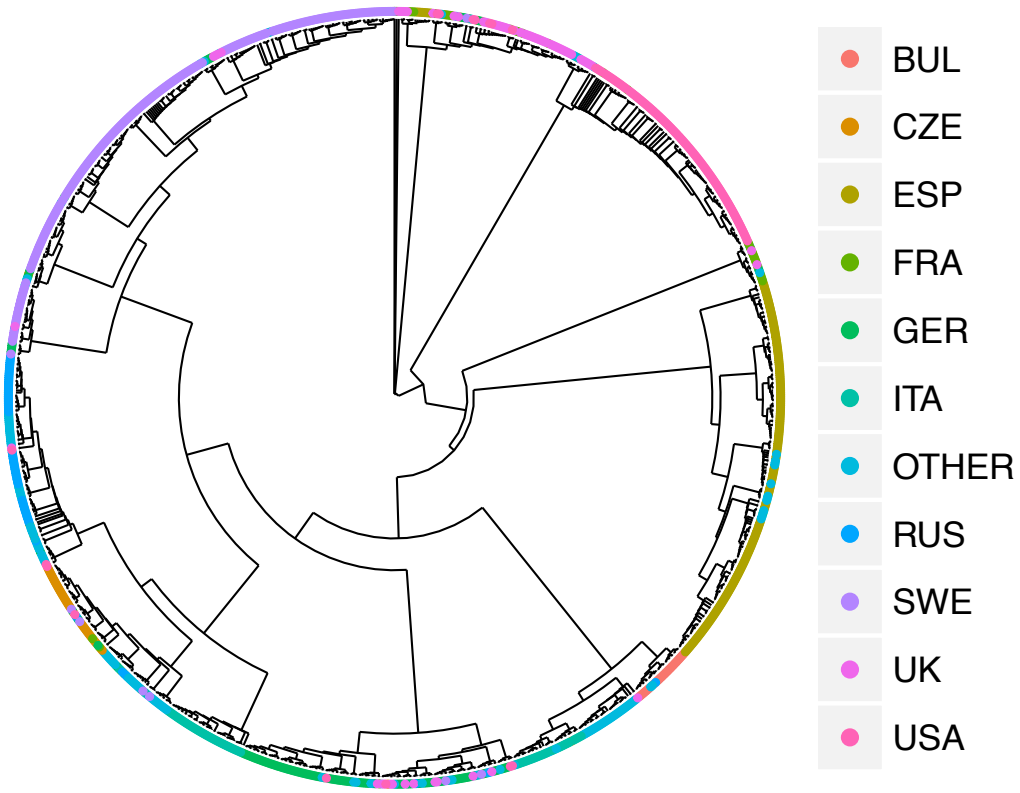


Figure D.10: **The phylogeny inferred from the 1,135 *Arabidopsis* strains using RAxML.** Each tip in the phylogeny is colored according to its country code. The legend represents the different countries in the analysis (BUL: Bulgaria, CZE: Czech Republic, ESP: Spain, FRA: France, GER: Germany, ITA: Italy, OTHER: Other countries, RUS: Russia, SWE: Sweden, UK: United Kingdom, USA: United States of America). The R package phytools (Revell, 2012) was used to plot the phylogeny.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

Alexander, David H., John Novembre & Kenneth Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19(9). 1655–1664. doi:10.1101/gr.094052.109. http://genome.cshlp.org/content/early/2009/07/31/gr.094052.109.abstract.

Alonso-Blanco, Carlos, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M. Borgwardt, Jun Cao, Eunyoung Chae, Todd M. Dezwaan, Wei Ding, Joseph R. Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela M. Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Cheng-Ruei Lee, Dazhe Meng, Todd P. Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Yu. Novikova, F. Xavier Picó, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salomé, Karl J. Schmid, Robert J. Schmitz, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel & Xuefeng Zhou. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166(2). 481 – 491. doi:http://dx.doi.org/10.1016/j.cell.2016.05.063. http://www.sciencedirect.com/science/article/pii/S0092867416306675.

Amasino, Richard M & Scott D Michaels. 2010. The timing of flowering. *Plant Physiology* 154(2). 516–520.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin & Gavin Sherlock. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1). 25–29. doi:10.1038/75556. http://dx.doi.org/10.1038/75556.

Astle, William & David J Balding. 2009. Population structure and cryptic relatedness in genetic association studies. *Statistical Science* 24(4). 451–471.

Baack, Eric J & Loren H Rieseberg. 2007. A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics & Development* 17(6). 513–518.

Bapteste, Eric, Leo van Iersel, Axel Janke, Scot Kelchner, Steven Kelk, James O. McInerney,

David A. Morrison, Luay Nakhleh, Mike Steel, Leen Stougie & James Whitfield. 2013. Networks: expanding evolutionary thinking. *Trends in Genetics* 29(8). 439–441.

Barrett, Jeffrey C, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M Michael Barmada, Alain Bitton, Themistocles Dassopoulos, Lisa Wu Datta, Todd Green, Anne M Griffiths, Emily O Kistner, Michael T Murtha, Miguel D Regueiro, Jerome I Rotter, L Philip Schumm, A Hillary Steinhart, Stephan R Targan, Ramnik J Xavier, Cecile Libioulle, Cynthia Sandor, Mark Lathrop, Jacques Belaiche, Olivier Dewit, Ivo Gut, Simon Heath, Debby Laukens, Myriam Mni, Paul Rutgeerts, Andre Van Gossum, Diana Zelenika, Denis Franchimont, Jean-Pierre Hugot, Martine de Vos, Severine Vermeire, Edouard Louis, Lon R Cardon, Carl A Anderson, Hazel Drummond, Elaine Nimmo, Tariq Ahmad, Natalie J Prescott, Clive M Onnie, Sheila A Fisher, Jonathan Marchini, Jilur Ghori, Suzannah Bumpstead, Rhian Gwilliam, Mark Tremelling, Panos Deloukas, John Mansfield, Derek Jewell, Jack Satsangi, Christopher G Mathew, Miles Parkes, Michel Georges & Mark J Daly. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* 40(8). 955–962.

Beceiro, Alejandro, María Tomás & Germán Bou. 2013. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clinical Microbiology Reviews* 26(2). 185–230.

Benjamini, Y. & Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57(1). 289–300.

Brent, R. P. 1973. *Algorithms for minimization without derivatives*. Mineola, New York: Dover Publications.

Bryant, David, Remco Bouckaert, Joseph Felsenstein, Noah A Rosenberg & Arindam RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29(8). 1917–1932.

Bryant, David & Vincent Moulton. 2004. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255–265.

Bush, William S. & Jason H. Moore. 2012. Genome-wide association studies. *PLoS Computational Biology* 8(12). e1002822.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Gunther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Muller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J Schmid & Detlef Weigel. 2011. Whole-

genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics* 43(10). 956–963. http://dx.doi.org/10.1038/ng.911.

Chifman, Julia & Laura Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30(23). 3317–3324.

Consortium, The Heliconious Genome. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487(7405). 94–98. http://dx.doi.org/10.1038/nature11041.

Davidson, Ruth, Pranjal Vachaspati, Siavash Mirarab & Tandy Warnow. 2015. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics* 16(Suppl 10). S1.

Degnan, James H & Noah A Rosenberg. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genetics* 2(5). 1–7. doi:10.1371/journal.pgen.0020068. https://doi.org/10.1371/journal.pgen.0020068.

Degnan, James H & Noah A Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* 24(6). 332–340.

Degnan, James H. & Laura A. Salter. 2005. Gene tree distributions under the coalescent process. *Evolution* 59(1). 24–37. doi:10.1111/j.0014-3820.2005.tb00891.x. http://dx.doi.org/10.1111/j.0014-3820.2005.tb00891.x.

DeLong, Elizabeth R, David M DeLong & Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3). 837–845.

Delsuc, Frédéric, Henner Brinkmann & Hervé Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* 6. 361–375. http://dx.doi.org/10.1038/nrg1603.

Devlin, B & Kathryn Roeder. 1999. Genomic control for association studies. *Biometrics* 55(4). 997–1004.

Drevinek, P & E Mahenthiralingam. 2010. *Burkholderia cenocepacia* in cystic fibrosis: epidemiology and molecular mechanisms of virulence. *Clinical Microbiology and Infection* 16(7). 821–830.

Durand, Eric Y., Nick Patterson, David Reich & Montgomery Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28(8). 2239–2252. http://mbe.oxfordjournals.org/content/28/8/2239.abstract.

Edwards, Albert O, Robert Ritter, Kenneth J Abel, Alisa Manning, Carolien Panhuysen & Lindsay A Farrer. 2005. Complement factor H polymorphism and age-related macular degeneration. *Science* 308(5720). 421–424.

Edwards, Scott V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63(1). 1–19.

Ellstrand, Norman C., Richard Whitkus & Loren H. Rieseberg. 1996. Distribution of spontaneous plant hybrids. *Proceedings of the National Academy of Sciences* 93(10). 5090–5093.

Ewing, Gregory & Joachim Hermisson. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26(16). 2064–2065.

Felsenstein, Joseph. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology* 27(4). 401–410.

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

Fontaine, Michael C., James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang, Andrew B. Hall, Flaminia Catteruccia, Evdoxia Kakani, Sara N. Mitchell, Yi-Chieh Wu, Hilary A. Smith, R. Rebecca Love, Mara K. Lawniczak, Michel A. Slotman, Scott J. Emrich, Matthew W. Hahn & Nora J. Besansky. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347(6217). 1258524. doi:10.1126/science.1258524. http://science.sciencemag.org/content/347/6217/1258524.

Gan, Xiangchao, Oliver Stegle, Jonas Behr, Joshua G. Steffen, Philipp Drewe, Katie L. Hildebrand, Rune Lyngsoe, Sebastian J. Schultheiss, Edward J. Osborne, Vipin T. Sreedharan, Andre Kahles, Regina Bohnert, Geraldine Jean, Paul Derwent, Paul Kersey, Eric J. Belfield, Nicholas P. Harberd, Eric Kemen, Christopher Toomajian, Paula X. Kover, Richard M. Clark, Gunnar Ratsch & Richard Mott. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477(7365). 419–423. http://dx.doi.org/10.1038/nature10414.

Gatesy, John & Mark S Springer. 2013. Concatenation versus coalescence versus "concatalescence". *Proceedings of the National Academy of Sciences* 110(13). E1179–E1179.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich & Svante Pääbo. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979). 710–722. http://www.sciencemag.org/content/328/5979/710.abstract.

Grenfell, Bryan T., Oliver G. Pybus, Julia R. Gog, James L. N. Wood, Janet M. Daly, Jenny A. Mumford & Edward C. Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303(5656). 327–332. http://www.sciencemag.org/content/303/5656/327.abstract.

Grenier, Jennifer & Scott Weatherbee. 2001. *From DNA to diversity: Molecular genetics and the evolution of animal design*. Wiley-Blackwell.

Guénet, Jean-Louis & François Bonhomme. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends in Genetics* 19(1). 24–31. doi:10.1016/S0168-9525(02)00007-0. http://www.sciencedirect.com/science/article/pii/S0168952502000070.

Hagmann, Jörg, Claude Becker, Jonas Müller, Oliver Stegle, Rhonda C. Meyer, George Wang, Korbinian Schneeberger, Joffrey Fitz, Thomas Altmann, Joy Bergelson, Karsten Borgwardt & Detlef Weigel. 2015. Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genetics* 11(1). 1–18. doi:10.1371/journal.pgen.1004920. http://dx.doi.org/10.1371%2Fjournal.pgen.1004920.

Haines, Jonathan L., Michael A. Hauser, Silke Schmidt, William K. Scott, Lana M. Olson, Paul Gallins, Kylee L. Spencer, Shu Ying Kwan, Maher Noureddine, John R. Gilbert, Nathalie Schnetz-Boutaud, Anita Agarwal, Eric A. Postel & Margaret A. Pericak-Vance. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308(5720). 419–421.

Hao, Weilong & Gary Golding. 2004. Patterns of bacterial gene movement. *Molecular Biology and Evolution* 21(7). 1294–1307.

Hein, Jotun, Mikkel Schierup & Carsten Wiuf. 2004. *Gene genealogies, variation and evolution:*

a primer in coalescent theory. Oxford: Oxford University Press.

Hejase, Hussein A & Kevin J Liu. 2016a. Mapping the genomic architecture of adaptive traits with interspecific introgressive origin: a coalescent-based approach. *BMC Genomics* 17(1). S8.

Hejase, Hussein A. & Kevin J. Liu. 2016b. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics* 17(1). 422. doi:10.1186/s12859-016-1277-1. http://dx.doi.org/10.1186/s12859-016-1277-1.

Hejase, Hussein A, Natalie Vande Pol, Gregory M Bonito, Patrick P Edger & Kevin J Liu. 2017a. Coal-Miner: A statistical method for GWA studies of quantitative traits with complex evolutionary origins. In *Proceedings of the 8th acm international conference on bioinformatics, computational biology, and health informatics*, 107–114. ACM.

Hejase, Hussein A, Natalie VandePol, Gregory A Bonito & Kevin J Liu. 2017b. FastNet: Fast and accurate inference of phylogenetic networks using large-scale genomic sequence data. *bioRxiv* doi:10.1101/132795. http://www.biorxiv.org/content/early/2017/05/01/132795.

Horton, Matthew W, Angela M Hancock, Yu S Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N Wayan Muliyati, Alexander Platt, F Gianluca Sperone, Bjarni J Vilhjalmsson, Magnus Nordborg, Justin O Borevitz & Joy Bergelson. 2012. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics* 44(2). 212–216. http://dx.doi.org/10.1038/ng.1042.

Hudson, Richard R. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* 23(2). 183–201.

Hudson, Richard R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2). 337–338. doi:10.1093/bioinformatics/18.2.337. http://bioinformatics.oxfordjournals.org/content/18/2/337.abstract.

Hudson, Richard R. & Norman L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1). 147–164. http://www.genetics.org/content/111/1/147.abstract.

Huerta-Sánchez, Emilia, Xin Jin, Asan, Zhuoma Bianba, Benjamin M. Peter, Nicolas Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixiang Ni, Bo Wang, Xiaohua Ou, Huasang, Jiangbai Luosang, Zha Xi Ping Cuo, Kui Li, Guoyi Gao, Ye Yin, Wei Wang, Xiuqing Zhang, Xun Xu, Huanming Yang, Yingrui Li, Jian Wang, Jun Wang & Rasmus Nielsen. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*

512(7513). 194–197.

Huo, Heqiang, Shouhui Wei & Kent J. Bradford. 2016. DELAY OF GERMINATION1 (DOG1)regulates both seed dormancy and flowering time through microRNA pathways. *Proceedings of the National Academy of Sciences* 113(15). E2199–E2206. doi:10.1073/pnas.1600558113. http://www.pnas.org/content/113/15/E2199.abstract.

Huson, Daniel H & Celine Scornavacca. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* 61(6). 1061–1067.

Jensen-Seaman, Michael I, Terrence S Furey, Bret A Payseur, Yontao Lu, Krishna M Roskin, Chin-Fu Chen, Michael A Thomas, David Haussler & Howard J Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research* 14(4). 528–538.

Jukes, Thomas H. & Charles R. Cantor. 1969. *Evolution of protein molecules* 21–132. Academic Press.

Kanehisa, Minoru & Susumu Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28(1). 27–30.

Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-yee Kong, Nelson B. Freimer, Chiara Sabatti & Eleazar Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42(4). 348–354. doi:10.1038/ng.548. http://dx.doi.org/10.1038/ng.548.

Kang, Hyun Min, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly & Eleazar Eskin. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178(3). 1709–1723.

Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellaker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jerome Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, Jose Afonso Guerra-Assuncao, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint & David J. Adams. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477(7364). 289–294. doi:10.1038/nature10413. http://dx.doi.org/10.1038/nature10413.

Kellis, Manolis, Nick Patterson, Matthew Endrizzi, Bruce Birren & Eric S. Lander. 2003. Se-

quencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937). 241–254. http://dx.doi.org/10.1038/nature01644.

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

Kurland, C. G., B. Canback & Otto G. Berg. 2003. Horizontal gene transfer: A critical view. *Proceedings of the National Academy of Sciences* 100(17). 9658–9662.

Leaché, Adam D, Rebecca B Harris, Bruce Rannala & Ziheng Yang. 2014. The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology* 63(1). 17–30.

Lechner, Marcus, Sven Findeiß, Lydia Steiner, Manja Marz, Peter F Stadler & Sonja J Prohaska. 2011. Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics* 12(1). 124.

Lefebre, Matthew D & Miguel A Valvano. 2002. Construction and evaluation of plasmid vectors optimized for constitutive and regulated gene expression in *Burkholderia cepacia* complex isolates. *Applied and Environmental Microbiology* 68(12). 5956–5964.

Lieberman, Tami D, Jean-Baptiste Michel, Mythili Aingaran, Gail Potter-Bynoe, Damien Roux, Michael R Davis, David Skurnik, Nicholas Leiby, John J LiPuma, Joanna B Goldberg, Alexander J McAdam, Gregory P Priebe & Roy Kishony. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics* 43(12). 1275–1280. http://dx.doi.org/10.1038/ng.997.

Liu, K., S. Raghavan, S. Nelesen, C. R. Linder & T. Warnow. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324(5934). 1561–1564.

Liu, Kevin J., Jingxuan Dai, Kathy Truong, Ying Song, Michael H. Kohn & Luay Nakhleh. 2014. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology* 10(6). e1003649. http://dx.doi.org/10.1371%2Fjournal.pcbi.1003649.

Liu, Kevin J., Ethan Steinberg, Alexander Yozzo, Ying Song, Michael H. Kohn & Luay Nakhleh. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences* 112(1). 196–201. doi:10.1073/pnas.1406298111. http://www.pnas.org/content/112/1/196.abstract.

Long, Anthony D & Charles H Langley. 1999. The power of association studies to detect the

contribution of candidate genetic loci to variation in complex traits. *Genome Research* 9(8). 720–731.

Long, Quan, Fernando A Rabanal, Dazhe Meng, Christian D Huber, Ashley Farlow, Alexander Platzer, Qingrun Zhang, Bjarni J Vilhjalmsson, Arthur Korte, Viktoria Nizhynska, Viktor Voronin, Pamela Korte, Laura Sedman, Terezie Mandakova, Martin A Lysak, Umit Seren, Ines Hellmann & Magnus Nordborg. 2013. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics* 45(8). 884–890. http://dx.doi.org/10.1038/ng.2678.

Ma, Jian. 2011. Reconstructing the history of large-scale genomic changes: biological questions and computational challenges. *Journal of Computational Biology* 18(7). 879–893.

Maddison, Wayne P. 1997. Gene trees in species trees. *Systematic Biology* 46(3). 523–536.

Marchini, Jonathan, Lon R Cardon, Michael S Phillips & Peter Donnelly. 2004. The effects of human population structure on large genetic association studies. *Nature Genetics* 36(5). 512–517.

McPherson, Ruth, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R. Cox, David A. Hinds, Len A. Pennacchio, Anne Tybjaerg-Hansen, Aaron R. Folsom, Eric Boerwinkle, Helen H. Hobbs & Jonathan C. Cohen. 2007. A common allele on chromosome 9 associated with coronary heart disease. *Science* 316(5830). 1488–1491.

Méndez-Vigo, Belén, F Xavier Picó, Mercedes Ramiro, JoséM Martínez-Zapater & Carlos Alonso-Blanco. 2011. Altitudinal and Climatic Adaptation Is Mediated by Flowering Traits and FRI, FLC, and PHYC Genes in Arabidopsis. *Plant Physiology* 157(4). 1942–1955. doi:10.1104/pp.111.183426. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3327218/.

Metzker, Michael L. 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics* 11(1). 31–46.

Michener, Charles D & Robert R Sokal. 1957. A quantitative approach to a problem in classification. *Evolution* 11(2). 130–162.

Mirarab, Siavash, Md Shamsuzzoha Bayzid, Bastien Boussau & Tandy Warnow. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346(6215). 1250463.

Mirarab, Siavash, Md Shamsuzzoha Bayzid & Tandy Warnow. 2014b. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology* 65(3). 366–380.

Mirarab, Siavash, Rezwana Reaz, Md S Bayzid, Theo Zimmermann, M Shel Swenson & Tandy Warnow. 2014c. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17). i541–i548.

Mirarab, Siavash & Tandy Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31(12). i44–i52.

Moran, Gary P, David C Coleman & Derek J Sullivan. 2011. Comparative genomics and the evolution of pathogenicity in human pathogenic fungi. *Eukaryotic Cell* 10(1). 34–42. doi: 10.1128/EC.00242-10. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3019795/.

Moret, Bernard M. E., Luay Nakhleh, Tandy Warnow, C. Randal Linder, Anna Tholse, Anneke Padolina, Jerry Sun & Ruth Timme. 2004. Phylogenetic networks: Modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(1). 13–23.

Munikumar, Manne, I Vani Priyadarshini, Dibyabhaba Pradhan, Amineni Umamaheswari & Bhuma Vengamma. 2013. Computational approaches to identify common subunit vaccine candidates against bacterial meningitis. *Interdisciplinary Sciences* 5(2). 155–164.

Nakhleh, Luay. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution* 28(12). 719–728. doi:http://dx.doi.org/10.1016/j.tree.2013.09.004. http://www.sciencedirect.com/science/article/pii/S0169534713002139.

Nakhleh, Luay, Jerry Sun, Tandy Warnow, C Randal Linder, Bernard ME Moret & Anna Tholse. 2003. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Pacific Symposium on Biocomputing*, vol. 8, 315–326. World Scientific.

Neafsey, Daniel E., Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, Bruno Arcà, Peter Arensburger, Gleb Artemov, Lauren A. Assour, Hamidreza Basseri, Aaron Berlin, Bruce W. Birren, Stephanie A. Blandin, Andrew I. Brockman, Thomas R. Burkot, Austin Burt, Clara S. Chan, Cedric Chauve, Joanna C. Chiu, Mikkel Christensen, Carlo Costantini, Victoria L. M. Davidson, Elena Deligianni, Tania Dottorini, Vicky Dritsou, Stacey B. Gabriel, Wamdaogo M. Guelbeogo, Andrew B. Hall, Mira V. Han, Thaung Hlaing, Daniel S. T. Hughes, Adam M. Jenkins, Xiaofang Jiang, Irwin Jungreis, Evdoxia G. Kakani, Maryam Kamali, Petri Kemppainen, Ryan C. Kennedy, Ioannis K. Kirmitzoglou, Lizette L. Koekemoer, Njoroge Laban, Nicholas Langridge, Mara K. N. Lawniczak, Manolis Lirakis, Neil F. Lobo, Ernesto Lowy, Robert M. MacCallum, Chunhong Mao, Gareth Maslen, Charles Mbogo, Jenny McCarthy, Kristin Michel, Sara N. Mitchell, Wendy Moore, Katherine A. Murphy, Anastasia N. Naumenko, Tony Nolan, Eva M. Novoa, Samantha O'Loughlin, Chioma Oringanje, Mohammad A. Oshaghi, Nazzy Pakpour, Philippos A. Pap-

athanos, Ashley N. Peery, Michael Povelones, Anil Prakash, David P. Price, Ashok Rajaraman, Lisa J. Reimer, David C. Rinker, Antonis Rokas, Tanya L. Russell, N'Fale Sagnon, Maria V. Sharakhova, Terrance Shea, Felipe A. Simão, Frederic Simard, Michel A. Slotman, Pradya Somboon, Vladimir Stegniy, Claudio J. Struchiner, Gregg W. C. Thomas, Marta Tojo, Pantelis Topalis, José M. C. Tubio, Maria F. Unger, John Vontas, Catherine Walton, Craig S. Wilding, Judith H. Willis, Yi-Chieh Wu, Guiyun Yan, Evgeny M. Zdobnov, Xiaofan Zhou, Flaminia Catteruccia, George K. Christophides, Frank H. Collins, Robert S. Cornman, Andrea Crisanti, Martin J. Donnelly, Scott J. Emrich, Michael C. Fontaine, William Gelbart, Matthew W. Hahn, Immo A. Hansen, Paul I. Howell, Fotis C. Kafatos, Manolis Kellis, Daniel Lawson, Christos Louis, Shirley Luckhart, Marc A. T. Muskavitch, José M. Ribeiro, Michael A. Riehle, Igor V. Sharakhov, Zhijian Tu, Laurence J. Zwiebel & Nora J. Besansky. 2015. Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science* 347(6217). doi:10.1126/science.1258522. http://science.sciencemag.org/content/347/6217/1258522.

Noor, Mohamed A. F. & Jeffrey L. Feder. 2006. Speciation genetics: evolving approaches. *Nature Reviews Genetics* 7(11). 851–861.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* 29(1). 59–75.

Ochman, Howard, Jeffrey G Lawrence & Eduardo A Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784). 299–304.

Paradis, Emmanuel, Julien Claude & Korbinian Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2). 289–290.

Philippe, Herve, Henner Brinkmann, Dennis V Lavrov, D Timothy J Littlewood, Michael Manuel, Gert Wörheide & Denis Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology* 9(3). 1–10.

Porter, Stephanie S, Peter L Chang, Christopher A Conow, Joseph P Dunham & Maren L Friesen. 2017. Association mapping reveals novel serpentine adaptation gene clusters in a population of symbiotic mesorhizobium. *The ISME Journal* 11(1). 248–262. doi:10.1038/ismej.2016.88. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5315480/.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick & David Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8). 904–909.

Price, Alkes L, Noah A Zaitlen, David Reich & Nick Patterson. 2010a. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* 11(7).

459–463.

Price, Morgan N., Paramvir S. Dehal & Adam P. Arkin. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* 26(7). 1641–1650. doi:10.1093/molbev/msp077. http://mbe.oxfordjournals.org/content/26/7/1641.abstract.

Price, Morgan N., Paramvir S. Dehal & Adam P. Arkin. 2010b. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3). e9490. doi:10.1371/journal.pone.0009490.

Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2). 945–959. http://www.genetics.org/content/155/2/945.

Rambaut, A. & N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13(3). 235–238.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin & Svante Paabo. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327). 1053–1060. http://dx.doi.org/10.1038/nature09710.

Revell, Liam J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3(2). 217–223.

Robinson, David F & Leslie R Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53(1). 131–147.

Rost, Simone, Andreas Fregin, Vytautas Ivaskevicius, Ernst Conzelmann, Konstanze Hortnagel, Hans-Joachim Pelz, Knut Lappegard, Erhard Seifried, Inge Scharrer, Edward G. D. Tuddenham, Clemens R. Muller, Tim M. Strom & Johannes Oldenburg. 2004. Mutations in *Vkorc1* cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* 427(6974). 537–541. http://dx.doi.org/10.1038/nature02214.

Saitou, Naruya & Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4). 406–425.

Salichos, Leonidas & Antonis Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449). 327–331. http://dx.doi.org/10.1038/nature12130.

Sanderson, Michael J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19(2). 301–302.

Schliep, KP. 2009. *Some applications of statistical phylogenetics*: Massey University dissertation.

Schmitz, Robert J., Matthew D. Schultz, Mark A. Urich, Joseph R. Nery, Mattia Pelizzola, Ondrej Libiger, Andrew Alix, Richard B. McCosh, Huaming Chen, Nicholas J. Schork & Joseph R. Ecker. 2013. Patterns of population epigenomic diversity. *Nature* 495(7440). 193–198. http://dx.doi.org/10.1038/nature11968.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2). 461–464.

Scuteri, Angelo, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, Mariano Dei, Sandra Lai, Andrea Maschio, Fabio Busonero, Antonella Mulas, Georg B Ehret, Ashley A Fink, Alan B Weder, Richard S Cooper, Pilar Galan, Aravinda Chakravarti, David Schlessinger, Antonio Cao, Edward Lakatta & Gonçalo R Abecasis. 2007. Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genetics* 3(7). e115.

Shriner, Daniel. 2011. Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity* 107(5). 413–420.

Silva, Inês N, Pedro M Santos, Mário R Santos, James E A Zlosnik, David P Speert, Sean W Buskirk, Eric L Bruger, Christopher M Waters, Vaughn S Cooper & Leonilde M Moreira. 2016. Long-Term evolution of *Burkholderia multivorans* during a chronic cystic fibrosis infection reveals shifting forces of selection. *mSystems* 1(3). e00029–16.

Sober, Elliott. 1983. Parsimony in systematics: philosophical issues. *Annual Review of Ecology and Systematics* 14(1). 335–357.

Solís-Lemus, Claudia & Cécile Ané. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics* 12(3). 1–21.

Song, Ying, Stefan Endepols, Nicole Klemann, Dania Richter, Franz-Rainer Matuschka, Ching-Hua Shih, Michael W. Nachman & Michael H. Kohn. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology* 21(15). 1296

– 1301. http://www.sciencedirect.com/science/article/pii/S0960982211007160.

Sousa, Sílvia A, Joana R Feliciano, Tiago Pita, Soraia I Guerreiro & Jorge H Leitão. 2017. *Burkholderia cepacia* complex regulation of virulence gene expression: A review. *Genes* 8(1). 43.

Speed, Doug & David J Balding. 2014. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* 24(9). 1550–1557.

Staubach, Fabian, Anna Lorenc, Philipp W. Messer, Kun Tang, Dmitri A. Petrov & Diethard Tautz. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genetics* 8(8). e1002891.

Supple, Megan A, Heather M Hines, Kanchon K Dasmahapatra, James J Lewis, Dahlia M Nielsen, Christine Lavoie, David A Ray, Camilo Salazar, W Owen McMillan & Brian A Counterman. 2013. Genomic architecture of adaptive color pattern divergence and convergence in *Heliconius* butterflies. *Genome Research* 23(8). 1248–1257.

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

Than, Cuong, Derek Ruths & Luay Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9(1). 322.

Thomas, Christopher M & Kaare M Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* 3(9). 711–721.

Vachaspati, Pranjal & Tandy Warnow. 2015. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics* 16(10). 1–13. doi:10.1186/1471-2164-16-S10-S3. http://dx.doi.org/10.1186/1471-2164-16-S10-S3.

Voight, Benjamin F & Jonathan K Pritchard. 2005. Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics* 1(3). e32.

Wang, Liewei, Howard L McLeod & Richard M Weinshilboum. 2011. Genomics and drug response. *The New England Journal of Medicine* 364(12). 1144–1153.

Westesson, Oscar & Ian Holmes. 2009. Accurate detection of recombinant breakpoints in whole-genome alignments. *PLoS Computational Biology* 5(3). e1000318. doi:10.1371/journal.pcbi.1000318.

Yang, Hyuna, Yueming Ding, Lucie N. Hutchins, Jin Szatkiewicz, Timothy A. Bell, Beverly J. Paigen, Joel H. Graber, Fernando Pardo-Manuel de Villena & Gary A. Churchill. 2009. A customized and versatile high-density genotyping array for the mouse. *Nature Methods* 6(9). 663–666. doi:10.1038/nmeth.1359. http://dx.doi.org/10.1038/nmeth.1359.

Yang, Hyuna, Jeremy R. Wang, John P. Didion, Ryan J. Buus, Timothy A. Bell, Catherine E. Welsh, Francois Bonhomme, Alex Hon-Tsen Yu, Michael W. Nachman, Jaroslav Pialek, Priscilla Tucker, Pierre Boursot, Leonard McMillan, Gary A. Churchill & Fernando Pardo-Manuel de Villena. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics* 43(7). 648–655. doi:10.1038/ng.847. http://dx.doi.org/10.1038/ng.847.

Yang, Jimmy & Tandy Warnow. 2011. Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics* 12(9). S4.

Yang, Ziheng & Bruce Rannala. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13(5). 303–314.

Yeager, Meredith, Nick Orr, Richard B Hayes, Kevin B Jacobs, Peter Kraft, Sholom Wacholder, Mark J Minichiello, Paul Fearnhead, Kai Yu, Nilanjan Chatterjee, Zhaoming Wang, Robert Welch, Brian J Staats, Eugenia E Calle, Heather Spencer Feigelson, Michael J Thun, Carmen Rodriguez, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Fredrick R Schumacher, Edward Giovannucci, Walter C Willett, Geraldine Cancel-Tassin, Olivier Cussenot, Antoine Valeri, Gerald L Andriole, Edward P Gelmann, Margaret Tucker, Daniela S Gerhard, Joseph F Fraumeni, Robert Hoover, David J Hunter, Stephen J Chanock & Gilles Thomas. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics* 39(5). 645–649.

Yu, Yun, Robert M. Barnett & Luay Nakhleh. 2013a. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology* 62(5). 738–751.

Yu, Yun, James H. Degnan & Luay Nakhleh. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics* 8(4). e1002660.

Yu, Yun, Jianrong Dong, Kevin J. Liu & Luay Nakhleh. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences* 111(46). 16448–16453. doi:10.1073/pnas.1407950111. http://www.pnas.org/content/111/46/16448.abstract.

Yu, Yun & Luay Nakhleh. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16(Suppl 10). S10.

Yu, Yun, Nikola Ristic & Luay Nakhleh. 2013b. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics* 14(Suppl 15). S6.

Yu, Yun, Tandy Warnow & Luay Nakhleh. 2011. Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *Journal of Computational Biology* 18(11). 1543–1559. doi:10.1089/cmb.2011.0174. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3216099/`.

Yun, Yu. 2014. *Models and methods for evolutionary histories involving hybridization and incomplete lineage sorting*: Rice University dissertation.

Zeggini, Eleftheria, Michael N. Weedon, Cecilia M. Lindgren, Timothy M. Frayling, Katherine S. Elliott, Hana Lango, Nicholas J. Timpson, John R. B. Perry, Nigel W. Rayner, Rachel M. Freathy, Jeffrey C. Barrett, Beverley Shields, Andrew P. Morris, Sian Ellard, Christopher J. Groves, Lorna W. Harries, Jonathan L. Marchini, Katharine R. Owen, Beatrice Knight, Lon R. Cardon, Mark Walker, Graham A. Hitman, Andrew D. Morris, Alex S. F. Doney, The Wellcome Trust Case Control Consortium (WTCCC), Mark I. McCarthy & Andrew T. Hattersley. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316(5829). 1336–1341.

Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas & Edward S Buckler. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42(4). 355–360.

Zhou, Xiang & Matthew Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44(7). 821–824.